S

Safety

▸ Privacy, Dataveillance, and Crime Prevention

Sample Point, Simple Event, Elementary Event

▸ Theory of Probability, Basics and Fundamentals

Sampling Effects in Social Network Analysis

Rick Grannis
Department of Sociology, UCLA, Los Angeles, CA, USA

Synonyms

Network sampling; Nonlinearity; Phase transition; Social networks

Glossary

Social Actor An entity with the ability to create or nullify social relations

Social Relation Anything that exists, occurs, or flows between two social actors

Dyad A set of two social actors potentially or actually connected by a social relation

Triad A set of three social actors all of whom are potentially or actually connected to each other by social relations

First Neighbors Social actors who directly connect to each other via a social relation

Second Neighbors Social actors who do not directly connect to each other via a social relation but whom both directly connect to the same intermediary

Mth **Neighbor** Social actors who connect to each other via a path consisting of m (but no fewer) social relations involving $m-1$ (but no fewer) intermediaries

Clustering The finding, common to social networks, that two social actors who share a common first neighbor have a disproportionately great probability of also being first neighbors of each other (also referred to as transitivity)

Phase Transition A sharp combinatorial transition point such that only a relatively small increase in network density (or some other individual-level network characteristic) transitions the network from the situation in which virtually no network member connects via a path to any other network member to the situation in which most network members are connected via a path to most other network members (also referred to as a critical point or double jump threshold)

Giant Component A single component that connects the majority of network members

R. Alhajj, J. Rokne (eds.), *Encyclopedia of Social Network Analysis and Mining*,
DOI 10.1007/978-1-4614-6170-8,

Chain-referral Sampling A sampling technique that involves selecting initial respondents and then generating all future respondents by following along the contact network of those who have already sampled (also referred to as snowball, link-tracing, or random walk samples)

Definition

While a great many useful statistical models and visualizations have been developed to explore large-scale complex networks, fewer have attempted to relate these models to data generated by samples. While, in many fields, complete (or near-complete) data is widely available, and while the Internet has made even more readily available, complete data about large-scale complex networks sufficient to answer many compelling social science questions does not exist and cannot be reasonably generated. In these cases, sampling theory must be used to connect data to models. This proves difficult for a variety of reasons such as: data collection methodologies which, while attempting to overcome nonresponse bias, deviate from standard sampling practices; and, the non-independence of both first neighbors and second neighbors. This entry reviews some of the different ways which have been created to faithfully translate data sampled from large-scale complex networks into useful statistics.

Introduction

One of the fundamental goals of social science is to understand how the interactions of individuals translate into the characteristics of the social systems they comprise (Schelling 1978). Network models have a potentially powerful role in this task.

To model the social world in network terms is to focus on social entities and the social relations among them, on the patterning of relations among social actors rather than the correlation among social actors' attributes. Often these entities forming the basis of social networks are individual people, but they can also be households, business firms, nations, or any social actor with the ability to create or nullify social relations. A social relation can mean anything that occurs or flows between two social actors such as information (advice, gossip, conversation, etc.), tangible resources (being exchanged, stolen, etc.), and emotional affect (Borgatti et al. 2009). Social relations never involve only a single social actor, and most, although not all, of them require at least the implied consent of both parties to exist. Social actors a and b determine the quality and quantity of their interaction with each other, including none at all, based not only on the motivations to interact with each other but also based on the social resources available to them through their other relations. Social relations are not independent.

Some research into large, complex networks ignores this nonindependence characteristic of social networks, and other research models its effects only with first-order neighbors and not second-order neighbors, threatening their theoretical reliability. Methodologies have been developed, however, to overcome these statistical obstacles in some cases and to fruitfully exploit them in others. This entry reviews these methods.

Key Points

Unlike many other types of network data, gathered from large databases or the Internet, social network data typically derives from samples and thus requires mathematical theory to reliably translate these sample statistics into estimates of population parameters.

When sampling from a social network, the non-independence of the sampled individuals can be both a resource to be exploited and an obstacle to be overcome.

This entry reviews three distinct ways in which social network researcher have dealt with nonindependence:

1. Well-defined statistical models faithfully deriving local-level network properties from sample data

2. Network sampling techniques which attempt to exploit the social network structure to yield more robust estimates of population parameters
3. Statistical models translating local-level information derived from traditional samples into reliable estimates of large-scale network realities

Deriving Local-Level Network Properties from Sample Data

Many useful sample estimators have been derived for local-level network properties (Granovetter 1976). As an example, Frank (1978) showed that a dyad count, C, a count of the number of distinct types of the $\binom{n}{2}$ possible dyads which can be induced from the network, has an unbiased estimator given by

$$\hat{C} = \frac{N\,(N-1)\,Z}{n(n-1)}$$

where N is the total number of individuals in the population, n is the number of individuals in the sample, and Z is the dyad count within the sample and $n \geq 2$. If N is large, $n \geq 4$, and the sampling fraction n/N is a relatively small nonzero number p, the variance

$$\sigma^2 \approx \frac{C}{p^2}$$

has an unbiased estimator $\hat{\sigma}^2$:

$$\hat{\sigma}^2 \approx \frac{\sum_{i=1}^{N} Z_i^2 - Z}{p^4}$$

Frank (1978) further showed that a triad count C, a count of the number of distinct types of the $\binom{n}{3}$ possible dyads which can be induced from the network, has an unbiased estimator given by $\hat{C} = Z/p_3$, where Z is the triad count within the sample, $p_r = n^{(r)}/N^{(r)}$, and $n \geq 3$.

If the sampling fraction p is a small number and N is large and $n \geq 6$, we have $\sigma^2 \approx \mathrm{C}/p^3$ and the variance has an unbiased estimator

$$\hat{\sigma}^2 \approx \frac{Z - \sum_{i=1}^{N}\sum_{j=1}^{N}\frac{Z_{ij.}^2}{2} + \sum_{i=1}^{N}\frac{Z_{i..}^2}{4}}{p^6}$$

For decades, these types of statistics, among others, have proved foundational to the development of social network sampling theory. Other sample estimators have been derived which can be used to connect data sampled from a network to still other local-level network parameters (i.e., phenomena only depending on a single node or their immediate contacts). Deriving sample estimators for global-level network properties has proved more troublesome, however; the inherent non-independence of the sampled individuals and the relations interconnecting them inhibits analysis. Global-level properties, such as centrality and network centralization, for example, have proven to be highly sensitive to microlevel changes (Butts 2006, 2009).

Using Network Samples to Effectively Analyze Population Characteristics

While the nonindependence of sampled individuals, inherent to social networks, can potentially impede analysis when sampling from a network, under some circumstances, this same lack of independence can prove advantageous. One of these key circumstances involves estimating the characteristics of "hidden populations" such as drug users, the homeless, or artists (Heckathorn 1997), as well as the enormous set of nonresponders (both actual and potential) who would choose to opt out if a survey was requested and frustrate virtually all survey efforts (Grannis et al. 2011).

Standard sampling and estimation techniques, which require the researcher to select sample members with a known probability of selection, necessitate that researchers have a sampling frame listing all members in the population;

however, for many populations of interest, such a list does not exist. Chain-referral (also known as snowball (Coleman 1958) or link-tracing) samples, which select initial respondents but then generate all future respondents by following along the contact network of those who have already been sampled, have been shown to be effective at penetrating hidden populations even without a sampling frame.

Such samples' use has been limited, however, due to the difficulty of making statistical inferences. Since members of the population to be sampled do not have the same probability of selection, those with many contacts are more likely to be included in the sample than social isolates. Also, biases in those "seeding" the sample, those first selected and from whom all subsequent respondents were indirectly recruited, may compound in unknown ways as the sampling process continued.

Because of this, chain-referral samples have often been considered to be nonprobability or convenience samples "which can only be assessed by subjective evaluation" (Kalton 1983) and "conventional wisdom among sociologists, public health researchers, and statisticians is that chain-referral sampling holds great promise for a number of problems, especially the study of hidden populations, but that it is so hopelessly biased that it cannot be used to make reliable estimates" (Salganik and Heckathorn 2004, p. 197).

Salganik and Heckathorn (2004), however, showed that, for any reciprocal relation, we can recover the proportion of any population belonging to a group A, PP_A, with knowledge only of the network structure connecting the population, the type of data which chain-referral samples most reliably generate:

$$PP_A = \frac{\widehat{D_B}\widehat{C_{B,A}}}{\widehat{D_A}\widehat{C_{A,B}} + \widehat{D_B}\widehat{C_{B,A}}}$$

where $\widehat{D_A}$ is the estimate of the average number of relations an individual of type A is involved with and $\widehat{C_{AB}}$ is the proportion of relations originating from an individual of type A which connect with an individual of type B.

They showed that the numerator and denominator of this statistic are both unbiased (Brewer and Hanif 1983) estimates of the population parameters and that the ratio of these two unbiased estimators is asymptotically unbiased with bias on the order of n^{-1}, where n is the sample size (Cochran 1977).

The mean degree of a chain-referral sample would be higher than the mean degree of the population since these methods overrepresent people with high degree (Erickson 1979; Kalton and Anderson 1986; Eland-Goosensen et al. 1997); therefore Salganik and Heckathorn (2004) used two distinct mathematical approaches to show that, assuming only that nodes are drawn with probability proportional to their degree, the mean degree of the population can be estimated by

$$\widehat{D_A} = \frac{n_A}{\sum_{i=1}^{n_A} \frac{1}{d_i}}$$

where n_A is the number of individuals of type A in the sample, and d_i indexes over the number of individuals of type A in the sample.

Furthermore, Salganik and Heckathorn (2004) showed that since recruitments originating from a person of type A can be categorized into two sets, those that connect with another person in group A, r_{AA} and those which connect with a person in group B, r_{AB} and since the observed recruitments are a random sample from all edges, an unbiased estimate for $C_{A,B}$ is given by

$$\widehat{C_{A,B}} = \frac{r_{AB}}{r_{AA} + r_{AB}}$$

Thus, unlike a conventional probability sampling design, one can use data generated from a chain-referral sample not to directly estimate population parameters but rather to analyze the population *specifically as a network.* One uses the sample to make estimates about the network connecting the population and only then is this information about the network used to derive population proportions. By not attempting to

estimate directly from the sample proportions to population proportions, one avoids many of the well-known problems with chain-referral samples (Heckathorn 2002).

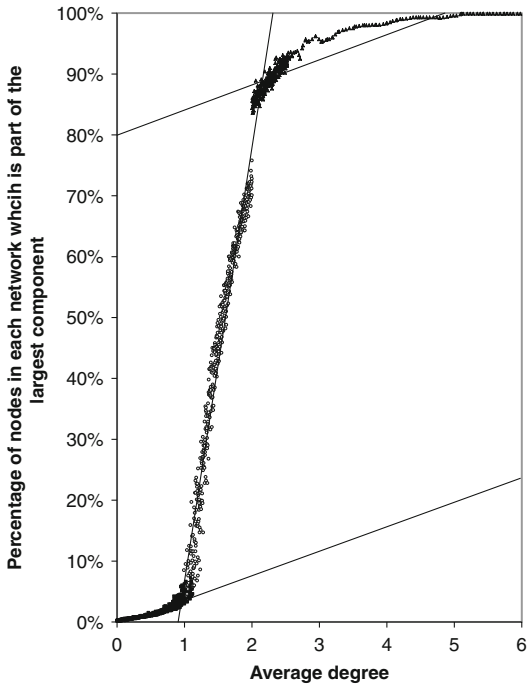
Modeling the Phase Transition from Sample Data

Arguably, the most fundamental global-level property of a large-scale, complex network concerns whether it contains a giant component, a single component that connects the majority of network members. If it does not, if the network is not connected, most other global network properties, like centralization, have little meaning. Furthermore, the only assumption about the structure of the network underlying chain-referral estimation procedures such as respondent-driven sampling is that the population basically consists of one connected component and that there exists a path between every person and every other person (Salganik and Heckathorn 2004).

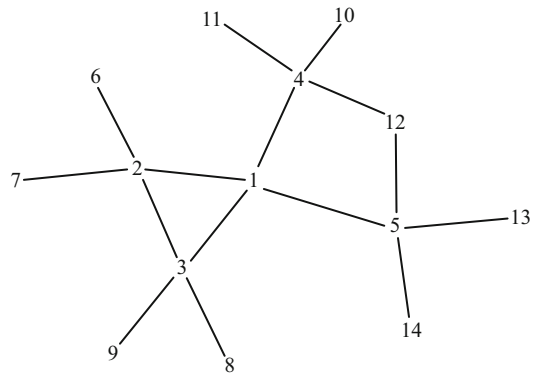
While they depend to some extent on a network's individual-level properties, giant components, however, do not emerge as a linear response to individual-level changes but rather subtle changes in relations potentially produce extraordinarily different macro-level outcomes. As the average number of relations among individuals increases, the size of the components that they form does not grow relatively smoothly from small to large. Instead, in virtually all networks, a sharp threshold point exists combinatorially such that only a relatively small increase in the proportion of relations transitions the network from the situation in which virtually no network member is connected via a path to any other network member to the situation in which most network members are connected via a path to most other network members (Erdos and Rényi 1960; Janson et al. 2002). This has been referred to as the "critical point" or "the double-jump threshold" (Molloy and Reed 1995, 1998) or the "phase transition."

In thermodynamics, where the concept of a phase transition originated, it refers to an abrupt change in physical properties resulting from a relatively small change in temperature. Readers will be familiar with the phase transition occurring when water entirely transforms from a crystalline solid (ice) to a liquid over a relatively small temperature threshold. As temperature rises, water molecules remain organized in a crystalline structure until, during a very short interval of degrees, they completely transform into a liquid form. A relationship, similar to that between temperature and molecular structure, exists between relational density and social network structure. If one imagines a sparse network with nodes existing only in small, disconnected components, as the density of relations increases, there would be no large-scale effects until a threshold point was achieved, when the addition of a relatively few relations transforms the population from many small, disconnected, insular communities into a network composed primarily of one dominant, comprehensive community whose constituent members are mutually reachable via paths.

Figure 1 illustrates this. It represents 10,000 randomly generated networks, each with 1,000 nodes and each with an average degree varying between zero and six. It plots the proportion of nodes in each network that is part of the largest component as a function of the average degree of the network. The three lines drawn on the graph represent three distinct theoretical slopes clearly evidenced in this plot: before the phase transition, during the phase transition, and after the phase transition. Clearly, as the average degree rises, so does the size of the largest component in the network. While the increase is rather gradual as the average degree rises to one, the slope changes spectacularly, increasing almost 20-fold, as the average degree goes to two and then, just as dramatically, returns to a slope similar to its previous one. The size of the largest component, which is gradually increasing along most of the continuum of increasing degree, suddenly "jumps" to a new threshold, one that it would not have achieved until the network was 20 times as dense as it currently is, if the phase transition had not occurred.



Sampling Effects in Social Network Analysis, Fig. 1 The **phase transition** simulated by 10,000 randomly generated networks, each with 1,000 nodes. *Lines* represent theoretical slopes: before, during, and after the phase transition



Sampling Effects in Social Network Analysis, Fig. 2 An illustration of **clustering**. Node *1* has four first neighbors, labeled *2* to *5*, and nine second neighbors, neighbors of neighbors, labeled *6* to *14*

Complications Unique to "Social" Networks

As a result of the extreme sensitivity of the phase transition, and other global network properties, to relatively trivial changes in local-network properties, great care must be exercised when using sample data to understand global-level network properties since virtually all samples are taken of local-network properties. For example, if we wanted to estimate whether the phase transition had occurred in a social network for which only sample data was available, there are known local-level properties, unique to social networks, which must be accounted for. Most important among these is clustering (illustrated in Fig. 2). In Fig. 2, node 1 connects to four others, and each of these also connects to four others (assume that the network continues on past the nodes labeled 6 and higher but that those edges simply are not illustrated here). While node 1 has four first neighbors, if we assumed that the number of second neighbors, neighbors of neighbors, would simply be a function of the number of first neighbors (i.e., that individuals only create or dissolve network ties based on their immediate interactions), we would expect node 1 to have 12 (4×3) second neighbors. We would expect that each node connected to node 1 would have the same average number of neighbors as node 1 (four) but one less due to the fact that each has already spent one of their four relations connecting to node 1. Instead, node 1 has only nine second neighbors. This occurs because nodes 2 and 3, both of which are first neighbors of node 1, each spend one of their relations connecting to each other and because nodes 4 and 5, both first neighbors of node 1, share a common second neighbor in node 12.

To account for the importance of clustering such as this on large-scale network properties, Grannis (2010) distinguished the number of first and second neighbors as two distinct variables, identifying the mean number of neighbors of a typical randomly chosen node as f_1 while letting f_2 represent the mean number of distinct second neighbors, regardless of how this number arises, whether influenced by transitivity or clustering or any other process that acts upon the distribution of second neighbors. The variable f_2 is measured independently of f_1. Because this variable, f_2,

ignores those edges that do not contribute to unique second neighbors, it therefore explicitly accounts for clustering (as well as the necessary connection with the original node). Thus, the ratio $g = (f_2/f_1)$ equals the proportional increase in the number of new neighbors. Thus, in the network illustrated in Fig. 2, $f_1 = 4$, $f_2 = 9$, and $g = 2.25$.

Using this notation, we expect the average node has f_1 first neighbors, $f_2 = f_1 g^1$ second neighbors, $f_3 = f_1 g^2$ third neighbors, and $f_m = f_1 g^{m-1}$ mth neighbors. The total number of neighbors reached in l (or fewer) steps is given by the geometric series

$$\sum_{m=1}^{l} f_m = \sum_{m=1}^{l} f_1 g^{m-1} = f_1 \frac{g^l - 1}{g - 1}$$

The expected size of the connected component to which a typical node belongs equals one (itself) plus the number of neighbors it could reach after an infinite number of steps. Substituting ∞ for l into the formula above and adding one yields

$$f_1 \frac{g^\infty - 1}{g - 1} + 1$$

If $g < 1$, g^∞ asymptotically approaches zero and the expression reduces to

$$1 + \frac{f_1}{1 - g}$$

If $g > 1$, the first term (and thus the entire expression) approaches infinity; the average component size is infinite (i.e., a giant component has formed). If, however, $g = 1$, then the first term becomes indeterminate, the phase transition point. A giant component exists when $g > 1$ and does not exist when $g < 1$.

Intuitively, one can understand this as follows. Assume that Jacob connects to Sophia, Ben, and Hannah. We can consider these individuals as the starting points on branches originating from Jacob. Regardless of Sophia's, Ben's, or Hannah's initial degrees, they must use one tie connecting to Jacob (else they would not be Jacob's neighbor), and they may use some (perhaps none, perhaps all) of their other ties (if any) connecting to each other (i.e., clustering). Any remaining ties will ramify out into new branches. Assume that, after connecting to Jacob and perhaps to each other, Sophia, Ben, and Hannah have zero, two, and one remaining ties, respectively, available to connect to new nodes. By connecting to Sophia, the number of branches originating from Jacob has decreased; by connecting to Ben, the number of branches originating from Jacob has increased; and, by connecting to Hannah, the number of branches originating from Jacob has stayed the same. In general, if the neighbors which any typical node is likely to connect to will, after accounting for ties spent in clustering and the initial connection, on average yield more than one new branch each, this process will expand throughout the network and a giant component can be expected to form.

Illustrative Example: Sampling Core Discussion Networks from the General Social Survey

To illustrate, consider the General Social Survey (GSS) data on the confidants with whom Americans discuss important matters. Examination of trends in GSS discussion networks (which were collected in 1985, 1987, and 2004) at the individual level have reported important changes in the last generation. For example, McPherson et al. (2006, p. 353) noted that individual networks are a third smaller in 2004 than in 1985 (about two people instead of three) and that the number of people saying there is no one with whom they discuss important matters nearly tripled.

To estimate whether the phase transition has occurred and a giant component exists, we need to calculate the proportional increase in the number of new neighbors,

$$g = \frac{f_2}{f_1} (f_1 \neq 0)$$

If $g > 1$, then the phase transition has occurred. To perform this calculation, we need to know the number of a node's neighbors, f_1, as well

Sampling Effects in Social Network Analysis, Table 1 Values for individual-level predictors of the phase transition

		1985		2004	
		Preferential attachment	Random	Preferential attachment	Random
Proportional increase in new neighbors (g)	No clustering	2.975 (0.06935)	2.975 (0.05055)	2.565 (0.09166)	2.565 (0.09771)
	Especially close	1.815 (0.04975)	1.894 (0.03968)	1.505 (0.06020)	1.592 (0.06161)
	Neither especially close nor strangers	0.6858 (0.02996)	0.8277 (0.03326)	0.5046 (0.03302)	0.6818 (0.03902)
Number of respondents		1,525		1,466	
Average number of first neighbors (f_1)		2.980 (0.4409)		1.987 (0.04409)	

as the number of distinct second neighbors that are not also first neighbors, f_2. The GSS data provides information on the first variable, f_1, the number of others each individual nominates as someone with whom they discuss important matters. Data concerning the second variable, f_2, however, is not readily apparent and requires some calculation.

The simplest, although probably not the most accurate, way to do this would be to assume that one's discussion partners do not discuss important matters with each other (i.e., that there is no clustering) but, rather, that they link to others randomly with the only stipulation being that each discussion partner has used one tie connecting to the respondent; all other ties extend outward. Table 1 displays results generated using this model. It shows that the value of g to be 2.975 in 1985 and 2.565 in 2004.

Alternatively, one could theorize that some of an individual's discussion partners also discuss important matters with each other. The GSS did not ask respondents which of the people they knew, whom they discussed important matters with, also discussed important matters with each other. The GIS did, however, ask respondents to characterize the relationship between each pair of the people they mentioned into three categories: as "especially close, as close or closer" than their relationship to the respondent; "total strangers"; or somewhere in-between. Which, if any, of these corresponds to discussing important matters is unknown.

If we theorize that all pairs of individuals identified as "especially close" are, in fact, discussion partners, then members of pairs so identified each spend one of their ties connecting to the other. Thus, fewer ties will extend outward to others. Table 1 shows that, under this model, the value of g is 1.894 in 1985 and 1.592 in 2004. For the "no clustering" model as well as the model of those who were "especially close" as discussion partners, $g > 1$ indicates that a giant component clearly unites most isolates.

Some might assume that only the more exclusive "especially close" relation represents those who would have in fact nominated each other as someone they discuss important matters with if the GSS had surveyed them. However, while it seems reasonable to assume that not all pairs of individuals whom respondents categorized as neither "especially close" nor "total strangers" would have nominated each other as someone they discuss important matters with, it is certainly arguable that some of them might have, given that this intermediate category implicitly included those "almost as close." Thus, if we further theorize that not only those who are "especially close" but also those in the intermediate category, neither "especially close" nor "strangers," are also discussion partners, then an even larger number of pairs of those directly connected to the respondent will spend ties connecting to each other. Table 1 shows that, using this model, the value of g is 0.8277 in 1985 and 0.6818 in 2004. If this model is correct, $g < 1$ tells us that in

neither year has the phase transition occurred and all components are minuscule.

This definitional distinction has dramatic effects when one considers the network model it generates. The difference between these two scenarios is not merely that one component is somewhat larger, but rather it is a difference in orders of magnitude. Theoretically, it signals the difference between a society that is primarily united into a single discussion network and a society that has utterly disintegrated.

In the first case, it is possible that the typical person is involved in an extended discussion network (e.g., she discusses important matters with someone who discusses important matters with someone who discusses important matters with someone, etc.) that ultimately includes multiplied millions of people. While it is unlikely that the specifics of one's discussions transmit over any distance, it is possible that general norms or values could diffuse and a general awareness, if not consensus, could form.

The second case is quite distinct. To understand just how tiny the nonphased components are, we can use the formula derived above for calculating average component size when $g < 1$:

$$1 + \frac{f_1}{1 - g}$$

We find that, in this case, the size of the average discussion component is 18 in 1985 and 7 in 2004. Thus, in this case, most persons' discussion networks do not extend much beyond those they have direct discussions with. Instead of society consisting of an extended network diffusing norms and values, it would have been pulverized into tiny groups, perhaps no larger than a single individual's discussion network.

Summary

This entry has reviewed some of the many unique issues which arise when one uses sample data to model social networks. In some cases, such as with chain-referral sampling, social network conceptualizations may prove advantageous, as statistical methods have been created which allow researchers to translate otherwise questionable data into robust estimates of population parameters. In contrast, other cases demonstrate that when using conventional sample data to understand network processes, great care must be taken in how one theorizes, defines, and operationalizes local-level processes. Social actors, unlike nonsocial network nodes, are aware of and respond to the actions or inactions of both their first neighbors and their second neighbors (Friedkin 1983). Relatively trivial variations in the social responses by these individual actors may have dramatic effects on the theoretical understanding which results from analyzing the sampled data.

Cross-References

- ▸ Network Representations of Complex Data
- ▸ Probabilistic Analysis
- ▸ Probabilistic Graphical Models
- ▸ Research Designs for Social Network Analysis

References

Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. Science 323:892–895

Brewer KRW, Hanif M (1983) Sampling with unequal probability. Springer, New York

Butts CT (2006) Exact bounds for degree centralization. Soc Netw 28:283–296

Butts CT (2009) Revisiting the foundations of network analysis. Science 325:414–416

Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York

Coleman JS (1958) Relational analysis: the study of social organization with survey methods. Hum Organ 17: 28–36

Eland-Goosensen M, Van De Goor L, Vollemans E, Hendriks V, Garretsen H (1997) Snowball sampling applied to opiate addicts outside the treatment system. Addict Res 5(4):317–330

Erdos P, Rényi A (1960) On the evolution of random graphs. Publ Math Inst Hung Acad Sci 5:17–61

Erickson BH (1979) Some problems of inference from chain data. In: Karl FS (ed) Sociological methodology, vol 10. Jossey-Bass, San Francisco, pp 276–302
Frank O (1978) Sampling and estimation in large social networks. Soc Netw 1:91–101
Friedkin NE (1983) Horizons of observability and limits of informal control in organizations. Soc Forces 62:54–77
Grannis R (2010) Six degrees of 'who cares?'. Am J Sociol 115:991–1017
Grannis R, Freedy E, Freedy A (2011) Ultra-rapid social network sampling in cross-cultural environments. In: HSCB Conference proceedings: Integrating social science theory and analytic methods for operational use, Arlington
Granovetter M (1976) Network sampling: some first steps. Am J Sociol 81(6):1287–1303
Heckathorn DD (1997) Respondent-driven sampling: a new approach to the study of hidden populations. Soc Probl 44(2):174–199
Heckathorn DD (2002) Respondent-driven sampling II: deriving valid population estimates from chain-referral samples of hidden populations. Soc Probl 49(1):11–34
Janson S, Luczak T, Rucinski A (2002) The phase transition, Chap. 5. In: Random graphs. Wiley, New York, pp 103–138
Kalton G (1983) Introduction to survey sampling. Sage, Beverly Hills
Kalton G, Anderson DW (1986) Sampling rare populations. J R Stat Soc Ser A 149:65–82
McPherson M, Smith-Lovin L, Brashears ME (2006) Social isolation in America: changes in core discussion networks over two decades. Am Sociol Rev 71: 353–375
Molloy M, Reed B (1995) A critical point for random graphs with a given degree sequence. Random Struct Algorithms 6:161–179
Molloy M, Reed B (1998) The size of the giant component of a random graph with a given degree sequence. Comb Probab Comput 7:295–305
Salganik MJ, Heckathorn DD (2004) Sampling and estimation in hidden populations using respondent-driven sampling. Sociol Methodol 34:193–239
Schelling TC (1978) Micromotives and macrobehavior. Norton, New York

Recommended Reading

Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. Science 323:892–895
Doreian P, Woodard KL (1992) Fixed list versus snowball selection of social networks. Soc Sci Res 21:216–233
Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256
Newman MEJ (2010) Networks: an introduction. Oxford University Press, New York

Scalable Distributed Systems

▸ Cloud Computing

Scale-Free Distributions

▸ Scale-Free Nature of Social Networks

Scale-Free Nature of Social Networks

Piotr Fronczak
Faculty of Physics, Warsaw University of Technology, Warsaw, Poland

Synonyms

Complex networks; Network models; Scale-free distributions; Universal scaling

Glossary

Degree The degree of a node in a network is the number of edges or connections to that node
Node Degree Distribution The distribution function $P(k)$ that gives the probability that a node selected at random has exactly$k$ edges
Power-Law Distribution Has a probability function of the form $P(x) \sim x^{-a}$
Fat-Tailed Distributions Have tails that decay more slowly than exponentially. All power-law distributions are fat tailed, but not all fat-tailed distributions are power laws (e.g., the log-normal distribution is fat tailed but is not a power-law distribution)
SF Network The network with power-law distribution of node degrees
ER Graph The network model in which edges are set between nodes with equal probabilities
Scale-Freeness Feature of objects or laws that does not change if length scale is multiplied by a common factor, also known as scale invariance

Definition

The notion of scale-freeness and its prevalence in both natural and artificial networks have recently attracted much attention. In physics and mathematics, scale-freeness (or more formally – scale invariance) is a feature of objects or laws that does not change if length scale is multiplied by a common factor. The term gained large popularity in 1999 when Barabasi and Albert used it as a descriptor of networks in which node degrees (vertex connectivity) follow a power-law distribution (Barabasi and Albert 1999). Since the most large complex networks are characterized by the distributions which at least partially are reminiscent of power functions, the term "scale-free," applied to networks, losts its formal meaning and nowadays is widely used (albeit erroneously) to describe the network with fat-tailed node degree distribution.

The overwhelming number of studies conducted in the last decade made it clear that the scale-free network topology can have a strong impact on the dynamical processes taking place on these networks such as opinion formation (Aleksiejuk 2002), diffusion of information (Cohen et al. 2000), and epidemic spreading (Pastor-Satorras and Vespignani 2001). Nowadays, the recently acquired knowledge about the network structure revolutionizes not only many fields of science, like biology, computer science, and economics, but also the society and its perception of the ubiquitous networks.

Introduction

Scale-freeness is the property which is fascinating especially for physicists, since most phenomena studied by physicists are not scale invariant. Among seminal exceptions are phase transitions in thermodynamic systems which are associated with the emergence of power-law distributions of certain quantities (Yeomans 2002). Similarly, the phenomenon known as self-organized criticality (a property of dynamical systems which have a critical point as an attractor) displays the spatial and/or temporal scale-free nature of the critical point of a phase transition, but without the need to tune control parameters to precise values (Bak 1996).

In mathematics, scale invariance is an exact form of self-similarity where at any magnification, there is a smaller piece of the object that is similar to the whole. Self-similarity is a typical property of fractals.

A common aspect of both phase transitions and self-similar fractals is a universality, i.e., the observation that there are properties for a large class of different systems that are independent of the dynamical details of the particular system.

These reasons (universality and criticality) explain the excitement of scientists, when the power-law character of node degree distribution has been observed in drastically increasing number of real networks. The promise of the discovery of the universal character of surrounding us social, technological, and natural networks made the notion of scale-freeness frequently misused. Nevertheless, it is a notion that has clearly taken root with today's society effectively guiding the communicative patterns of different scientific communities. In the following paragraphs, we will use this notion in its less formal meaning as a short cut of the networks with fat-tailed (or almost power law) distribution of node degrees.

Despite the pure mathematical differences, the properties of idealized (scale-free) and realistic (almost scale-free) networks have the same implications for real-world applications.

Key Points

To understand the scale-free architecture of the networks, it is useful to contrast it with the other network model which dominated the network research for decades, namely, the model of network developed by Erdos and Renyi in 1959 (ER graph) (Erdos and Renyi 1960). The importance of the ER graph for modelling real-world networks is currently diminished; however, it is still fundamental model in the random network theory. In the following, we will briefly

introduce ER graphs and emphasize differences between them and SF networks. We will present the methods of detection of the scale-free character of the node degree distribution in networks. We will discuss the most popular model in which the growing network evolves into scale-free state. Finally, we will discuss the vulnerability of SF networks to epidemics and intentional attacks and their extreme tolerance on random failures.

Historical Background

Power-law distributions in nature and society were already known in the nineteenth century. Italian economist, Vilfredo Pareto, in 1897, was the first to discover that the distribution of income in society follows the power law (Barabasi 2002). In 1925, George Udny Yule proposed a stochastic process (later called the Yule process, but is now better known as preferential attachment – see the next section) that leads to a distribution with a power-law tail – in this case, the distribution of species and genera (Yule 1925). In 1965, Derek John de Solla Price demonstrated a power-law distribution of links in a network of scientists linked by citation de Solla Price (1965). Although D. Price was a physicist, his discovery was totally ignored in physical sciences. In physics, the lattices and random networks like ER graph were the main objects of study until the late 1990s, when Barabasi and others rediscovered the importance of SF networks in technology, nature, and society.

Properties of Scale-Free Networks

Two Opposing Models of Random Networks

The definition of ER graph is simple: in a graph with N nodes, each possible pair of distinct nodes is connected with an edge with probability p. In that model, the majority of nodes have a degree that is close to the average degree of the whole network, and this average has small variance (the number of nodes with a given degree decays exponentially fast away from the mean degree). In Fig. 1a, we show a typical representative of this model. As one can see, the sizes of all nodes (which reflect node degrees) are similar. For large N and infinitesimal p (i.e., for large and sparse networks), the node degree distribution follows a Poisson law

$$P(k) = e^{-pN} \frac{(pN)^k}{k!},$$

where k is a node degree and the average node degree $\langle k \rangle = pN$ (Newman et al. 2002). The characteristic bell shape of $P(k)$ around the average node degree is visible in Fig. 2a.

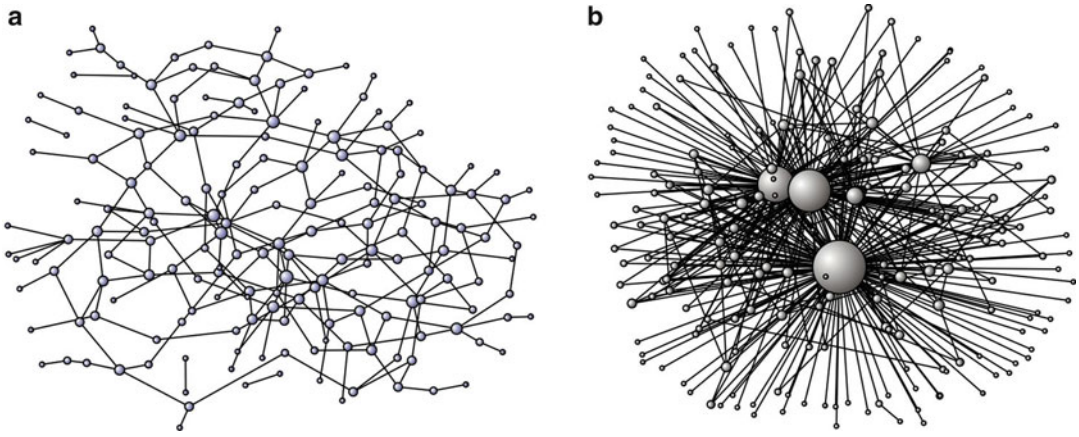
As we stated previously, recent studies show that most large complex networks are characterized by a connectivity distribution different to a Poisson distribution (among the exceptions are train networks or electrical power grids). For example, the World Wide Web, Internet, e-mail, and collaboration networks have a degree distribution that follows (at least in some range) a power-law relationship defined by

$$P(k) \sim k^{-\gamma},$$

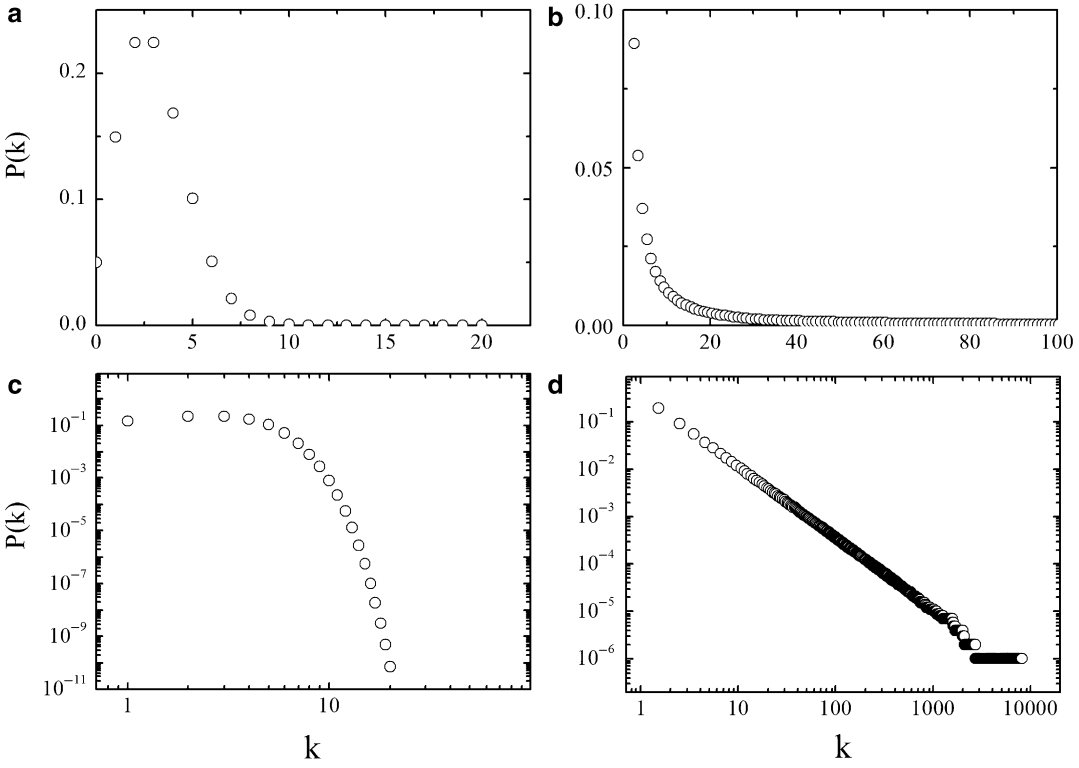
where γ, called *scale-free exponent*, ranges usually between 2 and 3 in real networks. Such networks have a very uneven distribution of connections. There are many nodes with only a few links and a few nodes with a large number of links. The difference between this type of network and a Poisson-like one is clearly visible in Fig. 1b, where some nodes act as "highly connected hubs" while the most of them has only one connection. The fat tail of this distribution, shown in Fig. 2b, is an evidence of an extreme heterogeneity of connections in the network.

Scale-Freeness of Networks with Power Law Distribution of Node Degrees

Why the networks with the power-law distributions of node degrees are called scale-free? It is because a power-law distribution is scale invariant. If we rescale a measure of connectivity



Scale-Free Nature of Social Networks, Fig. 1 Two realizations of an ER graph (**a**) and SF network (**b**) both with the same number of nodes and edges. Size of the nodes is proportional to their degrees



Scale-Free Nature of Social Networks, Fig. 2 Node degree distributions of ER graph (*left column*) and SF network (*right column*) in normal (*top row*) and double logarithmic (*bottom row*) scale

(e.g., counting how many tens of connections a node has, instead of counting all its connections), the connectivity distribution $P(10k)$ will be still proportional to the original $P(k)$. Mathematically, multiplying degree k by a constant c, the distribution remains the same and only scales the function $P(ck) = c^{\gamma}P(k)$, where $P(k) = ck^{-\gamma}$. To show that power-law

distribution is the only one, which fulfils this condition, we take the logarithm of its both sides:

$$\ln P(ck) = -\gamma \ln c + \ln P(k).$$

Now, we introduce a new function $R(k)$ defined as $R(\ln k) = P(k)$. This gives

$$\ln R(\ln ck) = -\gamma \ln c + \ln R(\ln k)$$

and, after rearrangement,

$$\frac{\ln R(\ln c + \ln k) - \ln R(\ln k)}{\ln c} = -\gamma.$$

In the limit $\ln c \to 0$, the left side becomes a derivative

$$\frac{d \ln R(\ln k)}{d \ln k} = -\gamma.$$

Since the right side is constant, integrating the equation gives

$$\ln R(\ln k) = -\gamma \ln k + \text{const}$$

and finally

$$R(\ln k) = P(k) = \text{const} \cdot k^{-\gamma}.$$

An important difference between fat-tailed and Poisson-like distributions is that moments of the former (i.e., mean $\langle k \rangle$ and variance $\delta^2(k)$) poorly characterize the distribution (in fact, they are undefined for certain power-law distributions). The moments μ_m of order m are defined as follows:

$$\mu_m = \sum_{k=0}^{k_{\max}} k^m P(k).$$

From the definition, in infinite networks (i.e., when $k_{\max} \to \infty$), all higher moments of order $m \geq \gamma - 1$ of the power-law distribution diverge. Since a mean and a variance are the moments of the first and the second order, respectively, a variance is infinite for γ in the range of typical real-world networks ($2 \leq \gamma \leq 3$):

$$\delta^2 = \mu_2 \sim \sum_{k=0}^{\infty} k^2 k^{-\gamma} = \infty, \text{ for } \gamma \leq 3.$$

Although the real networks are finite, the variance can be still several orders larger than the mean. Since a variance describes the error of measured mean node degree, its enormously large value questions the quality of the measurement and assigning the scale (related to the mean degree) to the network is a misuse.

Plotting Scale-Free Distributions

Since many fat-tailed distributions look similarly as in Fig. 2b (e.g., log-normal or stretched exponential distributions), to better expose the power-law nature of the node degree distribution, one usually plots the data on a double logarithmic scale. In that case, the power law transforms into a straight line with a slope of $-\gamma$ (see Fig. 2d and compare it with Poisson distribution shown in Fig. 2c), as follows:

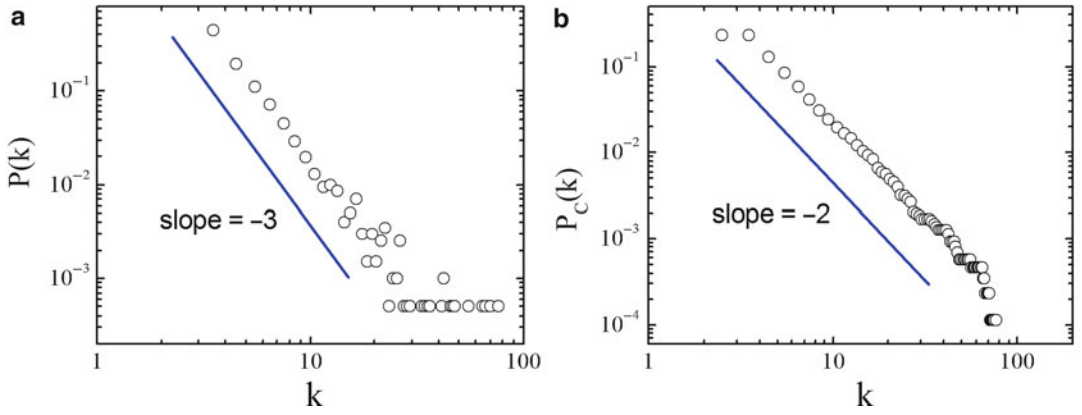
$$\begin{aligned} P(k) &= a \cdot k^{-\gamma} \\ \ln P(k) &= \ln(a \cdot k^{-\gamma}) \\ \ln P(k) &= \ln a + \ln(k^{-\gamma}) \\ \ln P(k) &= \ln a - \gamma \ln k \\ Y &= A - \gamma X, \end{aligned}$$

where X and Y are transformed variables and A is a transformed constant.

In practice, measuring a slope directly from Fig. 2d is usually very erroneous, due to the poor statistics at the tail of the distribution. Direct histograms are almost always noisy in this region. The solution is to construct a histogram in which the bin sizes increase exponentially with degree. The number of samples in each bin is then divided by the width of the bin to normalize the measurement. Plotting histogram in a logarithmic degree scale, one obtains the even widths of the bins.

An even more discriminating method to verify potential power-law character of the node degree distribution is to plot the complementary cumulative distribution function

$$P_C(k) = \int_k^{\infty} P(k) dk \sim k^{-(\gamma+1)}$$



Scale-Free Nature of Social Networks, Fig. 3 Node degree distribution of SF network (**a**) and its cumulative distribution (**b**)

which is the probability that the degree of a randomly chosen node is greater than or equal to k. Such a plot has the advantage that all the original data are represented. When we make a conventional histogram by binning, any differences between the values of data points that fall in the same bin are lost. The cumulative distribution function does not suffer from this problem. The cumulative distribution also reduces the noise in the tail, what is clearly illustrated in Fig. 3.

Seminal Model of Preferential Attachment

Soon after the discovery of the scale-free structure of the World Wide Web, it has been realized that many other real networks also show power-law distribution of node degrees. This feature has been observed in the Internet, communication, and transportation networks (Albert et al. 1999; Guimerà et al. 2005), as well as in many social networks, such as networks of scientific citations Redner (1998), e-mail networks (Ebel et al. 2002), or even sexual contact networks (Liljeros et al. 2001). The initial surprise of omnipresence of SF networks quickly turned into a question: Why so many networks have the same scale-free character of connections? When a feature appears in many systems that do not have an obvious connection to each other, you should suspect that there is a common causal principle, which can be described in the most general terms, without reference to the details of this or any other system. Is a scale invariance of complex networks a result of some universal rules that govern the dynamics of these systems?

Although there are many different processes which can give rise to the same power-law structure of complex networks, the one deserves particular attention at least for the two reasons. Firstly, its universal character allows to adapt the process to many social but also technological and natural networks. Secondly, it has been independently rediscovered several times in different fields and ages. The process is currently known as Matthew effect, Yule process, Dulbecco's law, rich gets richer, or preferential attachment (Barabasi and Albert 1999). Since it is used so widely across domains, the claim about its universality is reasonable.

The process, adopted to networks, comprises of two complementary mechanisms: network growth and preferential rules of joining nodes. Barabasi and Albert, who introduced the process to the modern science of complex networks, stated that real networks are not formed as a result of purely random process, in which a completely randomly selected nodes are connected by the edges. The most of the social and technological networks grow and change over time – they evolve. In networks, the newly added nodes prefer to create connections with such ones that already have a lot of other connections. The mechanisms

underlying this preference can be different. For example, new actors are more likely to play supporting roles in films with established stars, than in those where there are only other unknown actors. Thus, the more famous you are, the more probably is that you will attract new connections. The same principle seems to govern the structure of citation network. Preferential attachment corresponds to the feature that a publication with a large number of citations continues to be well cited in the future merely by virtue of being well-cited now. In the network of acquaintances, my friends introduce me to their friends. The more friends I have, the more recognized I am and the more chances to meet new people I have. In WWW, the more pages link to a web page, the more Internet users visit that site and the greater the likelihood that they will place a link to this page on their own website.

The algorithm of the discussed process consists of two steps:

1. Starting with a small number m_0 of nodes, at every time step, we add a new node with $m \leq m_0$ edges that link the new node to m different nodes already present in the system.
2. When choosing the nodes to which the new node connects, we assume that the probability $\prod$ that a new node will be connected to node i depends on the degree k_i of node i, such that

$$\prod(k_i) = \frac{k_i}{\sum_j k_j}.$$

After t time steps, this algorithm results in a network with $N = t + m_0$ nodes and mt edges. In Fig. 4, first steps of the network evolution have been shown. Already after several steps, the hubs in the network become clearly visible.

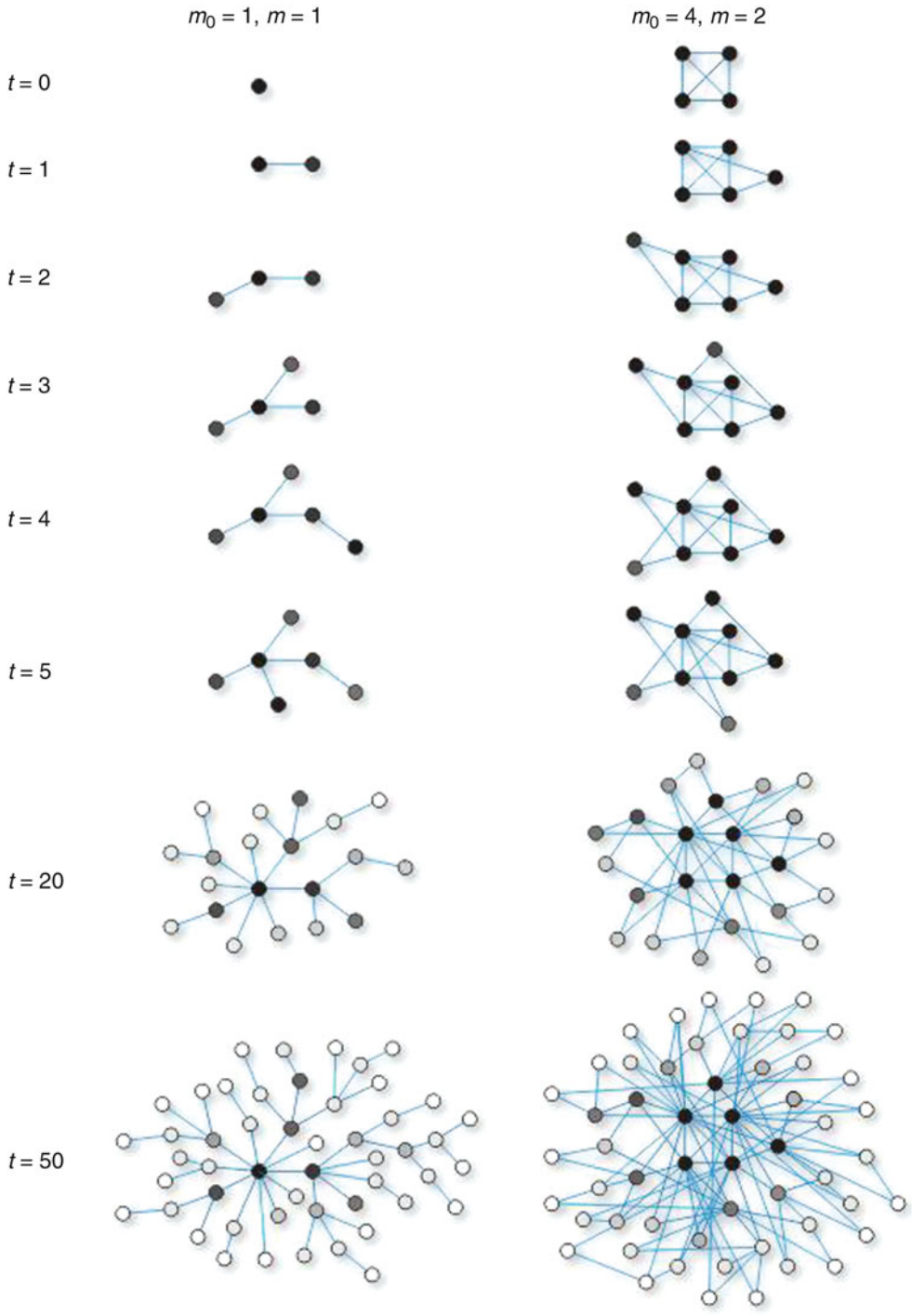
Mathematical derivations show that the node degree distribution of the network evolves into a scale-free one with the scale-free exponent $\gamma = 3$ independently of m, the only parameter in the model.

One has to keep in mind that the presented model does not share all properties observed in the real-world networks, e.g., it is less clustered. Soon, after this model was introduced, a large number of similar models, all based on some type of connecting preference, emerged, all leading to a power-law distribution of node degrees but also demonstrating a better agreement with real networks with reference to other network metrics.

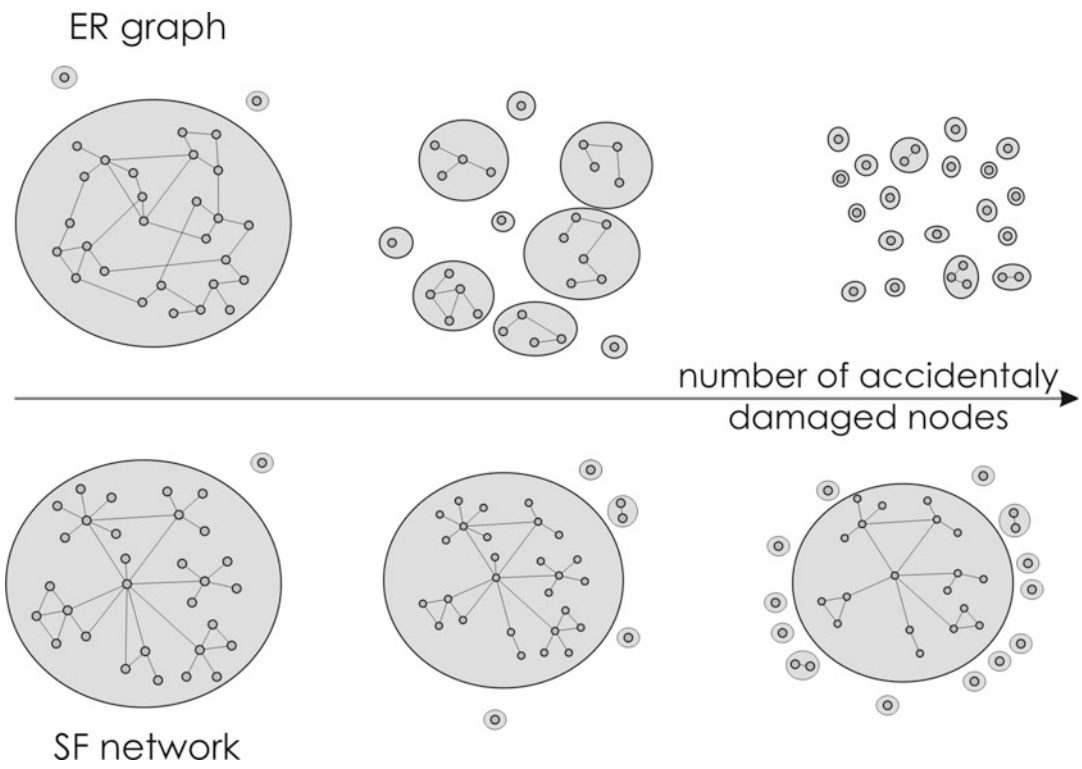
Preferential attachment is not the only possible explanation for the formation of scale-free structure of connections in complex networks. Among others, there are rewiring processes (Aiello et al. 2002), optimization-based models (Valverde et al. 2002), and also static constructions (Park and Newman 2004; Goh et al. 2001).

Resilience and Vulnerability of SF Networks to Failures and Attacks

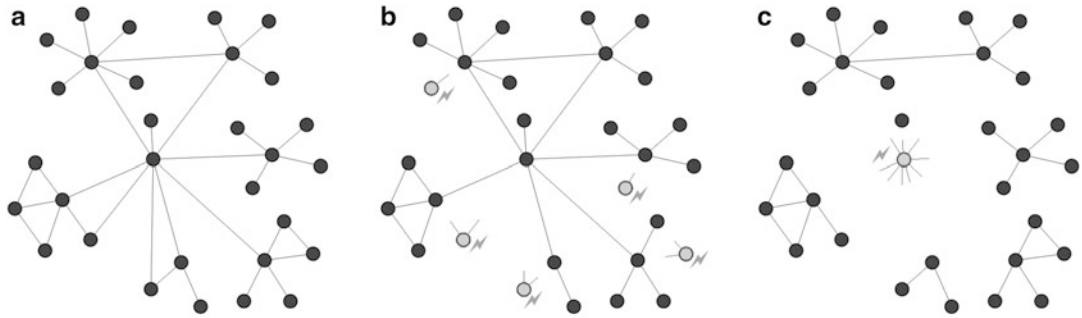
It has appeared recently that scale-free structure of complex networks has an important influence on their resilience to failures and attacks. In particular, SF networks seem much more robust than ER graphs in case of failures (modelled by a random removal of nodes or links) (Cohen et al. 2000), while they are more sensitive to attacks (modelled by the targeted removal of selected nodes or links) (Cohen et al. 2001). By the resilience, we understood that despite removed nodes, the main part of the network (so-called giant component) is still interconnected (i.e., any two nodes in that part are connected to each other by paths). If the node elimination proceeds, then at some critical moment, the network breaks apart into small disconnected parts. The moment when this dramatic breakdown occurs strongly depends on the network structure as well as on the method of node's elimination (random or targeted). In the random removal case, the critical moment of destruction occurs much earlier in ER graphs, in opposite to SF networks (see Fig. 5). It means that SF network is much more resilient to accidental damages. However, in case of intentional attack, when the nodes of the network are removed in decreasing order of their degree, SF network appears to be much more vulnerable than ER graph (since the removal of the hubs results in the largest possible damage, see Fig. 6). This vulnerability of SF networks to intentional attacks has been described as their Achilles' heel.



Scale-Free Nature of Social Networks, Fig. 4 Example of realization of two different growing networks in preferential attachment model. The *colors* of the nodes represent their age



Scale-Free Nature of Social Networks, Fig. 5 ER graphs break apart into small disconnected parts much faster that SF networks if the nodes are removed accidentally



Scale-Free Nature of Social Networks, Fig. 6 Random and targeted elimination of nodes. Original SF network (**a**), the network with randomly damaged nodes (**b**), and the network with the damaged hub

Epidemic Spreading in SF Networks

Scale-free nature of social networks has a great implication for understanding the spread of information, diseases, opinions, and innovations in society. Standard epidemiological models usually consider networks with the well-defined average node degree, such as ER graphs. In those networks, the models predict a critical threshold for the propagation of a contagion throughout a population. This epidemic threshold is determined by the virulence of the infection. In other words, if the spreading rate is larger than the threshold, the infection spreads and becomes persistent. Below the threshold, the infection dies out.

It turns out that in SF networks, the above statement is no longer correct. In 2001, Pastor-Satorras and Vespignani found that in that case, the threshold is zero (Pastor-Satorras and Vespignani 2001). It means, that all viruses, even those that are weakly contagious, will spread and persist in the system. The main reason is that the presence of hub nodes can facilitate epidemic spreading due to the large numbers of neighbors. Infected hub passes the infection to numerous other nodes, faster than the typical node recovers.

Specifically, in SF network, the traditional random immunization could easily fail because nearly everyone would have to be treated to ensure that the hubs were not missed. New immunization strategies have to be developed to recover the epidemic threshold. It turns out that one of the most efficient approaches is to selectively immunize hub nodes. Such a strategy is known as targeted immunization (Pastor-Satorras and Vespignani 2002).

Future Directions

The structure, topological properties, and appropriate measures were the main research topics in complex networks domain in recent years. Currently, dynamical processes taking place in the networks are quite intensively studied. It is believed that further understanding of dynamics on complex networks is the general direction of the field. There is a continuous shift from studies of networks in general and features that are common to most of them to more application-driven studies of increasingly narrow classes of networks. After a decade of mostly descriptive studies and just potential applications, there is a final need to transfer an acquired knowledge into concrete market applications. Complex networks research society should provide the manageable solutions to global challenges, like vaccination campaigns against serious viruses, risk reduction of financial crises, and preventing cascading bankruptcies among interlinked economies.

Cross-References

▸ Exponential Random Graph Models
▸ Network Models

References

Aiello W, Chung F, Lu L (2002) Random evolution of massive graphs. In: Pardalos PM, Abello J, Resende MGC (eds) Handbook of massive data sets. Kluwer Academic, Dordrecht/London, pp 97–122

Albert R, Jeong H, Barabasi A-L (1999) Diameter of the World Wide Web. Nature 401:130–131

Aleksiejuk A, Holyst JA, Stauffer D (2002) Ferromagnetic phase transition in Barabasi-Albert networks. Physica A 310:260–266

Bak P (1996) How nature works: the science of self-organized criticality. Copernicus, New York

Barabasi A-L (2002) Linked: the new science of networks. Perseus Books Group, Cambridge

Barabasi A-L, Albert R (1999) Emergence of scaling in random networks. Science 286:509–512

Cohen R, Erez K, ben Avraham D, Havlin S (2000) Resilience of the Internet to random breakdowns. Phys Rev Lett 85:4626

Cohen R, Erez K, ben Avraham D, Havlin S (2001) Breakdown of the Internet under intentional attack. Phys Rev Lett 86:3682

de Solla Price DJ (1965) Networks of scientific papers. Science 149:510–515

Ebel H, Mielsch LI, Bornholdt S (2002) Scale-free topology of e-mail networks. Phys Rev E 66:035103

Erdos P, Renyi A (1960) On the evolution of random graphs. Publ Math Inst Hung Acad Sci 5: 17–61

Goh K-I, Kahng B, Kim D (2001) Universal behaviour of load distribution in scale-free networks. Phys Rev Lett 87:278701

Guimerà R, Mossa S, Turtschi A, Amaral LAN (2005) The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles. Proc Natl Acad Sci 102:7794–7799

Liljeros F, Edling CR, Amaral LAN, Stanley HE, Aberg Y (2001) The web of human sexual contacts. Nature 411:907–908

Newman ME, Watts DJ, Strogatz SH (2002) Random graph models of social networks. Proc Natl Acad Sci USA 99(Suppl 1):2566–2572

Park J, Newman MEJ (2004) Statistical mechanics of networks. Phys Rev E 70:066117

Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale-free networks. Phys Rev Lett 86: 3200

Pastor-Satorras R, Vespignani A (2002) Immunization of complex networks. Phys Rev E 65:036104

Redner S (1998) How popular is your paper? An empirical study of the citation distribution. Eur Phys J B 4: 131–134

Valverde S, Ferrer Cancho R, Solé RV (2002) Scale-free networks from optimal design. Europhys Lett 60:512

Yeomans JM (2002) Statistical mechanics of phase transitions. Oxford University Press, New York

Yule UG (1925) A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Philos Trans R Soc Lond B Contain Pap Biol Character 213:21–87

Scaling Subgraph Matching Queries in Huge Networks

Matthias Bröcheler[1], Andrea Pugliese[2], and V. S. Subrahmanian[1]
[1]Computer Science Department, University of Maryland, College Park, MD, USA
[2]DIMES Department, University of Calabria, Rende, Italy

Synonyms

Graph matching; Subgraph identification; Subgraph isomorphic queries

Glossary

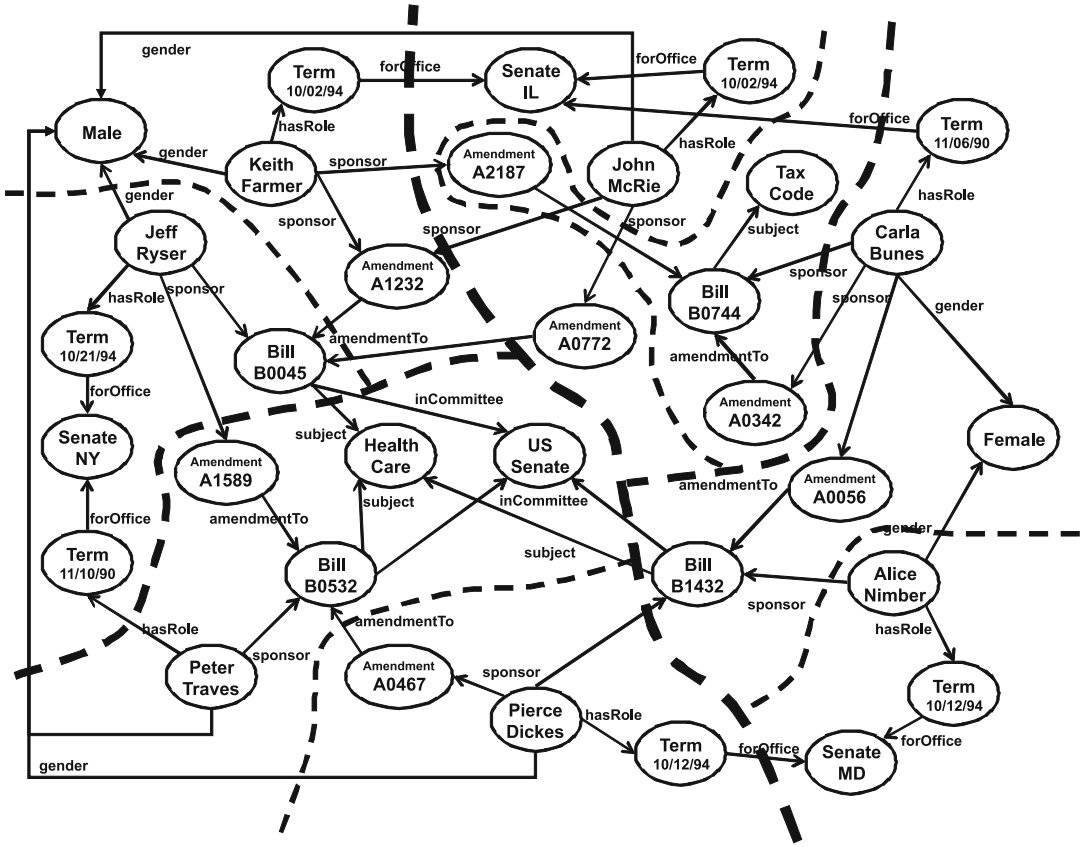
DOGMA Disk-Oriented Graph Matching Algorithm
COSI Cloud-Oriented Subgraph Identification
RDF Resource Description Framework
SPARQL SPARQL Protocol and RDF Query Language

Introduction

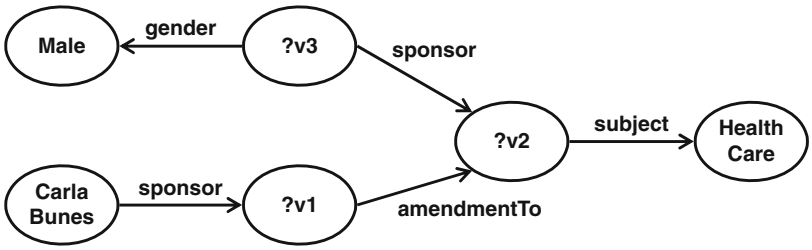
Both social network owners and social network users are interested in a variety of queries that involve subgraph matching. In addition, answering SPARQL queries in the Semantic Web's RDF framework largely involves subgraph matching. For example, the GovTrack dataset (2013) tracks events in the US Congress. In Fig. 1, we see that Jeff Ryster sponsored Bill B0045 whose subject is Health Care. A user who is using such a database might wish to ask queries such as that shown in Fig. 2. This query asks for all amendments ($?v1$) sponsored by Carla Bunes to bill ($?v2$) on the subject of health care that were originally sponsored by a male person ($?v3$). The reader can readily see that when answering this query, we want to find *all* matches for this query graph in the original graph. The reader who tries to answer this very simple query against this very tiny graph will see that it takes time to do so, even for a human being! In this entry, we show how to answer complex subgraph matching queries over huge graphs efficiently.

An important aspect of all past work is that it has focused on working in memory. Though memory prices are dropping while capacity is increasing, the increase in capacity is not even remotely large enough to store 1 % of Facebook or Twitter. As a consequence, in order to efficiently answer queries to social network graphs, we are forced to store the data, as well as indexes for the data, on disk. In section "The DOGMA Index" we provide a description of the DOGMA (Disk-Oriented Graph Matching Algorithm) index (Bröcheler et al. 2009) for building a disk-based index for huge networks. DOGMA is based on a simple observation: the size of any real-world social network graph is likely to be orders of magnitude larger than that of any subgraph matching query graph a user is likely to ask. This tells us that it should be possible to build an index for efficiently executing such queries that ensures that vertices in a social network graph that are "near" each other be stored together on a disk page.

Then, in section "Cloud-Oriented Subgraph Matching", we present the COSI (Cloud-Oriented Subgraph Identification) system (Bröcheler et al. 2010). COSI distributes a graph across multiple compute machines and answers subgraph matching queries in parallel using an asynchronous query-answering algorithm that does not rely on central orchestration. Thus, computation is completely distributed and our



Scaling Subgraph Matching Queries in Huge Networks, Fig. 1 Example graph



Scaling Subgraph Matching Queries in Huge Networks, Fig. 2 Example query

goal is to minimize communication between compute machines so as to save time. Finally, section "Experimental Results" shows some experimental results assessing the performance of DOGMA and COSI, and section "Related Work and Conclusions" briefly discusses related work and outlines conclusions.

Basic Notation

Throughout this entry, we assume the existence of an arbitrary but fixed set V whose elements are called vertices. For example, V might consist of all strings that can form a valid user ID and/or the set of all valid identifiers for comments in a social

network like Facebook. We also assume the existence of a finite set $\mathcal{P}$ of *predicate symbols*.

We model a social network graph $\mathcal{S}$ as a triple (V, E, λ) where V is the set of vertices, $E \subseteq V \times V$ is a *multiset* of edges from vertices to vertices, and $\lambda : E \rightarrow \mathcal{P}$ assigns a predicate symbol to each edge in E.

The *out neighborhood* of vertex v is the set $\text{out}(v) = \{u \mid (v, u) \in E\}$; the *in neighborhood* of node v is the set $\text{in}(v) = \{u \mid (u, v) \in E\}$. The neighborhood of v is the set $\text{ngh}(v) = \text{out}(v) \cup \text{in}(v)$. Each of these neighborhoods can be restricted to a particular predicate symbol p: for example, $\text{out}_p(v) = \{u \mid (v, u) \in E \wedge \lambda(v, u) = p\}$.

When formulating queries, we assume the existence of a set VAR of variable symbols ranging over V. Each variable symbol starts with a ?. A *query* Q is a triple (V_Q, E_Q, λ_Q) where $V_Q \subseteq V \cup \text{VAR}$, $E_Q \subseteq V_Q \times V_Q$ is a multiset of edges, and $\lambda_Q : E_Q \rightarrow \mathcal{P}$. We use VAR_Q to denote the set of variable vertices in query Q.

Suppose $\mathcal{S}$ is a social network graph and Q is a query. A *substitution* for query Q is a mapping $\text{VAR}_Q \rightarrow V$. If θ is a substitution for query Q, then $Q\theta$ denotes the replacement of all variables $?v$ in V_Q by $\theta(?v)$. Hence, the graph structure of $Q\theta$ is exactly like that of Q except that nodes labeled with variables are replaced by vertices in $\mathcal{S}$. A substitution θ is an *answer* for query Q w.r.t. $\mathcal{S}$ iff $Q\theta$ is a subgraph of $\mathcal{S}$. The *answer set* for query Q w.r.t. a $\mathcal{S}$ is the set $\{\theta \mid Q\theta$ is a subgraph of $\mathcal{S}\}$. For example, the substitution θ such that $\theta(?v1) =$ Amendment A0056, $\theta(?v2) =$ Bill B1432, and $\theta(?v3) =$ Pierce Dickes is the only answer for the query in Fig. 2.

The DOGMA Index

In this section we define the DOGMA index and describe an algorithm to take an existing social network graph and create the DOGMA index for it. Then, we describe algorithms to answer subgraph matching queries.

Before we define the DOGMA index, we first define what it means to merge two graphs. Suppose $G = (V, E, \lambda)$ is a graph, and $G_1 = (V_1, E_1, \lambda_1)$ and $G_2 = (V_2, E_2, \lambda_2)$ are two graphs such that $V_1, V_2 \subseteq V$ and k is an integer such that $k \leq \max(|V_1|, |V_2|)$. Graph $G_m(V_m, E_m, \lambda_m)$ is said to be a *k-merge of graphs* G_1, G_2 w.r.t. G iff: (i)$|V_m| = k$; (ii) there is a *surjective* (i.e., onto) mapping $\mu : V_1 \cup V_2 \rightarrow V_m$ called the *merge mapping* such that for all $v \in V_m$, $\text{rep}(v) = \{v' \in V_1 \cup V_2 \mid \mu(v') = v\}$, and $e_m = (v_1, v_2) \in E_m$ iff there exist $v_1' \in \text{rep}(v_1), v_2' \in \text{rep}(v_2)$ such that $e = (v_1', v_2') \in E$. The basic idea tying k-merges to the DOGMA index is that we want DOGMA to be a binary tree, each of whose nodes occupies a disk page. Each node is labeled by a graph that "captures" its two children in some way. As each page has a fixed size, the number k limits the size of the graph so that it fits on one page. The idea is that if a node N has two children, N_1 and N_2, then the graph labeling node N should be a k-merge of the graphs labeling its children.

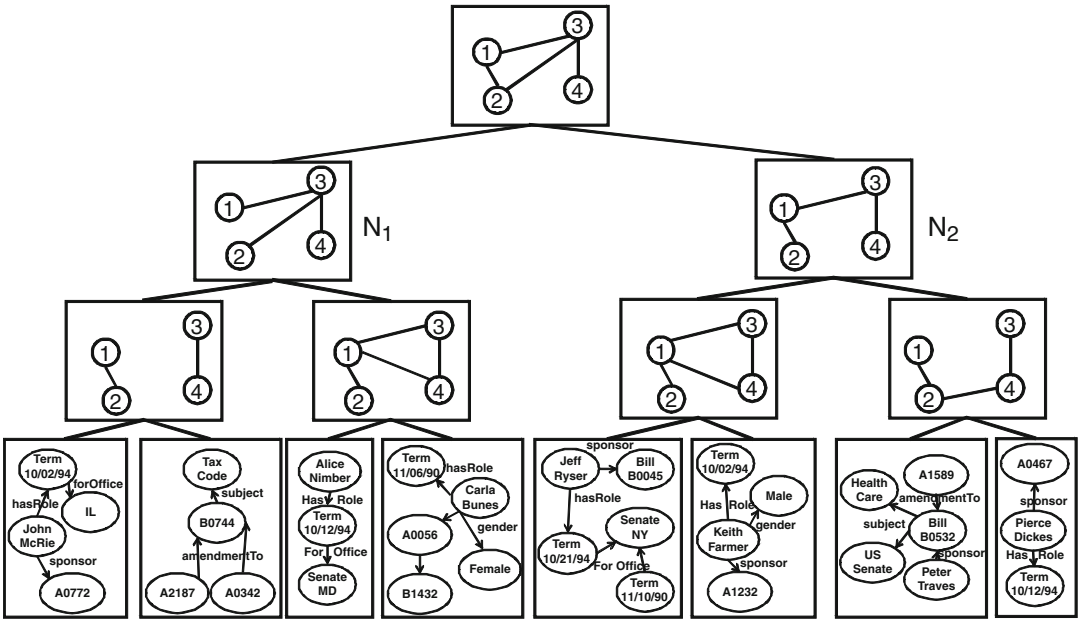
A DOGMA index for a social network graph $\mathcal{S}$ is a generalization of the well-known binary tree specialized to represent social network graphs in the following manner.

Definition 1 *A DOGMA index of order k ($k \geq 2$)* is a binary tree $\mathbf{D}_\mathcal{S}$ with the following properties:

1. Each node in $\mathbf{D}_\mathcal{S}$ equals the size of a disk page and is labeled by a graph.
2. $\mathbf{D}_\mathcal{S}$ is balanced.
3. The labels of the set of leaf nodes of $\mathbf{D}_\mathcal{S}$ constitute a partition of $\mathcal{S}$.
4. If node N is the parent of nodes N_1, N_2, then the graph G_N labeling node N is a k-merge of the graphs G_{N_1}, G_{N_2} labeling its children.

Note that a single social network database can have many DOGMA indexes.

Example 2 Suppose $k = 4$. A DOGMA index for the graph of Fig. 1 might split the graph into the eight components indicated by dashed lines in Fig. 1 that become the leaf nodes of the index (Fig. 3). Consider the two leftmost leaf nodes. They can be 4-merged together to form a parent node. Other leaf nodes can also be merged together (the results of k-merging are not shown in the inner nodes).



Scaling Subgraph Matching Queries in Huge Networks, Fig. 3 A DOGMA index for the graph of Fig. 1

Building DOGMA Indexes

Even though many different DOGMA indexes can be constructed for the same social network graph, we want to find a DOGMA index with as few "cross" edges between subgraphs stored on different pages as possible. In other words, if node N is the parent of nodes N_1, N_2, then we would like relatively fewer edges in $\mathcal{S}$ between some node in G_{N_1} and some node in G_{N_2}. The smaller this number of edges, the more "self-contained" nodes N_1, N_2 are and the less likely that a query will require looking at both nodes N_1 and N_2. DOGMA can employ any external graph partitioning algorithm (many of which have been proposed in the literature) that, given a weighted graph, partitions its vertex set in such a way that (i) the total weight of all edges crossing the components is minimized and (ii) the accumulated vertex weights are (approximately) equal for both components. In our implementation, we employ the *GGGP* graph partitioning algorithm proposed in Karypis and Kumar (1999).

In order to generate a DOGMA index for a social network $\mathcal{S}$, we can intuitively proceed through the two following phases (the fully detailed version of the algorithm can be found in Bröcheler et al. (2009)).

Iterative Coarsening. Iteratively "coarsen" $\mathcal{S}$ by merging nodes in $\mathcal{S}$. This generates a sequence of social network graphs $\mathcal{S} = \mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_k$ where $\mathcal{S}_{i+1}$ is obtained by randomly merging nodes (and corresponding edges) in $\mathcal{S}_i$ till the number of vertices in $\mathcal{S}_{i+1}$ is less than or equal to half of those in $\mathcal{S}_i$. We stop when we reach the smallest m such that the set V_m of vertices associated with S_m is small enough to fit on a disk page. Thus, m is proportional to $O(\log_2(|V|))$. When constructing V_{i+1} (and the corresponding E_{i+1}) from V_i and E_i, respectively, we keep track of which vertices (resp. edges) in V_i (resp. E_i) were merged into which vertices (resp. edges) in V_{i+1} (resp. E_{i+1}). The "root" of the DOGMA index now corresponds to S_m which, implicitly, represents the entire $\mathcal{S}$.

Hierarchical Decomposition. We now decompose S_m (the root) into two to get S_m's two children, using any standard graph partitioning algorithm. Suppose this partitioning splits S_m into S_m^1 and S_m^2. We then go back and see which

vertices in S_{m-1} got merged into the vertices in S_m^1 and replace the (merged vertices) in S_m^1 by the two vertices from which the merged vertex got created. We repeat this for S_m^2. This now gives us the two children of the root of the DOGMA index for social network database $\mathcal{S}$. This process is applied iteratively till we reach the leaf level of the DOGMA index (we will know when to stop because the vertices in the index can no longer be "unfolded").

Processing Queries with DOGMA

The basic query answering for answering graph matching queries using the DOGMA index, called DOGMA_basic (Bröcheler et al. 2009), is a recursive, depth-first algorithm which searches the space of all substitutions for the answer set to a given query Q w.r.t a graph $\mathcal{S}$. For each variable vertex v in Q, the algorithm maintains a set of constant vertices $R_v \subseteq V_S$ (called result candidates) to prune the search space; for each answer substitution θ for Q, we have $\theta(v) \in R_v$. In other words, the result candidates must be a superset of the set of all matches for v. Hence, we can prune the search space by only considering those substitutions θ for which $\theta(v) \in R_v$ for all variable vertices v in Q. The algorithm initializes the result candidates for all variable vertices v in Q which are connected to a constant vertex c in Q through an edge having the label specified in Q. Here we employ the fact that any answer substitution θ must be such that $\theta(v)$ is a neighbor of c, and thus, the set of all neighbors of c in $\mathcal{S}$ reachable by an edge labeled l are result candidates for v. We use the DOGMA index to efficiently retrieve the neighborhood of c. If v is connected to multiple constant vertices, we take the intersection of the respective constraints on the result candidates.

At each recursive invocation, the algorithm extends the given substitution and narrows down the result candidates for all remaining variable vertices correspondingly. To extend the given substitution θ, we greedily choose the variable vertex w with the smallest set of result candidates. This yields a locally optimal branching factor of the search tree since it provides the smallest number of extensions to the current substitution. In fact, if the set of result candidates is empty, then we know that θ cannot be extended to an answer substitution, and we thus directly prune the search. Otherwise, we consider all the possible result candidates m for w by deriving extended substitutions θ' from θ which assign m to w and then calling DOGMA_basic recursively on θ'. By assigning the constant vertex m to w, we can constrain the result candidates for all neighboring variable vertices as discussed above.

This basic query-answering algorithm only uses "short-range" dependencies, i.e., the immediate vertex neighborhood of variable vertices, to constrain their result candidates. While this suffices for most simple queries, considering "long-range" dependencies can yield additional constraints on the result candidates and thus improve query performance. For instance, the result candidates for v_1 in our example query not only must be immediate neighbors of "Carla Bunes": in addition, they must be at most at a distance of two from "Health Care". More formally, let $d_S(u, v)$ denote the length of the shortest path between two vertices $u, v \in V_S$ in the undirected counterpart of a graph $\mathcal{S}$, and let $d_Q(u, v)$ denote the distance between two vertices in the undirected counterpart of a query Q. A long-range dependency on a variable vertex $v \in V_Q$ is introduced by any constant vertex $c \in V_Q$ with $d_Q(v, c) > 1$.

We can exploit long-range dependencies to further constrain result candidates. Let v be a variable vertex in Q and c a constant vertex with a long range dependency on v. Then any answer substitution θ must satisfy $d_Q(v, c) \geq d_S(\theta(v), c)$ which, in turn, means that $\{m \mid d_S(m, c) \leq d_Q(v, c)\}$ are result candidates for v. This is the core idea of the DOGMA_adv algorithm (Bröcheler et al. 2009), which improves over and extends DOGMA_basic. In addition to the result candidates set R_v, the algorithm maintains sets of distance constraints C_v on them. As long as a result candidates set R_v remains uninitialized, we collect all distance constraints that arise from long-range dependencies on the variable vertex v in the constraints set C_v. After the result candidates are initialized, we ensure that all elements in R_v satisfy the distance constraints in C_v. Maintaining additional constraints therefore

reduces the size of R_v and hence the number of extensions to θ we have to consider.

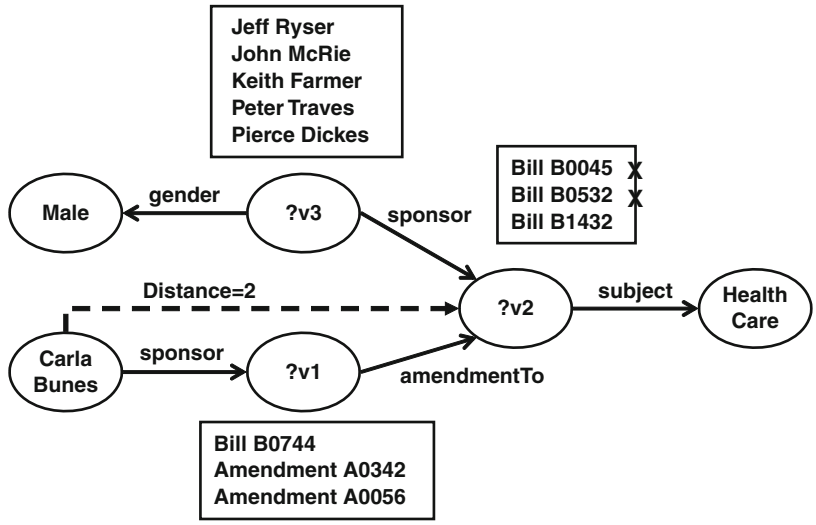
DOGMA_adv assumes the existence of a *distance index* to efficiently look up $d_{\mathcal{S}}(u, v)$ for any pair of vertices $u, v \in V_{\mathcal{S}}$, since computing graph distances at query time is clearly inefficient. But how can we build such an index? Computing all-pairs shortest path has a worst-case time complexity $O(|V_{\mathcal{S}}|^3)$ and space complexity $O(|V_{\mathcal{S}}|^2)$, both of which are clearly infeasible for large social network graphs. However, we do not need to know the *exact* distance between two vertices for DOGMA_adv to be correct. Since all the distance constraints in DOGMA_adv are *upper bounds*, all we need is to ensure that $\forall u, v \in V_{\mathcal{S}}$, the distance retrieved by the index is less than or equal to $d_{\mathcal{S}}(u, v)$.

Thus, we can extend the DOGMA index to include distance information and build two "lower bound" distance indexes, DOGMA_ipd and DOGMA_epd, that use approximation techniques to achieve acceptable time and space complexity. As seen before, the leaf nodes of the DOGMA index $\mathbf{D}_{\mathcal{S}}$ are labeled by subgraphs which constitute a partition of $\mathcal{S}$. For any node $N \in \mathbf{D}_{\mathcal{S}}$, let P_N denote the union of the graphs labeling all leaf nodes reachable from N. Hence, P_N is the union of all subgraphs in $\mathcal{S}$ that were eventually merged into the graph labeling N during index construction and therefore corresponds to a larger subset of $\mathcal{S}$. For example, the dashed lines in Fig. 1 mark the subgraphs P_N for all index tree nodes N of the DOGMA index shown in Fig. 3, whereas bolder lines indicate boundaries corresponding to nodes of lower depth in the tree.

The DOGMA *internal partition distance* (DOGMA_ipd) index stores, for each index node N and vertex $v \in P_N$, the distance to the outside of the subgraph corresponding to P_N. We call this the *internal partition distance* of v, N, denoted $\text{ipd}(v, N)$, which is thus defined as $\text{ipd}(v, N) = \min_{u \in V_{\mathcal{S}} \setminus P_N} d_{\mathcal{S}}(v, u)$. We compute these distances during index construction. At query time, for any two vertices $v, u \in V_{\mathcal{S}}$ we first use the DOGMA tree index to identify those distinct nodes $N \neq M$ in $\mathbf{D}_{\mathcal{S}}$ such that $v \in P_N$ and $u \in P_M$, which are at the same level of the tree and closest to the root. If such nodes do not exist (because v, u are associated with the same leaf node in $\mathbf{D}_{\mathcal{S}}$), then we set $d_{\text{ipd}}(u, v) = 0$. Otherwise, we set $d_{\text{ipd}}(u, v) = \max(\text{ipd}(v, N), \text{ipd}(u, M))$. It is easy to see that d_{ipd} is an admissible lower bound distance, since $P_N \cap P_M = \emptyset$. By choosing those distinct nodes which are closest to the root, we ensure that the considered subgraphs are as large as possible, and hence, $d_{\text{ipd}}(u, v)$ is the closest approximation to the actual distance.

Example 3 Consider the example of Figs. 1 and 2. Figure 4 shows the initial result candidates for each of the variable vertices in boxes. We can determine that there is a long-range dependency between "Carla Bunes" and variable vertex $?v_2$ at distance 2. The boldest dashed line in Fig. 1 marks the top-level partition and separates the sets P_{N_1}, P_{N_2}, where N_1, N_2 are the two nodes directly below the root in the DOGMA index in Fig. 3. We can determine that $\text{ipd}(\text{Carla Bunes}, N_2) = 3$, and since Bill B0045 and B0532 lie in the other subgraph, it follows that $d_{\text{ipd}}(\text{Carla Bunes}, \text{B0045/B0532}) = 3$, and therefore, we can prune both result candidates.

The DOGMA *external partition distance* (DOGMA_epd) index also uses the partitions in the index tree to compute a lower bound distance. However, it considers the distance to *other* subgraphs rather than the distance within the *same* one. For some fixed level L, let $\mathcal{N}_L$ denote the set of all nodes in $\mathbf{D}_{\mathcal{S}}$ at distance L from the root. As discussed above, $P = \{P_N\}_{N \in \mathcal{N}_L}$ is a partition of $\mathcal{S}$. The idea behind DOGMA_epd is to assign a color from a fixed list of colors C to each subgraph $P_N \in P$ and to store, for each vertex $v \in V_{\mathcal{S}}$ and color $c \in C$, the shortest distance from v to a subgraph colored by c. We call this the *external partition distance*, denoted $\text{epd}(v, c)$, which is thus defined as $\text{epd}(v, c) = \min_{u \in P_N, \phi(P_N) = c} d_{\mathcal{S}}(v, u)$ where $\phi : P \to C$ is the color assignment function. We store the color of P_N with its index node N so that for a given pair of vertices u, v we can quickly retrieve the colors c_u, c_v of the subgraphs to which u and v belong. We then compute $d_{\text{epd}}(v, u) = \max(\text{epd}(v, c_u), \text{epd}(u, c_v))$. It is easy to see that d_{epd} is an admissible lower bound distance.



Scaling Subgraph Matching Queries in Huge Networks, Fig. 4 Using DOGMA_ipd for query answering

Cloud-Oriented Subgraph Matching

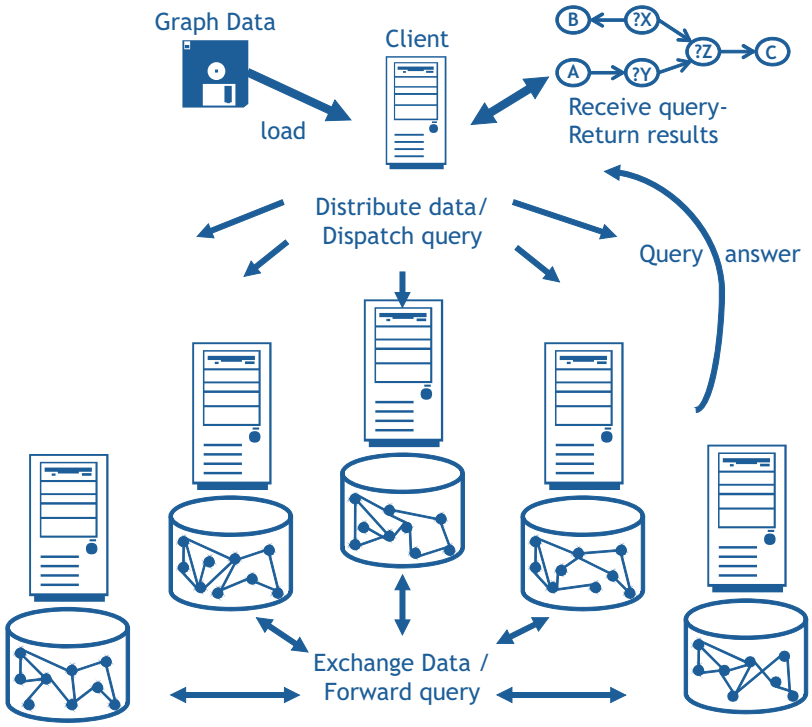
This section presents the COSI system which distributes a social network graph across multiple compute machines and answers subgraph matching queries in parallel using an asynchronous query-answering algorithm that does not rely on central orchestration.

Figure 5 shows a schematic view of the architecture of COSI. We assume that a compute cloud consists of k compute nodes and one "client" node. Compute nodes communicate *directly* without going through the client node, thus preventing the client node from becoming a communication bottleneck. The client node takes a query and directs it or parts of it to one or more compute nodes that complete the computation of the answer. The complete answer is then shipped to the client node. In Fig. 5, $k = 5$ and the compute machines are shown in the lower half of the figure. Each of those machines stores a fragment of the graph in a local graph database and responds to query requests specifically addressed to it. The architecture and system we present in this entry does not depend on any particular local graph database.

Distributing a Social Network Graph

We now address the question: How do we distribute a social network graph across a cloud so that we can efficiently process subgraph matching queries? In partitioning the social network data, we follow two objectives: (i) all k compute machines should store roughly the same amount of data to balance the load across machines, and (ii) the partition should minimize the expected query execution time.

At a high level, we achieve these objectives as follows. First, we transform the social network graph $\mathcal{S}$ into a simple weighted graph $\mathsf{WG}(\mathcal{S})$. Intuitively, the weight of an edge $e = (u, v)$ in $\mathsf{WG}(\mathcal{S})$ refers to the sum of the probability that v will be retrieved immediately after u and vice versa when an arbitrary query is processed. If this probability is (relatively) high, then the two vertices should be stored on the same compute node. Then, we use these to partition $\mathcal{S}$ across the k compute machines so that expected communication costs are minimized. In the remainder, we assume there is a probability distribution $\mathbb{P}$ over the space of all queries. Intuitively, $\mathbb{P}(Q)$ is the probability that a random subgraph matching query posed to a social network is Q. For any real-world online social network, $\mathbb{P}$ can be



Scaling Subgraph Matching Queries in Huge Networks, Fig. 5 Architecture of COSI

easily learned from frequency analysis of past query logs.

We start by introducing our formalization of *query plans* and *query traces*. A query plan $qp(Q)$ for a query Q is a sequence of two types of *operations*: the first type retrieves the neighborhood of vertex v (from whichever compute node it is on), and the second type performs some computation (e.g., check a selection condition or perform a join) on the results of previous operations. This definition is compatible with most existing definitions of query plans in the database literature. Now suppose $x = qp(Q)$ is a query plan for a query Q on a social network graph $\mathcal{S}$. The query trace of executing x on $\mathcal{S}$, denoted $qt(x, \mathcal{S})$, consists of (i) all the vertices v in $\mathcal{S}$ whose neighborhood is retrieved during execution of query plan x on $\mathcal{S}$ and (ii) all *pairs* (u, v) of vertices where immediately after retrieving u's neighborhood, the query plan retrieves v's neighborhood (in the next operation of x). When processing a query, we make the reasonable assumption that index retrievals are cached so that repeated vertex neighborhood retrievals are read from memory and hence the query trace $qt(x, \mathcal{S})$ can be defined as a set rather than as a multiset. Traces contain consecutive retrievals of vertex neighborhoods – this allows us to store neighborhoods of both u and v on the same compute node, avoiding unnecessary communication.

The probability distribution $\mathbb{P}$ on queries can be used to infer a probability distribution $\tilde{\mathbb{P}}$ over the space of feasible query plans: $\tilde{\mathbb{P}}(x) = \sum_{Q \in \mathcal{Q}: qp(Q)=x} \mathbb{P}(Q)$. This says that the probability of a query plan is the sum of the probabilities of all queries which use that query plan. In the rest of the entry, we will abuse notation and denote both PDFs by $\mathbb{P}$. We can now define the *probabilities of retrieval and co-retrieval*. The probability, $\mathbb{P}(v)$, of retrieving v when executing a random query plan is $\sum_{x \in qp(\mathcal{Q}): v \in qt(x,\mathcal{S})} \mathbb{P}(x)$. Thus, the probability of retrieving v is the sum of the probabilities of all query plans that retrieve v. The probability $\mathbb{P}(v_1, v_2)$ of retrieving v_2 immediately after v_1 is

$\sum_{x \in qp(\mathcal{Q}):(v_1,v_2) \in qt(x,\mathcal{S})} \mathbb{P}(x)$. This says that the probability of retrieving v_2 immediately after v_1 is the sum of the probabilities of all query plans that retrieve v_2 immediately after v_1.

We can associate a weighted graph $\mathsf{WG}(\mathcal{S})$ with the graph $\mathcal{S} = (V, E, \lambda)$. The weighted graph is the complete graph $(V, V \times V, w)$ where $w(v_1, v_2) = \mathbb{P}(v_1, v_2) + \mathbb{P}(v_2, v_1)$. An *edge cut* C of a weighted graph is a partition of the vertices into components. An edge (u, v) in the graph is said to *cross the edge cut* C if u is in one component of the partition and v is in another. The size of an edge cut is the sum of the weights of the edges that cross the cut. C is said to be a *minimum cut* iff there is no other cut C' such that the size of C' is less than the size of C. The important theorem we gave in Bröcheler et al. (2010) shows that the minimal edge cut of $\mathsf{WG}(\mathcal{S})$ corresponds to the partition of $\mathcal{S}$ across k compute nodes that minimizes expected cost of executing a query.

Since computing minimal edge cuts is a well-known NP-hard problem, we develop heuristic techniques to partition the graph that allow us to obtain suboptimal partitions of high quality, without incurring in the expensive computational costs of obtaining the optimal ones. We start by defining the concept of *vertex force vector*. Let $\mathcal{P} = \{P_1, \ldots, P_k\}$ be a partition of $\mathcal{S}$ and consider any component P_i. The *vertex force vector*, denoted $\left|\vec{v}\right|$, of any vertex $v \in \mathcal{S}$ is a k-dimensional vector where $|\vec{v}|[i] = f_{\mathcal{P}}\left(\sum_{x \in \text{ngh}(v) \cap P_i} w((v, x))\right)$ and $f_{\mathcal{P}} : \mathbb{R}^+ \to \mathbb{R}$ is a function called the *affinity measure*.

The vertex force vector intuitively specifies the "affinity" between a vertex and each component as measured by the affinity measure $f_{\mathcal{P}}$. An affinity measure takes the connectedness between a vertex v and the respective component as an argument. The vertex force vector captures the strength with which each component "pulls" on the vertex and is used as the basis for a vertex assignment decision: intuitively, if an inserted edge introduces a new vertex v, we first compute the vertex force vector $\left|\vec{v}\right|$ and then assign v to the component P_j where $j = \text{argmax}_{1 \le i \le k} \left|\vec{v}\right|[i]$.

COSI uses an affinity measure that is a linear combination of three factors: *connectedness*, *imbalance*, and *size*. Obviously, evaluating the connectedness of a vertex v to a component P_i is crucial for edge cut minimization – we measure this as the number of edges that connect v to the vertices in P_i. Moreover, balanced partitions lead to even workload distribution, thus enhancing parallelism. Let $|P_i|_E = \sum_{x \in P_i} \deg(x)$ be the number of edges in P_i and let T be an estimate (even a bad one) of the total number of edges that a given graph is expected to be. Then a reasonable measure of imbalance is the standard deviation of $\frac{\{|P_i|_E\}_{1 \le i \le k}}{T}$. Finally, we regulate the size of components by comparing the actual size of a component to its expected one. If a component grows beyond its expected size, we punish such growth more aggressively than imbalance does alone by reducing the affinity further according to the metric $\min(-\frac{|P_i|_E - \frac{T}{k}}{T}, 0)$.

Consider now the case of a new set of edges to be inserted into a social network graph, given that a partition $\mathcal{P} = P_1, \ldots, P_k$ of the graph already exists (this can be used to create a partition for the first time by assuming $\mathcal{S} = \emptyset$). A naive GreedyInsert insertion algorithm would iterate over all new vertices v: for each vertex v it would compute the vertex force vector and assigns v to the component P_i such that $|\mathbf{v}|[i]$ is maximal – fortunately we can do better.

Our COSI_Partition algorithm (Bröcheler et al. 2010) leverages graph *modularity* (Blondel et al. 2008) to identify a strongly connected subgraph that is loosely connected to the remaining graph. However, modularity cannot be used blindly as our balance requirement must also be met. The *modularity* of a partition $\mathcal{P}$ of an undirected graph $G = (V, E)$ with weight function $w : E \to \mathbb{R}$ is defined as

$$\text{mod}(\mathcal{P}) = \sum_{P \in \mathcal{P}} \left(\frac{W(P, P)}{2\,|E|} - \frac{\deg_W(P)^2}{(2\,|E|)^2} \right)$$

where $\deg_w(v) = \sum_{x \in V} w((v, x))$ is the weighted degree of vertex v, $W(X, Y) = \sum_{x \in X, y \in Y} w((x, y))$ is the sum of edge weights connecting two sets of vertices $X, Y \subset V$, and

$\deg_W(X) = \sum_{x \in X} \deg_w(x)$ is the weighted degree of a set of vertices $X \subset V$. Intuitively, components with high modularity are densely connected subgraphs which are isolated from the rest of the graph. Our algorithm iteratively builds high modularity components and then assigns all vertices in a component to one compute node based on the vertex force vector. Let $B \subset V$ be a set of vertices. We generalize the notion of a vertex force vector by defining $\left|\vec{B}\right|[i] = f_{\mathcal{P}}\left(\sum_{v \in B} \sum_{x \in \text{ngh}(v) \cap P_i} w((v,x))\right)$. The intuition behind our partitioning algorithm is that assigning vertices at the aggregate level of isolated and densely connected components yields good partitions because (i) we respect the topology of the graph, (ii) most edges are within components and therefore cannot be cut, and (iii) force vectors of sets of vertices combine the connectedness information of many vertices leading to better assignment decisions.

Processing Queries with COSI

Our COSI_basic parallel query processing algorithm (Bröcheler et al. 2010) operates asynchronously and in parallel across all compute nodes. A user issues query Q to the client node which "prepares" the query. In particular, it selects one constant vertex c from Q and determines the compute node that hosts c – the prepared query is then forwarded to this node.

The algorithm proceeds depth first, substituting vertices for variables in Q one at a time. We maintain a set of result candidates for each variable in Q. The algorithm assumes there is an index retrieval function that retrieves $\text{ngh}_I(v)$ from the local index (which could be implemented many ways – we used a DOGMA index in the experiments) on the compute node. The algorithm arbitrarily chooses the next vertex to be substituted. Incoming queries come with a selected variable to be instantiated with a vertex ID. The algorithm updates the candidate result sets by retrieving the neighborhood of the newly substituted vertex from the index. It then checks if any results have been found or whether the current substitution cannot yield a valid result. If neither condition holds, the algorithm selects the next variable v' to be substituted and forwards the query to those compute nodes that host potential substitution candidates for v'. All query results are sent to the client which returns them to the user.

COSI_basic does not rely on central orchestration – it uses depth-first search so the branches of the search tree are traversed in parallel while ensuring that no branch gets explored multiple times. After forwarding the prepared query to a compute node, the client waits for incoming results of that query and forwards those to the user. As we explore branches in parallel, the client node cannot be notified when the search for query results has completed. Keeping track of the current number of parallel executions for each query would introduce significant synchronization cost. Instead, the client node keeps track of the time t_{last} at which the last result of a running query has come in. If the difference between the current time and t_{last} exceeds a threshold, the client node asks all compute nodes for a list of query IDs of all currently running queries. The client node merges these lists and closes all queries whose IDs are not contained. To avoid the case where a query is being forwarded to another compute node at the very moment that the client node asks for all query IDs, each compute node keeps query IDs in their local list up to a certain grace period.

The choice of the next variable to be instantiated has profound implications on the running time of COSI_basic, as some substitutions yield larger branching factors in the search than others. Our COSI_heur algorithm (Bröcheler et al. 2010) handles this by choosing the variable vertex v' which has the lowest cost according to a function h_{opt}. First, to reduce the branching factor, we could choose the variable vertex v' with the smallest number of result candidates. This heuristic only considers the branching factor of the immediate next iteration but is nevertheless an important metric to consider in the cost heuristic. Second, whenever we instantiate a vertex on a remote component, we have to send a message to the appropriate compute node which is expensive. Therefore, we consider the fraction of result candidates which are not stored locally as

a cost metric. When we have to send a query to remote nodes for further processing, we would like to distribute the workload evenly across all nodes. Hence, we also analyze the distribution of result candidates by node via the cost metric

$$ds(v) = \sqrt{\sum_{1 \le i \le k} \left(|R_v^i| - \frac{|R_v|}{k} \right)^2}$$

where R_v^i is the set of result candidates for vertex v restricted to those which reside on compute node i. Finally, we define

$$h_{\text{opt}}(v) = |R_v| \times (1 - \frac{|R_v^l|}{\alpha \times |R_v|}) \times (1 + \beta \times \frac{ds(v)}{|R_v|})$$

where l is the ID of the local compute node and α and β are constants that determine how much the model favors locality over parallelism.

Experimental Results

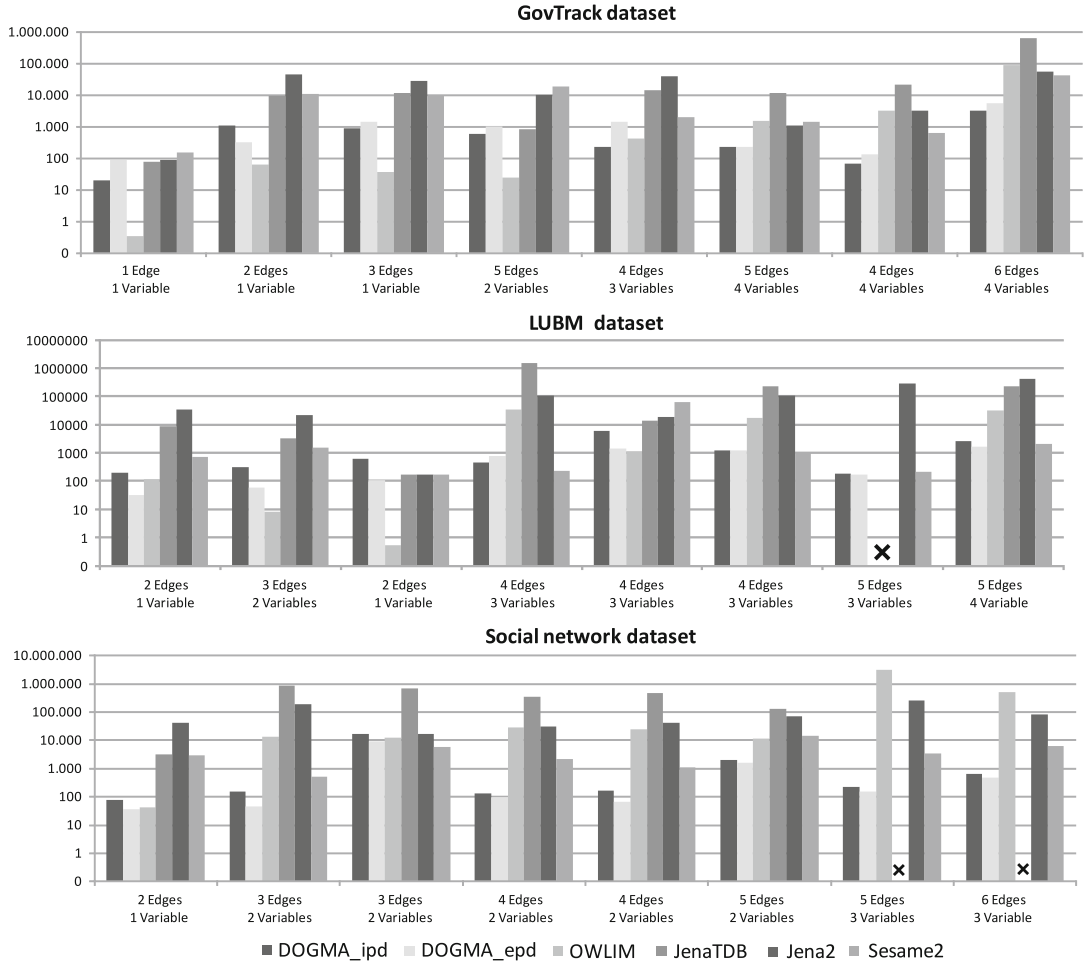
In this section we present the results of the experimental assessment we performed of both the DOGMA index and the COSI system.

DOGMA

We tested the DOGMA index on RDF data and compared its performance with four RDF database systems developed in the Semantic Web community that are most widely used and have demonstrated superior performance in previous evaluations (Lee et al. 2008): *Sesame2* (2013), *Jena2* (Wilkinson et al. 2003), JenaTDB (2013), and the internal memory version of *OWLIM* (Kiryakov et al. 2005). Moreover, we used three different RDF datasets. *GovTrack* (GovTrack dataset 2013) consists of more than 14.5 million triples describing data about the US Congress. The *Lehigh University Benchmark* (LUBM) (2013) is frequently used within the Semantic Web community as the basis for evaluation of RDF and ontology storage systems – we generated a dataset of more than 13.5 million triples. Finally, we used a fragment of the Flickr social network (2013) collected by researchers of the MPI Saarbrücken to analyze online social networks (Mislove et al. 2007). The fragment was anonymized and contains approximately 16 million triples. The GovTrack and Flickr datasets are well connected (with the latter being denser than the former), whereas the dataset generated by the LUBM benchmark is a sparse and almost degenerate graph containing a set of small and loosely connected subgraphs.

We designed a set of graph queries with varying complexity, where constant vertices were chosen randomly and queries with an empty result set were filtered out. Queries were grouped into classes based on the number of edges and variable vertices. We repeated the query time measurements multiple times for each query, eliminated outliers, and averaged the results. Finally, we averaged the query times of all queries in each class. All experiments were executed on a machine with a 2.4 GHz Intel Core 2 processor and 3 GB of RAM.

In a first round of experiments, we designed several relatively simple graph queries for each dataset, containing no more than six edges, and grouped them into eight classes. The results of these experiments are shown in Fig. 6 which reports the query times for each query class on each of the three datasets. Missing values in the figure indicate that the system did not terminate on the query within a reasonable amount of time (around 20 min). Note that the query times are plotted in logarithmic scale to accommodate the large discrepancies between systems. The results show that OWLIM has low query times on low-complexity queries across all datasets. This result is not surprising, as OWLIM loads all data into main memory prior to query execution. The performance advantage of DOGMA_ipd and DOGMA_epd over the other systems increases with query complexity on the GovTrack and the Flickr dataset, where our proposed techniques are orders of magnitude faster on the most complex queries. On the LUBM dataset, however, Sesame2 performs almost equally for the more complex queries. Finally, DOGMA_epd is slightly faster on the LUBM and Flickr dataset, whereas DOGMA_ipd has better performance on the GovTrack dataset.

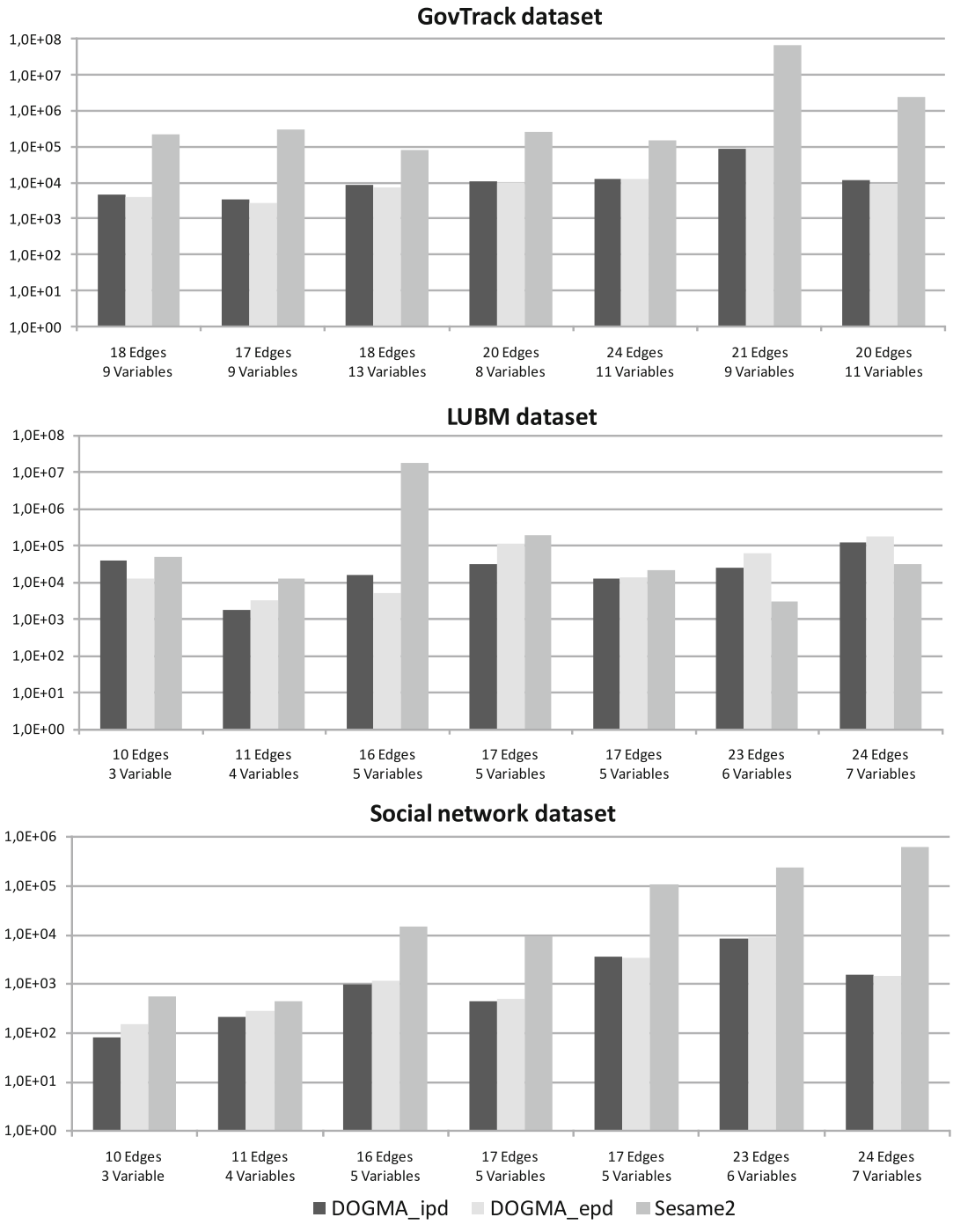


Scaling Subgraph Matching Queries in Huge Networks, Fig. 6 Query times (ms) for graph queries of low complexity

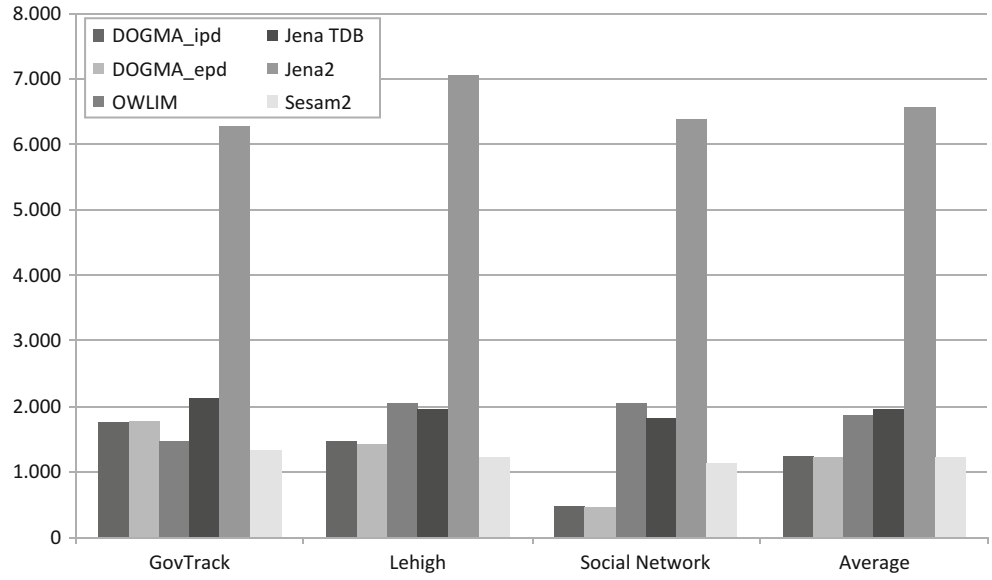
In a second round of experiments, we significantly increased the complexity of the queries, which now contained up to 24 edges. Unfortunately, the OWLIM, JenaTDB, and Jena2 systems did not manage to complete the evaluation of these queries in reasonable time, so we exclusively compared with Sesame2. The results are shown in Fig. 7. On the GovTrack and Flickr dataset, DOGMA_ipd and DOGMA_epd continue to have a substantial performance advantage over Sesame2 on all complex graph queries of up to 40,000 %. For the LUBM benchmark, the picture is less clear due to the particular structure of the generated dataset explained before.

Finally, Fig. 8 compares the storage requirements of the systems under comparison for all three datasets. The results show that DOGMA_ipd, DOGMA_epd and Sesame2 are the most memory efficient.

In conclusion, we can observe that both DOGMA_ipd and DOGMA_epd are significantly faster than all other RDF database systems under comparison on complex graph queries over non-degenerate graph datasets. Moreover, they can efficiently answer complex queries on which most of the other systems do not terminate or take up to 400 times longer while maintaining a satisfactory storage footprint. DOGMA_ipd and



Scaling Subgraph Matching Queries in Huge Networks, Fig. 7 Query times (ms) for graph queries of high complexity



Scaling Subgraph Matching Queries in Huge Networks, Fig. 8 Index size (MB) for different datasets

DOGMA_epd have similar performance, yet differences exist which suggest that each index has unique advantages for particular queries and graph structures.

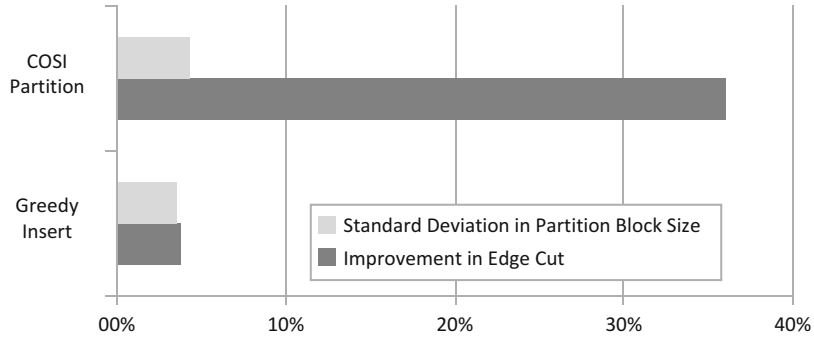
COSI

In the experiments with COSI we used the social network graph studied in Mislove et al. (2007). This graph contains 778M edges and describes personal relationships and group memberships crawled from Facebook, Orkut, Flickr, and LiveJournal. We fixed the coefficients for the affinity measure by hand. Both, the imbalance and excessive size metric, were given an equal weight of one. The connectedness measure was set relative to the number of edges we considered per batch. We experimented with different batch sizes and found best performance for half a million edges.

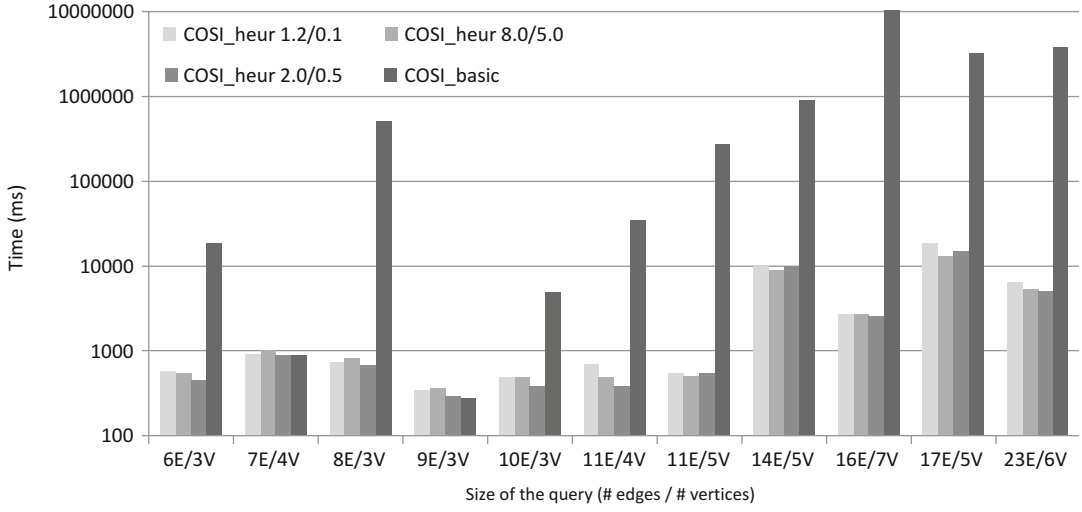
We developed a communication infrastructure for the compute nodes based on the Java NIO libraries which is used to send the graph data during the loading and the queries during the query-answering stages. The communication infrastructure handles contention at individual nodes and variations in network latency. It is optimized to ensure that the client node's requests for outstanding queries are answered quickly. In our experiments, we used a cluster of 16 compute nodes, out of which one served as a client node and the remaining 15 nodes served as compute nodes. All compute nodes had an identical hardware configuration with a 4-core 2.16 GHz Intel CPU, 4 GB of RAM, and 80 GB IDE 7,200 rpm hard drive. The client node's hardware differed slightly with an 8-core CPU and 8 GB of RAM.

Figure 9 compares COSI_Partition's performance with that of the GreedyInsert algorithm. To validate our experiments, we used a random partitioning scheme, which assigns vertices to compute nodes uniformly at random, as the naive baseline in our experiments and report all results in comparison to this baseline. COSI_Partition achieves a substantial 36 % improvement in edge cut over the naive baseline at a total running time of 10.5 h for all 778M edges. GreedyInsert only achieves a marginal improvement in edge cut. COSI_Partition significantly outperforms greedy batch insertion by 33 % with only slightly higher imbalance as measured in the standard deviation in component size relative to average size of a component.

Figure 10 compares COSI_basic against COSI_heur for three different parameter settings



Scaling Subgraph Matching Queries in Huge Networks, Fig. 9 Comparison of partitioning methods



Scaling Subgraph Matching Queries in Huge Networks, Fig. 10 Query times by query-answering algorithm on the 778M edge dataset

of function h_{opt}: ($\alpha = 1.2, \beta = 0.1$) which strongly favors locality over parallelism, ($\alpha = 8.0, \beta = 5.0$) which strongly favors parallelism over locality, and ($\alpha = 2.0, \beta = 0.5$) which balances locality and parallelism. The queries have increasing complexity as measured by the number of edges (E) and variables (V) in the query graph. All query times were averaged across six independent runs with complete system restarts after each run to empty caches. Note that the graph is plotted in logarithmic scale to accommodate the huge differences in query times.

COSI_heur drastically outperforms COSI_basic by up to four orders of magnitude on all but two queries, and the performance gap seems to grow exponentially with the query complexity. A close look at the difference in performance between the variants of COSI_heur reveals that the third configuration outperforms the first one on nine queries, with a tie on the remaining two, and outperforms the second configuration on eight queries, being slower only on three. These results suggest that a balanced choice of parameters leads to a better h_{opt}.

Related Work and Conclusions

The problem of efficiently evaluating subgraph matching queries over huge graphs/networks has been recently addressed in different scenarios, among which social network analysis and RDF database management play an important role (Martín and Gutierrez 2009). A wide variety of methods for social network analysis have been proposed (Borgatti et al. 2002; Nooy et al. 2005; Huisman and Duijn 2005). However, most algorithms operate solely in memory, loading the entire graph from disk and then executing the analysis. For social networks of the size of Facebook, Flickr, or Orkut, such an approach becomes infeasible. To handle social networks of such magnitude, one needs to store and query network data efficiently on disk. More importantly, complex queries involving even a few joins can quickly cause such approaches to run into trouble. Ronen and Shmueli (2009) introduce a social network-specific query language and show how such queries can be answered on moderately sized datasets. However, their query language is geared toward users of a social network in helping them communicate with friends.

Graph-structured RDF data has been studied in the Semantic Web community (Mahmoudi-Nasab and Sakr 2010). Initial approaches to RDF storage (Broekstra et al. 2003; Sintek and Kiesel 2006; Wilkinson et al. 2003) stored the graph in relational tables and then used a relational query engine to answer queries. Abadi et al. (2007) showed that storing RDF in a vertical database leads to significant query time improvements. Stocker et al. (2008) uses triple selectivity estimation techniques similar to those used in relational database systems. Pugliese et al. (2008) and Udrea et al. (2007) are the first to propose specific tree-structured indexes for RDF. All these approaches work on single machines. In response to the increasing need of scalability when facing extremely large RDF datasets, two approaches have essentially been proposed so far: *scale up* and *scale out*. In scaling up, existing RDF databases, such as RDF-3X (Neumann and Weikum 2008), Sesame (Broekstra et al. 2003), or YARS (Harth and Decker 2005), are simply run on more powerful machines. As such it requires no technological innovation but is very costly and limited by current hardware. In scaling out, multiple machines are utilized to store the data but all operations on the data are centrally executed. Parallel storage regimes, such as YARS2 (Harth et al. 2007), are cheaper but still limited in their scalability due to central execution. Our COSI system demonstrated efficient query answering across multiple machines without central orchestration.

Earlier work on database technologies for general graph data such as Lore (Goldman et al. 1999) considered much smaller graphs than the social networks we study here. More recent work (e.g., Cheng et al. 2009; Giugno and Shasha 2002; Ke et al. 2010; Sakr 2009; Zhang et al. 2010; Zhu et al. 2010) focuses on heuristics to predict the cost of answering strategies based on statistics about the dataset and the current state of query processing and then choose a strategy to minimize cost. However, due to the highly heterogeneous nature of network data (Newman 2003), such predictions can become inaccurate. Zou et al. (2009) proposes to transform vertices into points in a vector space, thus converting queries into distance-based multi-way joins over the vector space. In Cheng et al. (2008) the authors propose a two-step join optimization algorithm based on a cluster-based join index. *GADDI* is proposed in Zhang et al. (2009) that employs a structural distance-based approach and a dynamic matching scheme to minimize redundant calculations. GADDI can handle graphs with thousands of vertices, which are common in many biological applications. In Zhang et al. (2010) the authors propose *SUMMA*, which improves over GADDI and employs more advanced indices, becoming capable to handle graphs with up to tens of millions of vertices. The algorithm in Zhu et al. (2010) employs an aggressive pruning strategy based on an index storing label distributions. In Natale et al. (2010), the authors argue that existing indices over *sets* of data graphs do not support efficient pruning when they face graphs with tens of

thousands of vertices. They propose an index that is specifically targeted at this scenario.

In this entry we have described a disk-oriented index and a graph partitioning technique that make the processing of complex subgraph matching queries on very large graph data feasible. The DOGMA index is based on the simple observation that the size of any real-world social network graph is likely to be orders of magnitude larger than that of any subgraph matching query graph a user is likely to ask. Thus, it is possible to build an index for efficiently executing such queries that ensures that vertices in a social network graph that are "near" each other be stored together on a disk page. On the other hand, the COSI system is able to effectively distribute a social network graph across multiple compute machines and answer subgraph matching queries asynchronously in parallel. The experimental results confirm the feasibility of both approaches.

Cross-References

- ▸ Extracting and Inferring Communities via Link Analysis
- ▸ Graph Matching
- ▸ RDF
- ▸ SPARQL
- ▸ Subgraph Extraction for Trust Inference in Social Networks

References

Abadi DJ, Marcus A, Madden S, Hollenbach KJ (2007) Scalable semantic web data management using vertical partitioning. In: VLDB, Vienna, pp 411–422

Blondel V, Guillaume J, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008:P10008

Borgatti SP, Everett MG, Freeman LC (2002) Ucinet for windows: software for social network analysis. Harvard: Analytic Technologies

Bröcheler M, Pugliese A, Subrahmanian VS (2009) DOGMA: a disk-oriented graph matching algorithm for RDF databases. In: ISWC, Chantilly, pp 97–113

Bröcheler M, Pugliese A, Subrahmanian VS (2010) COSI: cloud oriented subgraph identification in massive social networks. In: Memon N, Alhajj R (eds) ASONAM, Odense. IEEE Computer Society, pp 248–255

Broekstra J, Kampman A, van Harmelen F (2003) Sesame: an architecture for storing and querying RDF data and schema information. In: Spinning the semantic web, Dieter Fensel, James A. Hendler, Henry Lieberman, and Wolfgang Wahlster (Eds.). MIT Press, pp 197–222

Cheng J, Yu JX, Ding B, Yu PS, Wang H (2008) Fast graph pattern matching. In: ICDE conference, Cancun, pp 913–922

Cheng J, Ke Y, Ng W (2009) Efficient query processing on graph databases. ACM Trans Database Syst 2(1–2):48

Flickr (2013). http://www.flickr.com

Giugno R, Shasha D (2002) Graphgrep: a fast and universal method for querying graphs. In: ICPR conference, Québec City, pp 112–115

Goldman R, McHugh J, Widom J (1999) From semistructured data to XML: migrating the Lore data model and query language. In: Proceedings of the 2nd international workshop on the web and databases (WebDB'99), Philadelphia, pp 25–30

GovTrack dataset (2013). http://www.govtrack.us

Harth A, Decker S (2005) Optimized index structures for querying RDF from the web. In: Proceedings of the 3rd Latin American web congress, Buenos Aires, pp 71–80

Harth A, Umbrich J, Hogan A, Decker S (2007) YARS2: a federated repository for querying graph structured data from the web. In: ISWC, Busan, pp 211–224

Huisman M, Duijn MAV (2005) Software for social network analysis. In: Carrington PJ, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, Cambridge/New York, pp 270–316

JenaTDB (2013). http://jena.apache.org

Karypis G, Kumar V (1999) A fast and high quality multilevel scheme for partitioning irregular graphs. SIAM J Sci Comput 20:359–392

Ke Y, Cheng J, Yu JX (2010) Querying large graph databases. In: DASFAA conference, Tsukuba, pp 487–488

Kiryakov A, Ognyanov D, Manov D (2005) OWLIM – a pragmatic semantic repository for OWL. In: WISE workshops, New York, pp 182–192

Lee C, Park S, Lee D, Lee J, Jeong O, Lee S (2008) A comparison of ontology reasoning systems using query sequences. In: Proceedings of the 2nd international conference on ubiquitous information management and communication, Suwon. ACM, pp 543–546

MahmoudiNasab H, Sakr S (2010) An experimental evaluation of relational RDF storage and querying techniques. In: DASFAA workshops, Tsukuba, pp 215–226

Martín MS, Gutierrez C (2009) Representing, querying and transforming social networks with RDF/SPARQL. In: ESWC conference, Heraklion, pp 293–307

Mislove A, Marcon M, Gummadi PK, Druschel P, Bhattacharjee B (2007) Measurement and analysis of

online social networks. In: Internet measurement conference, San Diego, pp 29–42
Natale RD, Ferro A, Giugno R, Mongiovì M, Pulvirenti A, Shasha D (2010) SING: subgraph search in non-homogeneous graphs. BMC Bioinform 11:96
Neumann T, Weikum G (2008) RDF-3X: a RISC-style engine for RDF. PVLDB 1(1):647–659
Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256
Nooy W, Mrvar A, Batagelj V (2005) Exploratory social network analysis with Pajek. Structural analysis in the social sciences, vol 27. Cambridge University Press, New York
Pugliese A, Udrea O, Subrahmanian VS (2008) Scaling RDF with time. In: WWW, Beijing, pp 605–614
Ronen R, Shmueli O (2009) Evaluating very large datalog queries on social networks. In: EDBT, Saint-Petersburg, pp 577–587
Sakr S (2009) GraphREL: a decomposition-based and selectivity-aware relational framework for processing sub-graph queries. In: DASFAA conference, Brisbane, pp 123–137
Sesame2 (2013). http://www.openrdf.org
Sintek M, Kiesel M (2006) RDFBroker: a signature-based high-performance RDF store. In: ESWC, Budva, pp 363–377
Stocker M, Seaborne A, Bernstein A, Kiefer C, Reynolds D (2008) SPARQL basic graph pattern optimization using selectivity estimation. In: Proceeding of the 17th international conference on World Wide Web, Beijing, pp 595–604. ACM
The Lehigh University Benchmark (2013). http://swat.cse.lehigh.edu/projects/lubm
Udrea O, Pugliese A, Subrahmanian VS (2007) GRIN: a graph based RDF index. In: AAAI, Vancouver, pp 1465–1470
Wilkinson K, Sayers C, Kuno H, Reynolds D (2003) Efficient RDF storage and retrieval in Jena2. Proc SWDB 3:7–8
Zhang S, Li S, Yang J (2009) GADDI: distance index based subgraph matching in biological networks. In: EDBT conference, Saint-Petersburg, pp 192–203
Zhang S, Li S, Yang J (2010) SUMMA: subgraph matching in massive graphs. In: CIKM conference, Toronto, pp 1285–1288
Zhu K, Zhang Y, Lin X, Zhu G, Wang W (2010) NOVA: a novel and efficient framework for finding subgraph isomorphism mappings in large graphs. In: DASFAA conference, Tsukuba, pp 140–154
Zou L, Chen L, Özsu MT (2009) Distancejoin: pattern match query in a large graph database. VLDB Conf 2(1):886–897

Scamming

▸ Online Social Network Phishing Attack

Scan Statistics

▸ Disease Surveillance, Case Study

Schema Matching

▸ Ontology Matching

Scholarly Communication

▸ Scholarly Networks Analysis

Scholarly Networks

▸ Scholarly Networks Analysis

Scholarly Networks Analysis

Erjia Yan[1] and Ying Ding[2]
[1]College of Computing and Informatics, Drexel University, Philadelphia, PA, USA
[2]Department of Information and Library Science, School of Informatics and Computing, Indiana University, Bloomington, IN, USA

Synonyms

Disciplinary; Knowledge flow; Scholarly communication; Scholarly networks; Science maps; Scientific collaboration; Scientific evaluation; Topic identification

Glossary

Node (in Scholarly Networks) Entities such as words, papers, patent, authors, journals, institutions, fields, or country

Edge (in Scholarly Networks) Citation, co-citation, co-word, coauthor, bibliographic coupling, or hybrid relations

Scholarly Network The combination of edge properties and node properties defines a scholarly network

Macro-level Approach Statistics that are used to identify the global structural features of the networks, including component, bicomponent, shortest distance, clustering coefficient, degree distribution, and error and attack tolerance

Meso-level Approach Approaches that focus on the behavior of a group of actors, including topic identification and community detection

Microlevel Approach Indicators that are useful to understand individual node's power, stratification, ranking, and inequality in social structures, including centrality measures and PageRank and its variants

Introduction

In recent years, we have witnessed a growing trend of studying various types of networks, such as social networks, information networks, technical networks, and biological networks (Newman 2003). These studies were informed by the social studies of human interactions, were accelerated by the discovery of small-world and scale-free properties, and were also enriched by various macro-level statistics, meso-level clustering techniques, and microlevel indicators.

Studying characteristics of scholarly communication is crucial for understanding and exploration of reasons for better scientific innovation, scientific collaboration, and scientific activities in general. Scholars have used different types of networks to answer a wide spectrum of questions related to research interaction, scholarly communication, and science policy making; these efforts have greatly advanced the scholarship of scientometrics and informetrics. The earliest well-defined network in scholarly communication is probably the paper bibliographic coupling network, proposed by Kessler in the 1960s (Kessler 1963). Since then, various types of networks have been proposed and examined, for instance, co-citation networks, citation networks, coauthorship networks, co-word networks, and hybrid networks. For these networks, the paper is usually the basic research unit and can be aggregated into several higher levels, such as the author, journal, institution, and field level. Network types define edge properties and aggregation levels define node properties. The combination of edge properties (i.e., citation, co-citation, co-word, coauthor, bibliographic coupling, or hybrid) and node properties (i.e., words, papers, patents, authors, journals, institutions, fields, or country) precisely defines a network. Such networks are referred to as scholarly networks in this entry.

Various types of scholarly networks provide an ideal research instrument to quantitatively study scholarly communication. In particular, scholarly networks have been employed to study several essential aspects of scholarly communication: conducting scientific impact evaluation (primarily through citation networks), studying scientific collaboration (primarily through collaboration networks), identifying research specialties and topics (primarily through co-occurrence networks), and studying knowledge flow patterns (primarily through citation networks).

Scholarly Networks as a Type of Networks

In an important review article on complex networks, Newman (2003) distinguished four kinds of real-world networks: social networks (e.g., collaboration networks), information networks (e.g., citation networks), technical networks (e.g., Internet router networks), and biological networks (e.g., protein networks). Based on such division, two types of scholarly networks can be distinguished: social networks vs. information networks. In social networks such as coauthorship networks, a node is a social actor (i.e., an author); in information networks, a node is usually an artifact, such as a paper, a journal, or an institution.

In addition to "social networks vs. information networks," another distinction can be made, which is "real connection-based networks vs. similarity-based networks." Coauthorship networks and citation networks are constructed based on real connections, whereas co-citation, bibliographic coupling, topical, and co-word networks are constructed based on similarity connections. These scholarly networks can also be viewed from their edge types: collaboration-based, citation-based, or word-based. Citation-based scholarly networks include citation networks, co-citation networks, and bibliographic coupling networks; word-based scholarly networks include topical networks and co-word networks; collaboration-based networks include coauthorship networks. Those distinctions (social networks vs. information networks, real connection-based networks vs. similarity connection-based networks, citation-based networks vs. non-citation-based networks) are helpful to understand how different types of scholarly networks relate to each other.

Yan and Ding (2012) constructed six types of scholarly networks aggregated at the institution level and found that topic networks and coauthorship networks have the lowest similarity and these two types of networks set two boundaries (social and cognitive) for all six types; co-citation networks and citation networks have high similarity; bibliographic coupling networks and co-citation networks have high similarity; co-word networks and topical networks have high similarity.

The Use of Scholarly Networks

Before network theories were introduced to scientometrics, accumulative citation counting was widely used in the area of scientific evaluation. In the same vein of research, several citation-based indicators were proposed, such as Journal Impact Factor and *h*-index (Hirsch 2005). The accumulative citation counting and citation-based indicators equated all citations to have the same weight, without consideration of the citing papers, citing authors, or citing journals. This equal counting mechanism has been questioned, as scholars (e.g., Pinski and Narin 1976; Bollen et al. 2006; Yan et al. 2011) argued that it is more reasonable to differentiate the weight of citations based on the source of endorsement. This tension has largely been alleviated by the construction of different types of scholarly networks and the invention of various network-based bibliometric indicators. Comparing to traditional citation counting, scholarly networks have the advantage to consider the source of the citation endorsement. In this way, scholarly networks can capture the complex research communication and interaction more precisely.

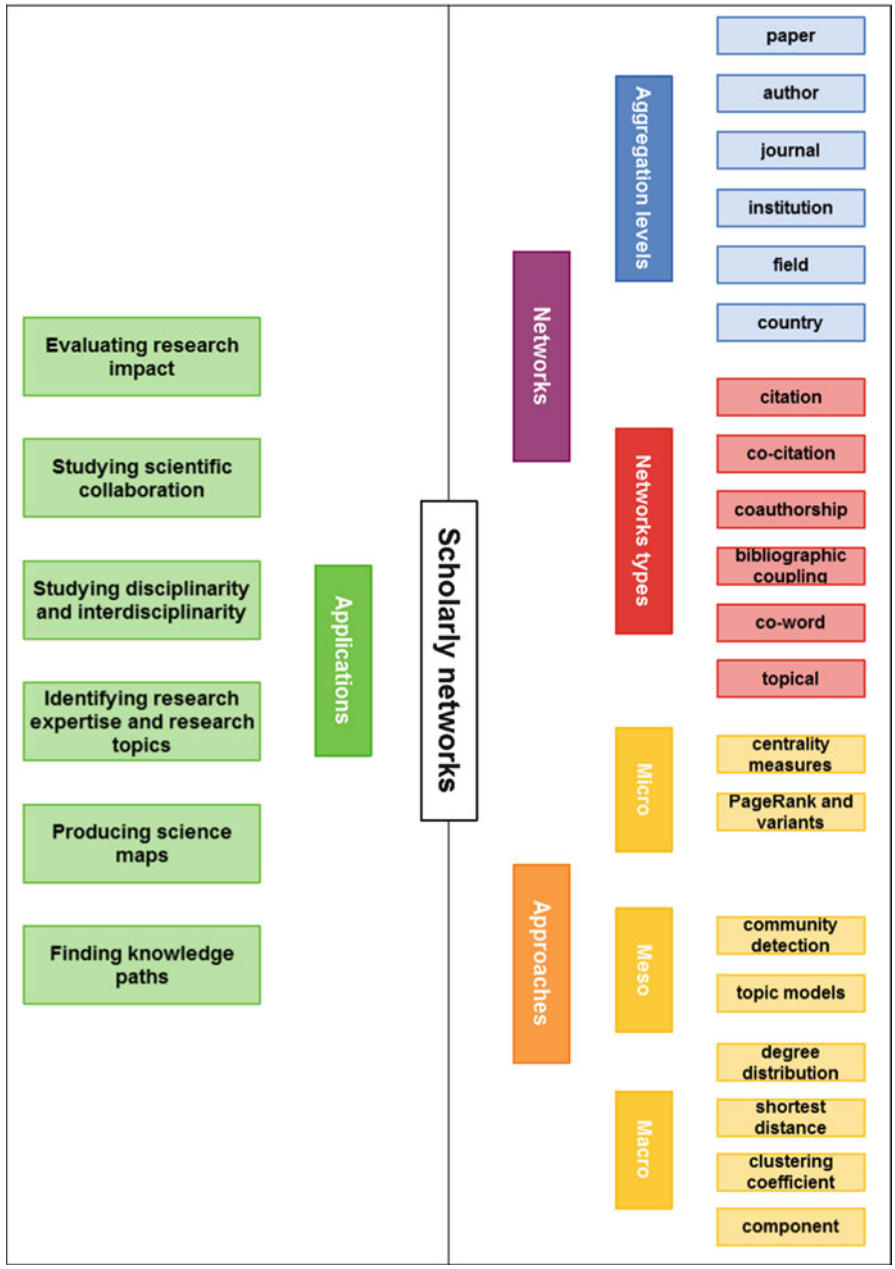
In addition to scientific evaluation, scholarly networks also contribute to other realms of scholarly communication and science policy making. For instance, coauthorship networks have been used to detect research communities and identify collaboration patterns (e.g., Newman and Girvan 2004); co-citation networks, bibliographic coupling networks, and co-word networks have been used to identify research specialties, examine interdisciplinarities, and map the backbone of science; and citation networks have been used to study knowledge flows and knowledge transfer in science and technology (e.g., Jaffe et al. 1993; Yan et al. 2013).

The Framework of Studying Scholarly Networks

Through scholarly network analysis, scientists and policy makers have gained unprecedented insights into the interaction of various research aggregates. The study of scholarly networks in general can be presented in a framework (Fig. 1), including approaches, network-network types, network-aggregation levels, and applications.

Approaches

Given that we have established a scholarly network, we can describe its properties on three levels, by macro-level metrics (global graph statistics), meso-level techniques (community characteristics), and microlevel metrics (individual actor properties). Macro-level metrics



Scholarly Networks Analysis, Fig. 1 A framework of scholarly network studies

seek to describe the global characteristic of a scholarly network as a whole with the aim to capture the generic structural features of a network. Commonly used measures include diameter, mean distance, components, and degree distribution. Meso-level techniques focus on identifying research communities and studying how communities interact with each other. Microlevel metrics relate to the analysis of the individual properties of network actors, for example, actor position, actor status, and distance to others, which informs us about "the differential constraints and opportunities facing individual actors which shape their social

behavior" (Yin et al. 2006, p. 1600). It zooms in to capture the features of the individual nodes/actors in a network with consideration of the topology of the network. Microlevel metric usually refers to centrality, which indicates how central the actor is to the network. Central actors are well connected to other actors, and metrics of centrality will measure an actor's degree (degree centrality), average distance (closeness centrality), or the degree to which geodesic paths between any pair of actors passes through the actor (betweenness centrality).

Macro-level Macro-level metrics are useful to identify the global structural features of the network. There are many ways of characterizing the structure of a network, such as component, bicomponent, k-core, mean distance, clustering coefficient, degree distribution, and error and attack tolerance of the network. In network analysis, connected graphs are called components.

- Component analysis can be used to learn about the macro-level structure of a network.
- In a bicomponent, no node can control the information flow between two other nodes completely because there is always an alternative path that information may follow (Nooy et al. 2005).
- The k-core of a network is a substructure in which each node has ties to at least k other nodes (Seidman 1983).
- A geodesic is the shortest path between two nodes.
- The degree of a node is the number of other nodes connected with it. Degree distribution measures the character of a network: a few nodes have many links and majority have smaller numbers of links.

Meso-level Meso-level scholarly network analyses focus on clustering various scholarly objects in the same groups based on certain clustering or modeling techniques. The clustering of papers, authors, institutions, journals, and subject categories is usually referred to as community detection; and the clustering of words and research topics is usually referred to as topic identification. Broadly perceived, clustering techniques fall into two branches: one yields discrete results where a node in a scholarly network is grouped into one or a couple of clusters; and the other branch yields fractional results where a node is grouped into clusters with certain probabilities. "Discrete" clustering techniques are traditional methods that include graph partitioning (e.g., Kernighan-Lin algorithm), hierarchical clustering, partitional clustering (e.g., k-means), and spectral clustering (e.g., algorithms utilizing Laplacian matrices). In this decade, more and more clustering tasks have used modularity-based methods that use modules to measure the strength of communities. "Fractional" clustering techniques use probabilistic models to assign papers, journals, or authors to clusters. The outcomes of topic models are probability distributions of words, papers, journals, or authors for each topic (e.g., Blei et al. 2003).

Micro-level Freeman (1979) elaborated four concepts of centrality in a social network, which have since been further developed into degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality. Eigenvector is based on the principle that the importance of a node depends on the importance of its neighbors. PageRank, on the other hand, is derived from the influence weights proposed by Pinski and Narin (1976); it is formally formulated by Brin and Page (1998), who developed a method for assigning a universal rank to Web pages based on a weight-propagation algorithm called PageRank. A page has high rank if the sum of the ranks of its backlinks is high. Actors in the PageRank of Web information retrieval systems are Web pages, and actors in the PageRank of coauthorship networks are authors. The underlying idea is that a citation from an influential publication, a prestigious journal, or a renowned author should be regarded as more valuable than a citation from an insignificant publication, an obscure journal, or an unknown author. It is sometimes argued that non-recursive indicators measure popularity and recursive indicators measure prestige.

Network Types

In addition to the different approaches, the interaction of research aggregates can be explored from different types of scholarly

networks. Each type of scholarly networks has its own use and can bring different perspectives to study research interaction and scholarly communication. For example, social networks such as coauthorship networks focus on finding collaboration patterns of contacts or interactions between social actors. Similarity-based networks such as co-citation networks, bibliographic coupling networks, and co-word networks focus on identifying research topics or schools of thoughts. In citation networks, each node is a piece of knowledge and a link denotes the knowledge flow.

Aggregation Levels

In these network types mentioned above, an article usually is a single research unit and can be aggregated into several higher levels. Figure 2 shows the different aggregation levels discussed in scholarly network studies.

The right side of the cascade is connected through "journal-ship" affiliation: a paper is published in a journal, a journal is classified into a subject category, and a subject category is further classified into a class. The left side of the cascade is connected through authorship affiliation: a paper is written by authors, an author is affiliated to an institution, and an institution is located in a country. Through studies of different research aggregates, we are provided with multiple focus lenses that allow us to zoom in and gain a concrete, detailed perspective on research interaction, while zooming out allows us to obtain a holistic and integrated view of the interacting institutions and disciplines.

Key Applications

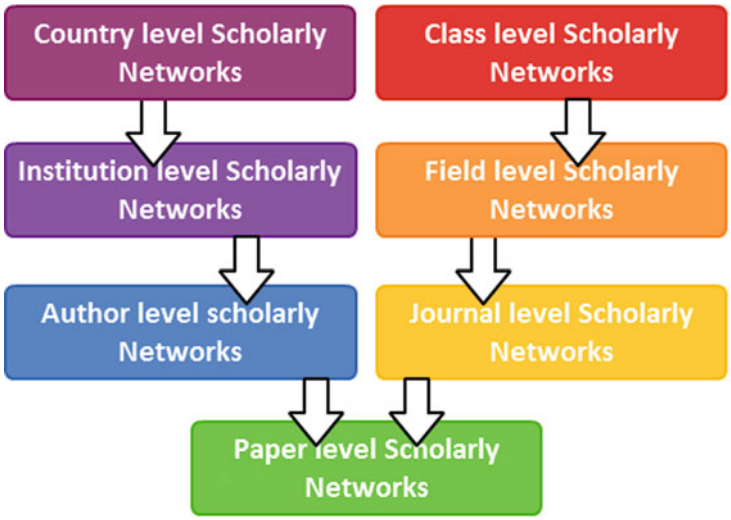
Scholarly networks have rich applications in the studies of scholarly communication and research interactions. Broadly perceived, six applications are apparent to us. In this section, brief introductions are given for each application.

Evaluating Research Impact

Impact evaluation has become an important issue in the science community. Scientists as well as policy makers now have a keen interest in evaluating scientific output. For scientists, evaluations of research impact help them find potential collaboration, discover new research topics, and locate appropriate venues to publish their work. For science policy makers, evaluations of research impact help inform them how to allocate research funds, promote emerging research fields, and monitor discipline developments. The traditional citation-based bibliometric indicators do not consider the source of the citation endorsement. However, in reality, being cited by a renowned author, a prestigious journal, and/or a highly influential paper differs from being cited by a remote author, a peripheral journal, and/or an obscure paper. Network-based bibliometric indicators are capable of considering the provenance of citation endorsement; specifically, PageRank and its variants have gained popularity in evaluating research impact. PageRank-like indicators denote a collection of algorithms based on Google's PageRank, such as Y-factor (Bollen et al. 2006), CiteRank (Walker et al. 2007), Eigenfactor (Bergstrom and West 2008), and SCImago Journal Rank (SCImago 2007). Among these network-based bibliometric indicators, citations are weighed differently depending on the status of the citing publication (e.g., Walker et al. 2007), the citing journal (e.g., Bollen et al. 2006; Pinski and Narin 1976), or the citing author (e.g., Radicchi et al. 2009).

Studying Scientific Collaboration

Scientific collaboration, as a large-scale real-world social phenomenon, has a particular charm to scientists and social scholars. Coauthorship networks provide an accurate and expedite medium, allowing scientists and scholars to explore various intriguing questions pertinent to this social phenomenon. Physicists and mathematicians have discovered the small-world and scale-free properties from coauthorship networks, for the first time providing a systematic inquiry into humans' social relationships. Later on, coauthorship networks have been used as a testing field for various modern clustering techniques (e.g., Newman and Girvan 2004).



Scholarly Networks Analysis, Fig. 2 Aggregation levels of scholarly networks

Such techniques are useful to examine scientific collaboration at a more granular level, providing insights to study the science of team science.

Studying Disciplinarity and Interdisciplinarity

The topic of interdisciplinarity has long been a research focus for social scientists. The quantitative study of interdisciplinarity has been enhanced by studying citation networks aggregated at the field level. Scholars usually chose some representative journals, or all journals from a field based on the ISI's classification of journals, and then measure the extent to which the publications of the chosen field cited the publications of other subject categories. Network-based indicators have also been proposed to measure how interdisciplinary disciplines are, using measures such as entropy (Zhang et al. 2010), integration and specialization (Porter et al. 2006), diversity and coherence (Rafols and Meyer 2010), and relative openness (Rinia et al. 2002).

Identifying Research Expertise and Research Topics

Human knowledge, in the form of scholarly publications, increases at a fast pace. How to effectively organize the expanding knowledge has become an important issue. Under such motivation, scholars have proposed various clustering techniques to group papers, authors, journals, institutions, and fields, with the aim to identify and organize research specialty in an effective way. For similarity-based scholarly networks such as co-citation networks and bibliographic coupling networks, the assumption is that if two research entities co-occurred frequently, then they are more likely to have similar characteristics. Therefore, co-occurrence networks can successfully achieve the goal of identifying and organizing scientific knowledge (e.g., White and McCain 1998; Boyack et al. 2005; Waltman et al. 2010).

Producing Science Maps

Clustering results can also be presented in science maps, and these maps are able to deliver richer and more informative messages to a broader audience body. Science maps on author and journal interactions are usually used to identify research topics (e.g., Boyack et al. 2005). As institutions are associated with geographical locations, science maps at the institution level are useful to illustrate the geographical distribution of scientific productivity (e.g., Leydesdorff and Persson 2010). Science maps at the field level provide a unique view on the backbone of science (e.g., Boyack et al. 2005) or on the knowledge flow in scientific disciplines (e.g., Rosvall and Bergstrom 2008).

Finding Knowledge Paths

The production and creation of knowledge is not dependent on a single isolated entity; instead, knowledge is diffused, exchanged, and circulated among various entities. Knowledge flow, in the past 20 years, is becoming more inter-sectoral, more interorganizational, more interdisciplinary, and more international. The issues of how do scientific and technological knowledge, innovative ideas, management skills, or certain influences transfer within different sectors, between different organizations, and between different scientific disciplines are pertinent to understanding patterns of knowledge transfer and dissemination. Citation networks serve as an ideal research instrument to uncover such patterns. In citation networks, a node is a research aggregate, and a link denotes a citation from the citing research aggregate to the cited research aggregate.

Future Directions

Studies on scholarly networks usually chose one type of network at one aggregation level. The choice of a type of network can be inconsistent or even arbitrary, and the findings have been discrete and cannot be generalized to address a wider spectrum of research questions. We recommend that, in order to capture varied aspects of research interactions, different types of networks need to be combined and thus form a hybrid network. Beyond hybrid approaches, scholars have proposed heterogeneous scholarly networks to incorporate different academic entities while keeping edge semantics. Study of the heterogeneous networks has evolved from bi-typed networks to star-typed heterogeneous networks. By adding more academic entities (e.g., authors, journals, articles, words), heterogeneous networks can better simulate the mutual engagement of various academic entities in the complex academic environment.

Therefore, future research on this topic would benefit from (1) constructing hybrid and heterogeneous scholarly networks and (2) evaluating different approaches on hybrid networks or heterogeneous scholarly networks through possible "golden standards" (such as award lists or expert judgments) in order to determine which approach can yield more precise clustering results and more useful information for scientific evaluations.

Cross-References

- ▸ Centrality Measures
- ▸ Components of the Network Around an Actor
- ▸ Similarity Metrics on Social Networks
- ▸ Social Interaction Analysis for Team Collaboration

References

Bergstrom CT, West JD (2008) Assessing citations with the Eigenfactor™ metrics. Neurology 71(23):1850–1851

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3(4–5):993–1033

Bollen J, Rodriguez MA, Van de Sompel H (2006) Journal status. Scientometrics 69(3):669–687

Boyack KW, Klavans AR, Börner K (2005) Mapping the backbone of science. Scientometrics 64(3):351–374

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst 30(1–7):107–117

Freeman LC (1979) Centrality in social networks: conceptual clarification. Soc Netw 1(3):215–239

Hirsch JE (2005) An index to quantify an individual's scientific research output. Proc Natl Acad Sci USA 102(46):16569–16572

Jaffe AB, Trajtenberg M, Henderson AD (1993) Geographical localization of knowledge spillovers by patent citations. Q J Econ 108(3):577–599

Kessler MM (1963) Bibliographic coupling between scientific papers. Am Doc 14(1):10–25

Leydesdorff L, Persson O (2010) Mapping the geography of science: distribution patterns and networks of relations among cities and institutes. J Am Soc Inf Sci Technol 61(8):1622–1634

Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256

Newman M, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev E 69(2):026113

Nooy W, Mrvar A, Batagelj V (2005) Exploratory social network analysis with Pajek. Cambridge University Press, Cambridge

Pinski G, Narin F (1976) Citation influence for journal aggregates of scientific publications: theory, with

application to the literature of physics. Inf Process Manag 12(5):297–312
Porter AL, Roessner JD, Cohen AS, Perreault M (2006) Interdisciplinary research: meaning, metrics and nurture. Res Eval 15(3):187–195
Radicchi F, Fortunato S, Markines B, Vespignani A (2009) Diffusion of scientific credits and the ranking of scientists. Phys Rev E 80(5):056103
Rafols I, Meyer M (2010) Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. Scientometrics 82(2): 263–287
Rinia EJ, Van Leeuwen TN, Bruins EEW, Van Vuren HG, Van Raan AFJ (2002) Measuring knowledge transfer between fields of science. Scientometrics 54(3): 347–362
Rosvall M, Bergstrom CT (2008) Maps of information flow reveal community structure in complex networks. Proc Natl Acad Sci USA 105(4):1118–1123
SCImago (2007) SJR: SCImago Journal & Country Rank. http://www.scimagojr.com. Retrieved 31 Aug 2009
Seidman SB (1983) Network structure and minimum degree. Soc Netw 5:269–287
Walker D, Xie H, Yan KK, Maslov S (2007) Ranking scientific publications using a simple model of network traffic. J Stat Mech: Theory Exp P06010. doi:10.1088/1742-5468/2007/06/P06010
Waltman L, Van Eck NJ, Noyons ECM (2010) A unified approach to mapping and clustering of bibliometric networks. J Informetr 4(4):629–635
White HD, McCain KW (1998) Visualizing a discipline: an author co-citation analysis of information science 1972–1995. J Am Soc Inf Sci 49(4):327–355
Yan E, Ding Y (2012) Scholarly network similarities: how bibliographic coupling networks, citation networks, co-citation networks, topical networks, coauthorship networks, and co-word networks relate to each other. J Am Soc Inf Sci Technol 63(7):1313–1326
Yan E, Ding Y, Sugimoto CR (2011) P-Rank: an indicator measuring prestige in heterogeneous scholarly networks. J Am Soc Inf Sci Technol 62(3): 467–477
Yan E, Ding Y, Cronin B, Leydesdorff L (2013) A bird's-eye view of scientific trading: dependency relations among fields of science. J Informetr 7(2):249–264
Yin L, Kretschmer H, Hanneman RA, Liu Z (2006) Connection and stratification in research collaboration: an analysis of the COLLNET network. Inf Process Manag 42(6):1599–1613
Zhang L, Liu X, Janssens F, Liang L, Glänzel W (2010) Subject clustering analysis based on ISI category classification. J Informetr 4(2):185–193

Science Maps

▸ Scholarly Networks Analysis

Science of Science (Sci2) Tool

▸ Plug-and-Play Macroscopes: Network Workbench (NWB), Science of Science Tool (Sci2), and Epidemiology Tool (EpiC)

Science of the Internet

▸ Web Science

Science of the Web

▸ Web Science

Scientific Collaboration

▸ Scholarly Networks Analysis

Scientific Communities

▸ Stability and Evolution of Scientific Networks

Scientific Evaluation

▸ Scholarly Networks Analysis

Search Engine

▸ Weblog Analysis

Search Missions

▸ Weblog Analysis

Security

▸ Reconnaissance and Social Engineering Risks as Effects of Social Networking

Self-Confidence

▸ Self-Efficacy vs. Expertise

Self-Disclosure

▸ Privacy and Disclosure in a Social Networking Community
▸ Reconnaissance and Social Engineering Risks as Effects of Social Networking

Self-Efficacy vs. Expertise

Donghee Yvette Wohn and Chandan Sarkar
Department of Telecommunication, Information Studies and Media, Michigan State University, East Lansing, MI, USA

Synonyms

Self-confidence; Self-perception

Glossary

Self-efficacy An individual's confidence about his or her skills
Expertise Knowledge and actual skills

Definition

Self-efficacy (Bandura 1977a,b) is an individual's self-perception of his or her ability. By placing importance on the individual's perception as opposed to the individual's actual skill, this construct can explain why people have different behaviors even if they have a similar skill set. In much of social-psychological research, self-efficacy serves as a good proxy of predicting people's behaviors because it looks not only at perceived expertise (knowledge and actual skills) about a certain behavior but also perceived confidence. However, in the context of behaviors required to ensure privacy and security in an online environment, confidence in one's ability may not necessarily be the best factor that explains behavior. Because privacy behaviors, such as changing privacy settings and employing preventive security measures, require a certain degree of technical expertise, *perceived* expertise and *perceived* confidence can be false indicators. For example, one may have extremely strong confidence in one's ability but could very well be overestimating that ability. In the context of Internet privacy, this differentiation is important as privacy protection requires certain technical skills.

Efficacy beliefs are the product of a complex process of self-persuasion that rely on cognitive processing of diverse sources of information. Research has found that self-efficacy is an important construct that explains an individual's attitude about privacy, which ultimately affects their behavior (Rifon et al. 2005). A limitation for studies that examine self-efficacy, however, is that there was little consideration for actual expertise.

In semi-structured in-depth interviews with young adults aged 20–30, we found that the reality of how users process information and behave in relation to privacy and security issues online was sometimes inconsistent with their self-efficacy, especially among those with low levels of expertise. Expertise was determined by an individual's actual understanding of the technical aspects of privacy issues through a set of questions to participants answered explaining constructs such as phishing, Trojans, cookies, and browser privacy settings. Self-efficacy was measured asking participants if they consider themselves knowledgeable about the security risks and threats to privacy that exist online. We found that the reality

of how users process information and behave in relation to privacy and security issues online was sometimes inconsistent with their self-efficacy, especially among those with low levels of expertise. In many cases, there was a difference between an individual's self-efficacy and expertise; novices were more likely to overestimate their self-efficacy while experts were more likely to underestimate it. It was expertise, not self-efficacy, that was a stronger predictor of their attitude and behavior. For example, participants with high self-efficacy but low expertise showed extreme caution and concern regarding the implication of privacy policies and were more upset about behavioral targeting than those with high knowledge, whereas participants with higher self-efficacy and high expertise showed accepting behavior towards behavioral targeting.

Individuals' abilities to change privacy settings, set up firewalls, and use security software among others were key indicators to how they perceived privacy issues and how they acted to deal with those issues. Novices were fearful of privacy and security threats and relied more on peripheral cues such as privacy seals and brand names to make their judgment. Experts were less concerned about such threats and interpreted peripheral cues differently from novices. Across different topics, we consistently saw differences between those who had expertise and those who did not.

Although this seems to challenge studies that show self-efficacy as a strong predictor of behavior, it may be that there are different dimensions of self-efficacy. For example, self-efficacy of behavior (how to run antivirus software) may be high, but self-efficacy of underlying concepts or mechanisms (how the software works) may be low. Thus, from a theoretical perspective, we suggest that in the context of privacy studies, researchers measure individuals' actual expertise in addition to self-efficacy, as expertise may play a moderating role in predicting behavior. From a practical perspective, the distinction between self-efficacy and expertise can inform how and what we should teach people about privacy and security issues. People using social networks face many challenges in terms of privacy and security threats. There are no regulations in terms of legal enforcement to what extent personal information can be gathered and used by the social network providers for purposes such as behavioral targeting. The average user has uncertainty about what kind of personal data is collected and for what purpose, let alone where that information is being sold or how they should protect themselves. Furthermore, most social network services collect network data, in which case information that a user reveals to another thinking it is private could still end up being collected and sold to third parties. Differentiating the true expertise from the self-perception of self-efficacy may enable us to identify individuals who are at high risk – those who think they know much but actually don't.

Cross-References

▸ Privacy in Social Networks, Current and Future Research Trends on
▸ User Behavior in Online Social Networks, Influencing Factors

References

Bandura A (1977a) Self-efficacy: toward a unifying theory of behavioral change. Psychol Rev 84(2):191–215
Bandura A (1977b) Self-efficacy: the exercise of control. Freeman, New York
Rifon NJ, LaRose R, Choi SM (2005) Your privacy is sealed: effects of web privacy seals on trust and false assurances. J Consum Aff 39(2):337–360

Self-Perception

▸ Self-Efficacy vs. Expertise

Self-Presentation

▸ Privacy and Disclosure in a Social Networking Community

Semantic Networks

▸ Combining Online Maps with Text Analysis
▸ Semantic Social Networks Analysis

Semantic Perception

▸ Twitris: A System for Collective Social Intelligence

Semantic Social Networks

Peter Gloor[1] and Jana Diesner[2]
[1]Center for Collective Intelligence, MIT, Cambridge, MA, USA
[2]The iSchool, University of Illinois at Urbana-Champaign, Champaign, IL, USA

Synonyms

Context networks; Ontologies; Taxonomies; Text networks; Topic networks

Glossary

Semantic Network Structured representations of knowledge that are used for reasoning and inference
RDF Resource Description Framework
FOAF friend of a friend
NLP Natural Language Processing

Introduction

Semantic networks represent the relationships between pieces of knowledge or information. They were originally designed to be used for performing inference and reasoning on the meaning of these data (Sowa 1992; Woods 1975). Social networks represent the interactions between social agents, typically people and organizations (Freeman 2004). The analysis of semantic networks and social networks are both active areas of research and innovation, while work at their intersection is less prevalent. However, it has been long recognized that combining both types of network data enables researchers and practitioners alike to ask more advanced yet highly relevant questions such as:

- Who is talking to whom (social network) about what (semantic network) (Danowski 1993; McCallum et al. 2007)?
- Does shared knowledge or interest in similar topics (semantic network) increase the likelihood of becoming acquaintances (social networks), or vice versa (Crandall et al. 2008; Wenger 1999)?
- How do information, opinions, and rumors (semantic network) emerge, spread, and vanish in society and on social networking sites (social network) (Adar and Adamic 2005; Leskovec et al. 2009)?
- Can social network data be made more accurate, complete, and useful by exploiting background information on social agents (Berners-Lee et al. 2001; Van Atteveldt 2008)?

The last one of these questions originates from a more specific use of the concept of "semantic social networks": in addition to referring to the combination of semantic and social networks, this term also describes the enhancement of relational data with background information on any type of node or entity. Again, these enhanced graphs can then be used for conducting inference and reasoning over the data. One prominent example for this approach is the semantic web (Berners-Lee et al. 2001). The key idea with the semantic web is to mark up data objects on the Web, such as words and relations, as they occur on webpages, by using a standardized annotation language (Resource Description Framework, or RDF). These enhanced data structures are then used to generate machine-readable definitions of data that can be interpreted by computers. An example for a stream of work that originates from the Semantic Web philosophy is the "friend

of a friend" (FOAF) framework. FOAF combines social network data with semantic network data, where both types of data structures are denoted according to a predefined, machine-readable description language, allowing for automated inference on the data. On a general level, such data can be used to recommend social ties between people who share some interests or are involved in the same events or to recommend activities and pieces of information that are new to a person whose friends have endorsed these things.

Historical Background on Semantic Social Networks

Efforts in combining social and semantic networks trace back their roots well before the advent of the computer and the Internet. Vannevar Bush (1945), advisor to US president Roosevelt during the Second World War, envisioned Memex, a device for organizing all knowledge of mankind in a structured way. In the 1960s, Ted Nelson coined the term "hypertext." In the 1970s and 1980s, a vibrant field around the concept of hypertext emerged in disciplines such as Artificial Intelligence and computer science. In the 1980s and 1990s, the ACM annual hypertext conference regularly drew between 500 and 1,000 researchers. Early on these researchers started combining hypertext with semantic networks (Brachman 1979). In 1991 Tim Berners-Lee, together with Robert Cailliau, presented the Web at the ACM hypertext conference in San Antonio, and just a few years later he broadened his concept to the Semantic Web (Berners-Lee et al. 2001). Shortly after, Hermann Maurer and his colleagues envisioned a hypermedia system called Hyper-G, which combines semantics and social networking (Andrews et al. 1995).

Combining Social and Semantic Networks

Combining social networks and semantic networks opens up novel ways for extracting meaning from social interaction. One strategy for combining social networks with semantics about the network data is to enhance a given social network with additional information about agents and their connections. This can be done by exploiting external data sources such as the Web, news archives, and domain-specific databases (Van Atteveldt 2008). For example, for the individuals being co-mentioned in a news article, one could search knowledge bases such as Wikipedia or the Web for further information on those people's roles and locations. Adding this information to the social network contextualizes the data, which allows for a richer and more fine-grained indexing and retrieval. This approach basically adds semantics to a social network. It can also help to disambiguate social agents, e.g., people who have the same name, but differ in their job, location, or date of birth as indicated on knowledge bases such as Wikipedia, and to specify the types of relationships between agents. The inverse of this principle, i.e., utilizing, searching, or suggesting connections between people who share some knowledge, interests, or likings, is also an active area of research and development. Examples for this approach include recommender systems, dating services, and social networking platforms.

Another common strategy for bringing together social networks with semantic information is to enhance a social network with the information produced, processed, or shared by members from within or outside the network (Diesner and Carley 2011). This information typically represents salient information from natural language text data, such as people's interests as indicated on their social networking profile or key terms and themes that are explicitly or implicitly contained in documents that people authored. For communication data, for instance, a social network can be built from the explicit information about communication partners (who talks to whom). Then, agents can be linked to nodes representing words and short phrases that occur with a high (weighted) frequency in the underlying text data. Suitable data sources used for this procedure include transcripts of conversations and meetings and online discussion forums. An early example for a tool that jointly analyzed the social network of e-mail senders

and receivers *and* the content of e-mail bodies is TeCFlow, now called Condor (Gloor and Zhao 2004). Today, a variety of methods and tools has been developed that support the joint collection, visualization, and analysis of relational social and semantic information from communication archives. Remaining challenges in this domain include the partitioning of people into meaningful groups prior to associating social clusters with content and selecting pieces of information to link to agents in a scalable yet non-arbitrary fashion (Diesner 2013).

Sometimes, the information about social networks is encoded in sources that are typically used for conducting semantic analysis, namely, unstructured, natural language text data. In these cases, social network data can be extracted from the text data (Diesner and Carley 2011; Roth and Yih 2002). Typical data sources include news wire data, interviews, communication data, and social media data, such as microblogging services and social networking sites. The main steps involved in this task are the identification of entities, i.e., nodes, and the relations between them. These entities are sometimes further categorized into different classes, such as people, organizations, and locations, and can entail one-word units as well as multi-word units (Diesner and Carley 2008). The types of relations can be defined over entity types, such as social networks between people or a membership network between people and organizations. Alternatively, applicable types of relations can be specified in an ontology or a taxonomy, which can be predefined or extracted from the data (Brin 1999; Roth and Yih 2002). An example would be to classify social network ties as representing friendship or antagonism. Highly accurate, automated, and scalable methods for relation extraction typically exploit a combination of lexical (words and their structure), semantic (meaning of words), syntactic (relationships between words and grammar), and statistical information from text data (Diesner and Carley 2008; Mihalcea and Radev 2011). These methods, which have been developed in the fields of Natural Language Processing and Computational Linguistics, typically combine routines from statistics and machine learning and sometimes also consider models and methods from socio-linguistics and sociology (Corman et al. 2002; Diesner and Carley 2008). Once such network data have been extracted from a text corpus, they can serve as input to regular network analysis (Carley et al. 2007; Corman et al. 2002). Used this way, relation extraction can serve as a complementary or alternative method for collecting data about social networks. These social networks can be combined with semantic networks that are also extracted from the text data. In fact, some models and methods consider "knowledge" or "information" as a node classes for relation extraction (Diesner and Carley 2008). Combining social networks and semantic networks extracted from text data or built from other sources can be useful for addressing the following types of questions (Barthelemy et al. 2005; Carley et al. 2007; Gloor et al. 2009):

- Which social agents are associated with what ideas, beliefs, or pieces of knowledge?
- Which agents are prominent with respect to their association with information? These people might function as information brokers or gatekeepers if they have a high betweenness centrality or be somewhere between well informed and overloaded with information if they feature a high degree centrality (for details on these metrics, see the section on centrality measures).
- Which agents are linked to too many knowledge items and thus might suffer from task overload?
- Which agents have exclusive access to some information? In an organizational context, such people might represent a vulnerability, which can be mitigated by converting tacit knowledge residing in people to information being documented in written form.

One caveat with distilling the network data from text data itself is that research on resembling ground truth data for social networks by exploiting the substance of text data has shown that the overlap between text-based social networks, e.g., those extracted from e-mail bodies, and social

networks constructed from associated metadata, e.g., e-mail headers, shows only minimal overlaps (Diesner 2013).

Key Applications

One prominent example for employing semantic social networks in practice is expert finder systems (Dooley et al. 2002). A specific instance is SmallBlue, a tool developed and deployed at IBM (Ehrlich et al. 2007). This system augments social network data with information on who knows what, allowing people to search for the shortest social path to knowledge through their wider and potentially remote network of coworkers. In such systems, the information on people's expertise can be pulled from internal data sources, such as organizational databases, and from public sources, such as blog posts and tags. Ehrlich et al. evaluated the SmallBlue system to be particularly useful for locating experts on very particular pieces of knowledge, which complements the general understanding about broadly regarded experts on certain topics. Generalizing from this idea, professional social network sites such as LinkedIn are based on the same premise: they allow people to search for experts on certain topics within the professional network of their immediate acquaintances. If a match has been found but the identified individual is not a contact of the person executing the search, one could mobilize their social capital to be introduced via the shortest social path of mutual acquaintances.

Another real-world example for semantic social networks is Wikipedia (Brandes et al. 2009; Crandall et al. 2008). This knowledge base not only provides a vast amount of socially vetted information but also entails metadata about the authors and detailed information about every single contribution. The metadata surrounding the content pages entail information about the *what* – the edits of pages, *when* – edits over time, *who* – which authors edited the pages, and *how* – which links to other pages inside and outside of Wikipedia. These types of data can be fused into dynamic, multimodal network data. Moreover, taken together, these data open new opportunities for investigating the processes that lie behind the life cycle of the creation of content and contributing knowledge to the public domain. Visualizations built on this information can provide maps of concepts, knowledge, and trends, which can be displayed by content domain, geophysical region, cultural background, etc. Analyzing and comparing these maps and semantic networks across time, space, and languages can contribute towards a better understanding of societies and cultures. In addition, constructing coauthorship networks where links between Wikipedia articles are drawn based on the same person editing different articles enables the identification of domain experts as well as trusted arbitrators. Furthermore, Wikipedia has an implicit social organization of its own, composed of networks of contributors. Analyzing this social network can help to understand if active Wikipedians operate under an implicit set of rules that has evolved within the Wikipedia community and might generalize to other open source production systems or to traditional organizations. Studying collaboration among Wikipedians also gives indications of the role of social capital for teams in organizations where members are collaborating virtually without much face-to-face contact. In the same way that social network surveys made visible the importance of the informal organization within large corporations, the analysis of Wikipedia editor networks enables the measuring of the role of social capital in voluntary online collaboration: social capital indeed seems to increase efficiency in this emerging organizational setting as well (Nemoto et al. 2011).

Future Directions

An example for ongoing research on semantic social networks is the measurement of team performance over time. In this work the performance and creativity of organizations are analyzed by correlating social network data, content-based

semantic analyses, and creative performance of work teams (Grippa et al. 2012). This approach is a first step towards articulating a systematic theory of social networks coming from the dynamic and causal dimensions of relationships. This work represents the general idea behind semantic social networks, namely, enabling the investigation of ties among community members not only under the quantitative aspect related to SNA metrics but also under the qualitative aspect related to the content of the ties. Such an emergent theory will give new meaning to the "relational" and "cognitive" dimensions of social capital (Stinchcombe 1990).

Cross-References

- ▸ Analysis and Mining of Tags, (Micro)Blogs, and Virtual Communities
- ▸ Combining Online Maps with Text Analysis
- ▸ Ontology Matching
- ▸ Recommender Systems, Semantic-Based
- ▸ Social Network Analysis and Company Linguistic Identity
- ▸ Social Web Search
- ▸ Sources of Network Data

References

Adar E, Adamic L (2005) Tracking information epidemics in blogspace. Paper presented at the 2005 IEEE/WIC/ACM international conference on web intelligence, Compiegne, Sept 2005

Andrews K, Kappe F, Maurer H (1995) Serving information to the web with Hyper-G. Paper presented at the third international world-wide web conference, computer networks and ISDN systems

Barthelemy M, Chow E, Eliassi-Rad T (2005) Knowledge representation. Issues in semantic graphs for relationship detection. Paper presented at the AAAI spring symposium on AI Technologies for Homeland Security, Stanford, http://pages.cs.wisc.edu/~eliassi/chow-aaai-ss2005.pdf

Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. Sci Am 284(5):34–43

Brachman RJ (1979) On the epistemological status of semantic networks. In: Findler NV (ed) Associative networks: representation and use of knowledge by computers. Academic, New York, pp 3–50

Brandes U, Kenis P, Lerner J, Van Raaij D (2009) Network analysis of collaboration structure in Wikipedia. Paper presented at the 18th international conference on world wide web, Madrid

Brin S (1999) Extracting patterns and relations from the world wide web. Paper presented at the the world wide web and databases, Valencia, 27–28 Mar 1998

Bush V (1945) As we may think. Atl Mon 176(1):101–108

Carley KM, Diesner J, Reminga J, Tsvetovat M (2007) Toward an interoperable dynamic network analysis toolkit. Decis Support Syst 43(4):1324–1347. Special Issue Cyberinfrastructure for Homeland Security

Corman SR, Kuhn T, Mchee RD, Dooley KJ (2002) Studying complex discursive systems: centering resonance analysis of communication. Hum Commun Res 28(2):157–206

Crandall D, Cosley D, Huttenlocher D, Kleinberg J, Suri S (2008) Feedback effects between similarity and social influence in online communities. Paper presented at the proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, Las Vegas

Danowski JA (1993) Network analysis of message content. Prog Commun Sci 12:198–221

Diesner J (2013) From Texts to Networks: Detecting and Managing the Impact of Methodological Choices for Extracting Network Data from Text Data. Künstliche Intelligenz/ Artificial Intelligence. 27(1):75–78. DOI: 10.1007/s13218-012-0225-0

Diesner J, Carley KM (2008) Conditional random fields for entity extraction and ontological text coding. J Comput Math Org Theory 14:248–262. doi:10.1007/s10588-008-9029-z

Diesner J, Carley KM (2011) Words and networks. In: Barnett G, Golson JG (eds) Encyclopedia of social networking. Sage, Thousand Oaks, pp 958–961

Dooley KJ, Corman SR, Mchee RD (2002) A knowledge directory for identifying experts and areas of expertise. Hum Syst Manag 21(4):217–228

Ehrlich K, Lin C, Griffiths-Fisher V (2007) Searching for experts in the enterprise: combining text and social network analysis. Paper presented at the 2007 international ACM conference on supporting group work, Sanibel Island

Freeman LC (2004) The development of social network analysis. Empirical Press, Vancouver

Gloor P, Zhao Y (2004) TeCFlow – a temporal communication flow visualizer for social networks analysis. Paper presented at the ACM CSCW conference, workshop on social networks, Chicago

Gloor P, Krauss J, Nann S, Fischbach K, Schoder D, Switzerland B (2009) Web science 2.0: identifying trends through semantic social network analysis. Sanibel Island Paper presented at the IEEE conference on social computing (SocialCom-09), Vancouver

Grippa F, Palazzolo M, Bucuvalas J, Gloor P (2012) Monitoring changes in the social network structure of clinical care teams resulting from team development efforts. Int J Org Des Eng 2(2):149–166

Leskovec J, Backstrom L, Kleinberg J (2009) Meme-tracking and the dynamics of the news cycle. Paper presented at the 15th ACM SIGKDD international conference on knowledge discovery and data mining
Mcallum A, Wang X, Corrada-Emmanuel A (2007) Topic and role discovery in social networks with experiments on Enron and academic email. J Artif Intell Res 30:249–272
Mihalcea RF, Radev DR (2011) Graph-based natural language processing and information retrieval. Cambridge University Press, Cambridge
Nemoto K, Gloor P, Laubacher R (2011) Social capital increases efficiency of collaboration among Wikipedia editors. Paper presented at the HT'11 22nd ACM conference on hypertext and hypermedia
Roth D, Yih W (2002) Probabilistic reasoning for entity and relation recognition. Paper presented at the international conference on computational linguistics (COLING), Taipei
Sowa J (1992) Semantic networks. In: Shapiro SC (ed) Encyclopedia of artificial intelligence, 2nd edn. Wiley, New York, pp 1493–1511
Stinchcombe AL (1990) Information and organizations. University of California Press, Berkley
Van Atteveldt W (2008) Semantic network analysis: techniques for extracting, representing, and querying media content. BookSurge Publishers, Charleston
Wenger E (1999) Communities of practice: learning, meaning, and identity. Cambridge University Press, New York
Woods W (1975) What's in a link: foundations for semantic networks. In: Bobrow D, Collins A (eds) Representation and understanding: studies in cognitive science. Academic, New York, pp 35–82

Semantic Social Networks Analysis

Christophe Thovex[1], Francky Trichet[1], and Bénédicte Le Grand[2]
[1]Laboratory of Computer Sciences (LINA, UR CNRS 6241), University of Nantes, Nantes, France
[2]Centre de Recherche en Informatique, Université Paris 1 Panthéon-Sorbonne, Paris, France

Synonyms

Detection of communities; Graph mining; Knowledge engineering; Networks dynamics analysis; Semantic networks; Social trends discovery; Text mining

Glossary

SNA Social network analysis (see *Definition* section)
SW Semantic Web (see *Historical background* section)
Knowledge Engineering Discipline studying, extracting, and managing knowledge implicitly defined within digital data structures
Graph Mining Extracting implicit information and knowledge from graphs
Data Mining Extracting implicit information and knowledge from numeric data
Text Mining Extracting implicit information and knowledge from text corpora
Social Capital Knowledge and skills owned by employees (human capital) when shared in a collaborative context and defining a network of professional interactions

Definition

Social networks analysis (SNA) enables to figure out the position of people and communities within social networks, represented as social graphs. It defines a set of methods and measures, such as graph clustering for community detection or closeness centrality and betweenness centrality, which identify and rank members or communities based on the statistical analysis of the connections found in these social graphs. When these kinds of methods and measures also take into account the semantics of the digital content shared within social networks or semantic information about people, SNA turns into **semantic social networks analysis** (SSNA).

Introduction

Standard SNA measures mostly consider ties and relationships within social networks and thus remain blind to the semantics of the digital content shared by their members and/or implicitly expressed by their profiles. Therefore, searching opinion leaders within a planetary network such as Facebook or MSN using SNA measures

generally returns the most mediatized people, whether they are journalists, politicians, or international artists.

Indeed, most SNA methods and measures are based on the statistical analysis of social graphs topology (Freeman et al. 1989). A graph $G(V, E)$ is a set of vertices and a set of edges. Each element of $G(V, E)$ is possibly weighted and/or labeled with one or more values. As a result, we find structure-based measures, such as the stress centrality defined in Shimbel (1953), and flow-based measures (Newman 2005) for undirected or directed graphs, integrating various metrics such as information flows or virality (Brandes and Fleischer 2005; Miramontes and Luque 2002).

Semantic SNA is mostly based on interdisciplinary models merging SNA and knowledge engineering (KE). On the one hand, it refines SNA measures and metrics in order to enhance the processing of data, text, and knowledge tied to the members of social networks. On the other hand, it refines KE principles, techniques, and methods such as linguistic statistics and ontologies, so as to provide KE capabilities adaptable to SNA models. Semantic SNA measures thus make it possible to retrieve opinion leaders within a large network, for specific topics or keywords. For instance, the semantic betweenness centrality defined in Thovex and Trichet (2012) enables to retrieve polyvalent experts in specific domains of professional activity defined by seized keywords, such as "database administration and website management," even if managers are much more connected and relay more communications than technical experts within the enterprise social network.

Key Points

This essay proposes an insight of theoretical aspects of semantic social networks analysis, of its epistemic extents, and of the applications it enables to develop. Based on the state of the art in the domain, we study the theoretical foundations of SSNA from their graphic aspects such as topology and flows, or static and dynamic behavior within social networks, to their KE aspects such as data mining, text mining, graph mining, or ontologies and the semantic Web. The main theoretical aspects are illustrated with an application of SSNA for enterprises and with examples of applications impacting our social life, economic life, professional life, and private life. Lastly, the future and theoretical directions of SSNA are presented under the epistemic aspects of the way paved by SSNA foundations, and future applicative directions are explored in terms of social, economic, and strategic outcomes.

Historical Background

Some premises of SNA have appeared in Moreno (1934) with the notions of sociogram and sociometry, then in Freeman et al. (1960), concerning the study of social relationships and leadership in communities. Introduced as a sociological discipline, SNA started to have recourse to mathematics and statistics to develop new measures adapted to large-scale analysis, mostly centrality and modularity measures (Shimbel 1953; Freeman 1977). These measures and metrics are now considered as standard in SNA, and while sociologists carry on studying socialization and group behaviors (Tajfel et al. 1971), the theoretical foundations of standard measures continually inspire new refinements (Brandes 2001; Miramontes and Luque 2002; Pearson and West 2003), so as to face the new challenges raised by the planetary networks of the social Web – e.g., Twitter, Facebook, MSN, and Orkut. As the Web started to be semantic before being social, semantic SNA is currently becoming a mainstream in SNA (Erétéo 2011; Thovex and Trichet 2012).

Increasing interest in Web information retrieval led to the semantic Web initiative (Berners-Lee et al. 2001) from the World Wide Web Consortium in 2001. Semantic standards have been widely used since then, even outside the scope of the Web. However, the main limit in the use of such techniques is the need for explicit semantics from users as fully automatic semantic annotation is not possible. The semantic Web is well fitted to merge with the social Web as both content and actors are generally strongly tied,

through semantics-to-content relationships and members-to-members relationships (e.g., finding appropriate files/endpoints within peer-to-peer networks). Defining enhanced capabilities for SNA and taking advantage of the semantic Web, semantic social network analysis now tends to provide decisional models for the semantic and social Web.

Semantic SNA Is Interdisciplinary

Sociology, mathematics, knowledge engineering, these three disciplines summarize the interdisciplinary aspect of semantic social networks analysis. In this section, we present SSNA from the standpoint of computer sciences, focusing on the theoretical aspects of standard SNA models then on knowledge engineering techniques, before summing up with an epistemic overview of the main conceptual bridges discovered in the presented domain.

Theoretical Aspects in SNA

Standard SNA models are mainly based on the study of topology and flows within social graphs. We differentiate static models and dynamic models.

Topology and Flows

A graph is identified by its vertices and its edges (connections in the case of social networks). When connections are distributed depending on a Gaussian law, the graph structure is named random graph (Erdos and Rényi 1959), and when they depend on a power law (i.e., the probability for a given node to be connected to k other nodes is proportional to $k^{-\gamma}$, where γ is a parameter generally comprised between 2 and 3), the structure is called scale-free network (Barabasi and Albert 1999); scale-free networks contain many nodes with a very low number of connections, and a few "hubs" connected to many other nodes. The Web and social networks are identified as scale-free networks depending on preferential attachments (Barabasi and Albert 1999). Standard SNA measures are sensitive to topology because they generally follow geodesic paths – i.e., shortest paths connecting pairs of nodes (i, j) within a graph – so as to proceed to pairwise comparisons of nodes such as in the betweenness centrality defined in Freeman (1977) as follows:

$$C_{B(P_k)} = \sum_{i>j}^{n} \sum_{ij(P_k)}^{b}$$

The definition above adds $b_{ij(P_k)} = 1$ to the betweenness centrality of a point P_k for each geodesic path between the pair of nodes (i, j) comprising P_k, and so on for each pair (i, j) of a social graph. It has been successfully implemented and experimented in Erétéo (2011), in the context of a project deploying semantic Web languages and tools (i.e., RDF, SPARQL) on a professional dataset based on semantic annotations and collaborative documents sharing – cf. *Ontologies and the Semantic Web* section.

Integrating flows values enhances the results of SNA models, because it fosters the discrimination of representative positions such as leaders or eccentric influencers within social networks. It enables to differentiate hubs regarding the information they share and to take into account various flows metrics such as read/written textual content, social media viewing/listening, shared knowledge, positive/negative opinions, or friendliness (Chen and Qi 2011; Zhuhadar et al. 2011).

In order to produce relevant flows values, knowledge engineering techniques enable to define semantic flows metrics based on the content shared within social networks. For instance, the study of professional skills and activities in enterprises and/or institutions social networks introduces metrics of semantic intensity ($SemI$) and semantic resistance ($SemR$) based on linguistic analysis techniques (Thovex and Trichet 2012), such as in the following definitions:

$$SemI_{U,T,D} = TF(T, D);\ SemR_{U,T,C} = IDF(T, C)$$

In these definitions, U, T, D, C represent respectively a node, a term, a document, and the corpus of text documents tied to the studied

social networks. TF and IDF are well-known measures in the domain of linguistic statistics, which are trivially defined as follows:

$$\text{Term Frequency}_{(\text{term})} = |\text{term occurences}| \,/\, |\text{terms} \in \text{document}|$$

$$\text{Inverse Document Frequency}_{(\text{term})} = |\text{documents} \in \text{corpus}| \,/\, ||\text{document} \ni \text{term}||$$

They are introduced in SSNA by Robertson and Sparck Jones (1976), where they are coupled with semantic indexation and research services so as to produce semantic metrics which enable to value the ties between people and terms within a social graph, depending on the endogenous content – i.e., the content generated and shared within the studied social network. With such a graph, when i represents an individual and j represents content, $b_{ij(P_k)} = 1$ is easily weighted using the semantic edge metrics $SemI$ and/or $SemR$ as factors, in order to define a new semantic betweenness centrality and new semantic centralities based on standard SNA measures and path walks (Newman 2005). As $SemI$ and/or $SemR$ are not calculated for all the edges, we have defined a dynamic model propagating the metrics in a coherent way within the whole graph.

Static Models and Dynamic Models

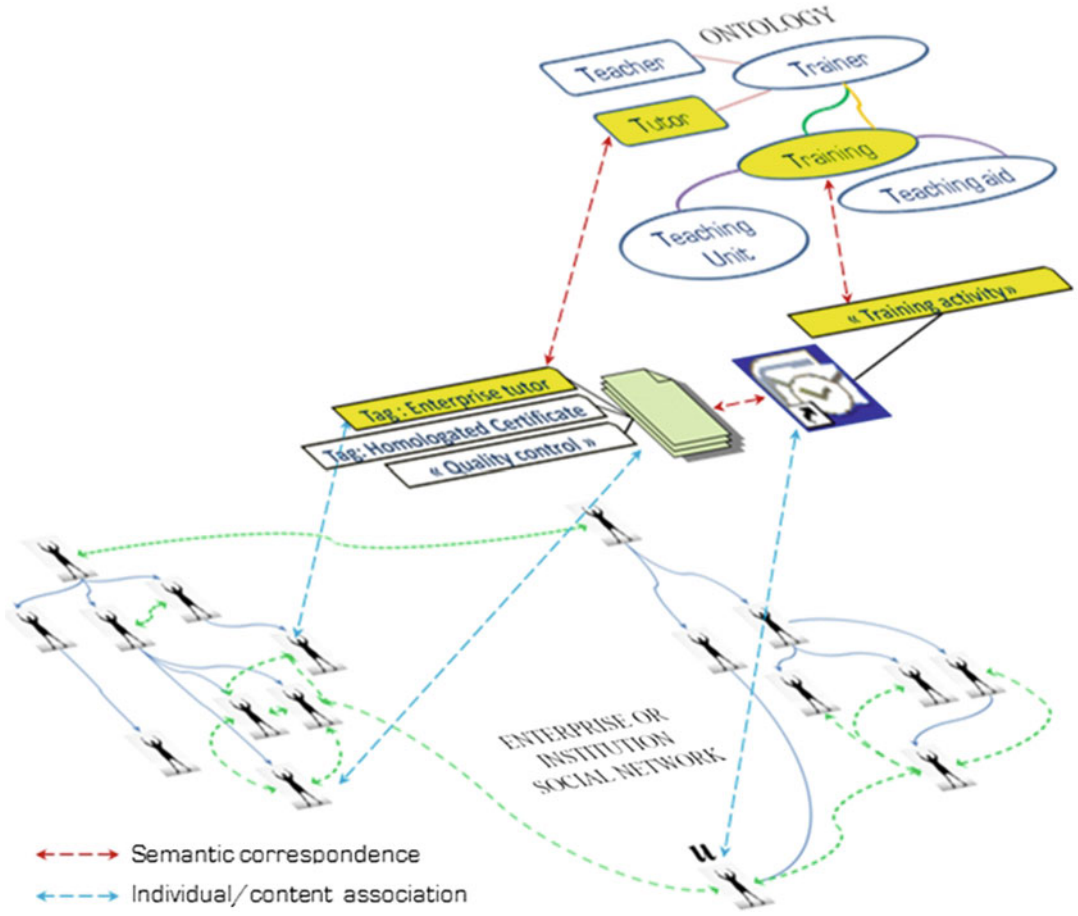
It seems essential to differentiate static SNA models, in which the values found within social graphs are not temporally dependent on each other (i.e., static values), from dynamic SNA models in which the values are temporally dependent on each other, like in electric circuits where the current of a part depends on the other parts. This metaphor is significant, regarding the main contributions to dynamic SNA based on the analogy between information and electronic flows (Newman 2005; Brandes and Fleischer 2005). It is also developed in physics, introducing SNA measures so as to prevent failures in electric power grids (Wang et al. 2010).

Static models are powerful when social graphs under study are fully weighted before being analyzed. For instance, applying the metrics $SemI$ and $SemR$ to a network representing the relationships between the members of an enterprise and the terms found in the mails they exchange, it is possible to define a social graph in which all edges are weighted by semantic values (Newman 2005). In such a context, a semantic referential such as an ontology representing the terms found in the network should increase the weighting relevance of each term, using semantic metrics such as defined in Aimé et al. (2010). Merging semantic networks and social networks fosters the development of relevant SSNA models. This example of social and semantic architecture is a case in point.

Figure 1 represents a multilayered view of semantic network and social network merged in a single structure. At the bottom of the picture, the dotted lines represent collaborative social relationships, and the full lines represent organizational relationships, within an enterprise social network. The individual u shares the content of an email comprising the expression "training activity." Through the term "training," represented in these mantic layer at the center of the picture, and thanks to the ontological relationships about this term, the email associated to the individual u is associated to the documentary resources comprising terms or annotations similar or close to "training" – e.g., "tutor." Individuals associated to these documentary resources are then more tightly associated to u, thanks to shared knowledge. So, the three individuals pointed by the individual/content associations become prominent nodes of a same semantic and social subgraph related to training and tutoring.

Moreover, the other expressions indexing the documentary resources of the socio-semantic subgraph detected (i.e., "homologated certificate," "quality control") can help in the automatic classification of endogenous resources, having recourse to semantic indexing and natural language processing techniques.

Dynamic SNA models enable to introduce heuristics based on natural behaviors, such as encountered in physics or biology (Galam 2008; Giugliano 2009). For instance, with the enterprise



Semantic Social Networks Analysis, Fig. 1 Semantic network and social network merged in a single structure

social graph that we previously took as an example, applying $SemI$ and $SemR$ only produces weighs on the edges, not on the nodes representing people or terms. Furthermore, the metrics are not always coherent within the whole graph and could transgress simple rules such as "for each node, the sum of incoming flows equals the sum of outgoing flows." The issue is solved by a dynamic SSNA model implementing Kirchoff's and Ohm's laws (Thovex and Trichet 2012). This model enables to weigh the whole graph by ensuring the coherence of all weighs, according to the natural balance of electronic flows in solid state circuits. This naturally coherent heuristic still does not take electromagnetic losses and interactions into account but improves previous epistemic approaches (Newman 2005; Brandes and Fleischer 2005), which introduce Kirchhoff's point law in dynamic SNA without integrating the Ohm's law, although it is a prerequisite in physics. The dynamic method of flow propagation defined in Thovex and Trichet (2012) owns two temporal aspects. On the one hand, it produces a coherent distribution through the whole graph, of the semantic values coming out from $SemI$ and $SemR$ on the edges connecting people to content. This phase enables to weight nodes and people-to-people and/or content-to-content edges, in a coherent way. On the other hand, temporal changes occurring within the input dataset might be processed, so as to compare the states of studied socio-semantic networks together and to produce temporal analysis of socio-semantic networks following a timeline.

Theoretical Aspects in Knowledge Engineering

Knowledge engineering aims at integrating knowledge in computer systems in order to lower the need for human intervention. Two main issues need to be addressed: knowledge discovery (i.e., mining techniques such as linguistic statistics) and knowledge representation and exchange – i.e., semantic formalisms. These issues are developed in the following sections.

Data Mining, Text Mining, and Graph Mining

Data mining aims at automatically finding patterns (such as rules or outliers) in large datasets.

The underlying motivation comes from the data explosion which has been taking place since several decades. Daily data generated by social networks contribute to the Big Data phenomenon. Moreover, a significant part of these data needs to be processed in real time, which represents an additional challenge to the one related to scalability.

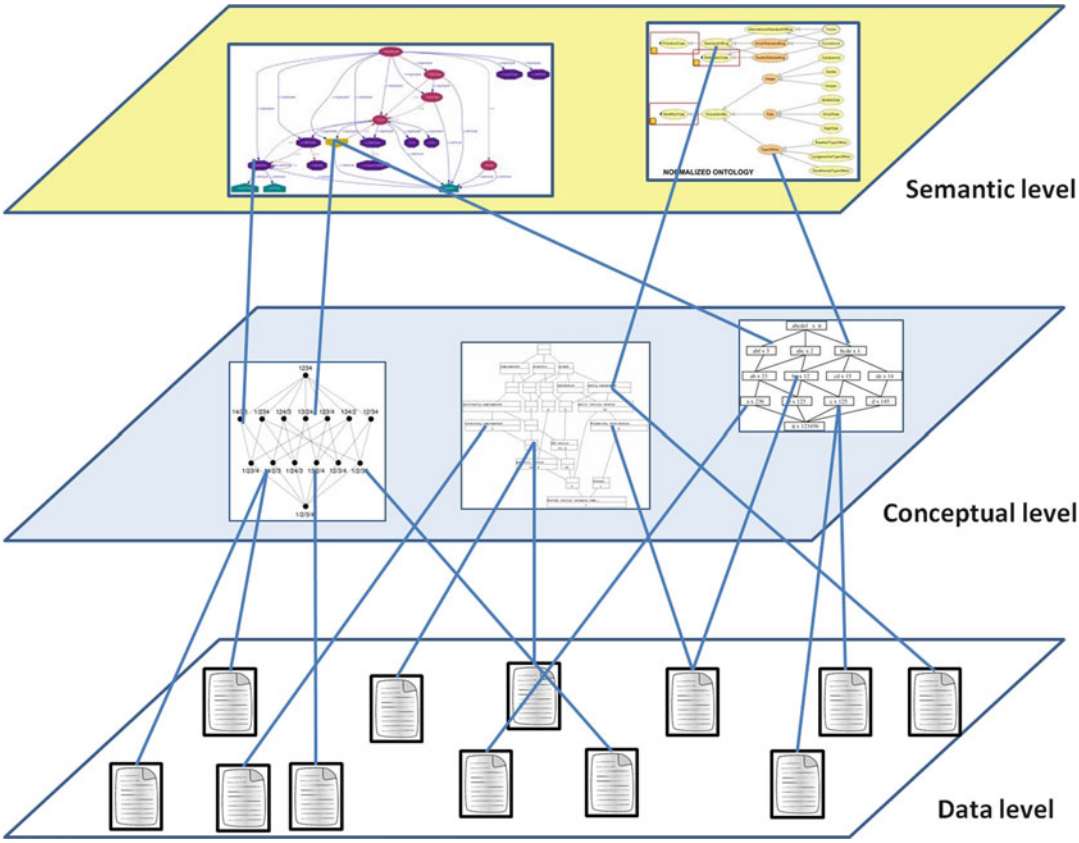
Many scientific areas contribute to data mining techniques, e.g., statistics, artificial intelligence, machine learning, and optimization. Data mining solutions include data classification or clustering, association rules generation, and outlier detection.

Data mining is applied to many sectors, among which text analysis, Web mining, marketing, financial or biological data analysis, or fraud detection. Moreover, the development of networked data such as computer, biological, or social networks has created new challenges for data mining and graph mining in particular. Indeed, these – often large and heterogeneous – real networks also called *complex networks* may be represented as graphs. Complex networks analysis has raised interest in the scientific community, and various graph mining techniques have therefore been developed in order to describe these real graphs and design models to generate realistic networks. Another trend in graph mining consists in identifying clusters of strongly connected nodes in the network, called communities (Fortunato 2010). Finally, very little is known about complex networks dynamics, and much remains to be done – e.g., study of communities evolution over time, ties, and interactions between data types.

Ontologies and the Semantic Web

Various formalisms exist within the semantic Web framework, with different levels of complexity and expressiveness, from simple annotation syntaxes to sophisticated reasoning capabilities. The eXtensible Markup Language (XML), the Resource Description Framework (RDF) (Lassila and Swick 1998), ontologies (Gruber 1993), rules, and logic all belong to the semantic Web picture. Many definitions of ontologies may be found in the literature; among them, Tom Gruber's (1993) is frequently referred to: "*An ontology is a formal specification of a shared conceptualization.*" An ontology basically describes concepts and the relationships among these concepts. A thesaurus may be seen as a light ontology as it also describes concepts, but the relationships among them are not specified as formally as in ontologies.

Conceptual graphs (Sowa 1976) constitute a way to represent and organize knowledge. Such graphs may be built from structured or unstructured data, for example, through the computation of Galois lattice based on formal concept analysis (Ganter and Wille 1998). From a set of elements (called *objects* in the FCA terminology) described by their properties (called *attributes*), a Galois lattice builds a partially ordered set of concepts, consisting each in objects sharing common attributes. Based on semantic attributes and relationships, it defines semantic networks and clusters which can be compared to the notion of community encountered in SNA. Semantic networks such as Galois lattices and ontologies represent the topology of semantic relationships between concepts enriched with various qualitative information. Hence, intrinsic features of members of a social network such as age, education, address, or hobbies (i.e., profiles) may be used for the identification of communities or for the recommendation of new contacts.



Semantic Social Networks Analysis, Fig. 2 Three-tier architecture: data, conceptual, and semantic layers

Conceptual graphs, ontologies, and thesauri are excellent candidates to support the convergence of semantic analysis and social network analysis, since both disciplines are based on graphic representations and heuristics. Conceptual graphs and semantic networks provide an intermediate layer between analyzed data and semantics, as shown on Fig. 2. Indeed, they have been successfully used for the analysis of social networks extracted from Myspace, Flickr, Dailymotion (Riadh 2009), and Twitter (Melo et al. 2012). A node of the conceptual layer may be linked to several nodes of the semantic layer, creating a bridge among various ontologies. New similarity metrics for ontology matching may also be derived from graph-based metrics. Conversely, a semantic node may be related to distinct concepts, allowing the navigation from a conceptual graph to another via the semantic layer.

Figure 2 illustrates how, in KE, a semantic network maps onto data. It represents the conceptual bases illustrated in Fig. 1 – which shows an ontology snippet mapped onto a social network *via* the content shared by its members. Similarity between both figures reveals how semantic networks are propitious to define SSNA models, but not only. The conceptual bridge it entails can be crossed from KE to SNA/SSNA, but also from SNA/SSNA to KE, in order to research new methods to build and/or populate semantic networks using SNA/SSNA models. In such a context, the possibility of discovering virtuous and self-learning models seems to be latent.

As explained earlier, the interest of semantics in social network analysis has been

acknowledged; conversely, SNA results may help maintain and enrich ontologies. For instance, communities identified through topological links within a social network may correspond to emerging concepts to be added in ontologies. Intra/inter-communitarian ties between members may help with ontology building in the interdisciplinary context of semantic and social networks analysis. This raises challenging research questions in order (1) to identify, within social networks and social content, useful and relevant ties for ontology building and/or matching and (2) to define bi-disciplinary self-learning processes in SSNA.

Semantic SNA: An Interdisciplinary Approach

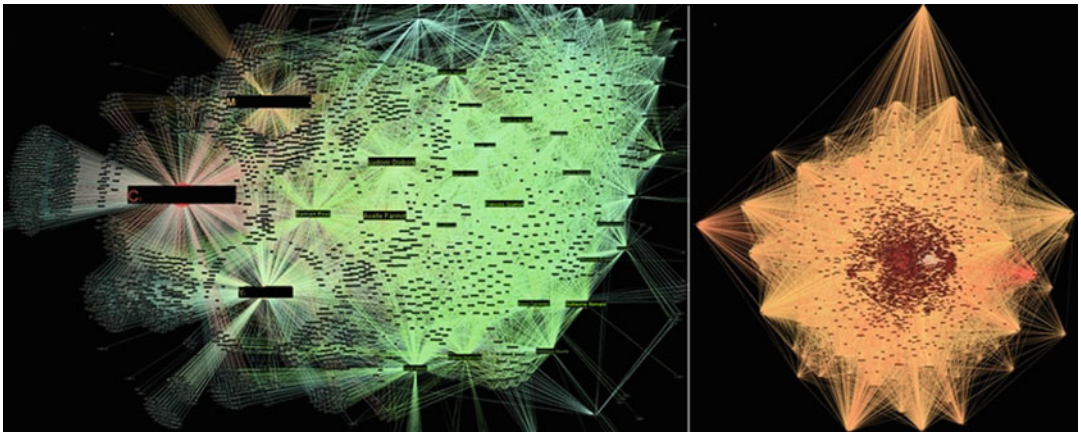
In the first part of the current section, we have explored the bases of social networks analysis and shown how semantic SNA enhances standard SNA – integrating semantics related to endogenous content into SNA measures – thanks to knowledge engineering methods such as linguistic statistics. In the second part, we have discovered a singular analogy between social networks and semantic networks, through a presentation of conceptual graphs, Galois lattices, ontologies, and thesauri – i.e., semantic representations based on graphs. We state that it opens an epistemic track paving the way for future and interdisciplinary SSNA models. We define SSNA as an interdisciplinary approach based on SNA and KE. SSNA introduces a generation of models which adapt the results of standard measures and metrics depending on the semantics found in the content shared within social networks. Current experimentations show a significant improvement of SNA results, thanks to SSNA models. We can imagine future extensions like merging opinion analysis into SSNA.

Unfortunately, current SSNA models are mostly dependent on the existence of text within the endogenous content, while social networks include more and more pictures, audio and/or video streams, bookmarks, or geographical locations. Before providing semantic data, the social Web requires a lot of various techniques for processing signal in visual and/or audio streams and for knowledge extraction in bookmarks and locations – e.g., face recognition, multilingual speech to text, musical pattern recognition, Web crawling, linguistic analysis, association rules, and fuzzy logic. Hence, without heavy preprocessing frameworks extracting textual representations and semantics from social media, SSNA omits a large part of the social content it is supposed to process. Though the explicit relationships within social networks and social media provide a turnaround, this seems to be one of the biggest obstacles to the deployment of SSNA in the socio-semantic Web, with more general problems such as processing very large datasets, mining the hidden/deep Web, or subtle psychosocial knowledge regarding human behaviors, hidden intentions, and subconscious opinions.

Key Applications

As social networks touch our social life, private life, economic life, and professional life, the application domain of SSNA is potentially vast and linked to SNA applications. One of the first major trends we have seen developed is criminal networks analysis for counterterrorism. Obviously, SNA/SSNA is an important decisional leverage for marketing agencies and strategies. In 2008, two founders of Facebook declared having to leave the enterprise to conceive a new kind of products that "will become to your work life what Facebook.com is to your social life" – cf. ASANA and http:/en.wikipedia.org/wiki/Dustin_Moskovitz. In the same time, enterprise social networks became as usual as mail exchanges in certain professional branches, and when they are consensually accepted, they are considered as tools fostering collaboration and productivity. They are also precious for human resources management and social capital management.

Experimenting the semantic metrics and measures defined in Thovex and Trichet (2012) on collaborative enterprise dataset, we have identified and ranked significant terms and teams within skills networks – i.e., socio-semantic networks representing professional collaborations.



Semantic Social Networks Analysis, Fig. 3 Visualization of SSNA results – samples

As a result, SSNA of skills networks provided a set of relevant indications helping in (1) self-managed collaboration and teams organization; (2) detection of critical topics, in terms of stress at work; and (3) redeployment of human resources, according to dynamic requirements in terms of competencies and workload. Evaluating our predictive and epistemic model with the experts involved in experimental phase, the produced recommendation enabled to retrieve a group of users sharing an anonymous account, though there was no explicit data allowing to identify these users in the studied dataset.

Future Directions

Semantic SNA: An Interdisciplinary Way to Be Paved

We have presented earlier the interdisciplinary dimension of SSNA. According to our knowledge of the domain, this dimension raises an unsuspected set of epistemic issues far beyond the analogy between electric current flows and information flows in social networks developed in Newman (2005), Brandes and Fleischer (2005), and Thovex and Trichet (2012). The interdisciplinary dimension of SSNA can be explored in depth as well as transversally.

Pursuing the in-depth exploration of epistemic equivalences between electrophysics and SSNA, we could intend to merge electromagnetic and thermodynamic principles into our current model (e.g., Maxwell's equations, Joule effect), so as to detect invisible information flows and semantic ties or risks of psychological burnout within social networks. Following its logical way, this in-depth epistemic immersion could lead to Schrödinger's and/or Dirac's equations – i.e., to relativist quantum electrodynamics.

Transversally exploring the interdisciplinary dimension of SSNA, we might discover epistemic connections between biological similarities and socio-semantic networks formulations, between knowledge networks and neuronal networks and/or brain dynamics, or between geography of social networks and knowledge networks, as an example. Figure 3 represents a sample of our experimental results, studying collaboration relationships within an enterprise. At the left on the picture, weighted degree centrality (i.e., the sum of all weights from edges connected to a node) based on $SemI*SemR$(named *semantic tension*) defines the size and color of nodes and the color of edges, from light blue for weak values to red for high values. It enables to identify the most important collaborators (largest hubs at the left) in terms of skills and knowledge (small nodes/terms around the hubs). The same dataset is represented at the right on the picture, based on semantic closeness centrality values such as defined in Thovex and Trichet (2012).

Orange color shows how semantic closeness is concentrated on average values, in this case. The most common terms are tied with most of the collaborators, represented as the core of the network. At the periphery, we find eccentric collaborators working on rare terms, sometimes with high semantic tension (red edges) which represent rare but important knowledge/skills within the enterprise.

From Social Outcomes to Strategic Outcomes

While social networks thoroughly describe our social, private, economic, and professional lives, SSNA outcomes are gradually turning into strategic outcomes. The sum of indications and recommendations they provide quickly becomes strategic for economy, politics, education, and information sharing all around the world. It also concerns our social, private, economic, and professional lives, through current and future SSNA applications for contextual social networks, cyber and/or cultural anthropology or geography, evolutions of social organizations, participative democracy, privacy, security and liberty, product purchase, empowering social ties within society, or participative and digital newspapers. Based on current facts and trends, we can reasonably hope that the benefits of SSNA will be larger than its possible perverse effects.

Cross-References

- ▸ Analysis and Mining of Tags, (Micro)Blogs, and Virtual Communities
- ▸ Classical Algorithms for Social Network Analysis: Future and Current Trends
- ▸ Community Detection, Current and Future Research Trends
- ▸ Dynamic Community Detection
- ▸ Managerial Networks
- ▸ Modeling Social Preferences Based on Social Interactions
- ▸ Ontology Matching
- ▸ Path-Based and Whole-Network Measures
- ▸ Recommender Systems, Semantic-Based
- ▸ Role Identification of Social Networkers
- ▸ Semantic Social Networks
- ▸ Social Capital
- ▸ Social Interaction Analysis for Team Collaboration
- ▸ Social Network Analysis in Organizational Structures Evaluation
- ▸ Social Recommendation in Dynamic Networks
- ▸ Virtual Team

References

Aimé X, Furst F, Kuntz P, Trichet F (2010) Prototypicality gradient and similarity measure: a semiotic-based approach dedicated to ontology personalization. J Intell Inf Manag 2(2):65–158. issn:2150–8194

Barabasi A-L, Albert R (1999) Emergence of scaling in random networks. Sci Mag 286(5439):509–512

Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. Sci Am Mag

Brandes U (2001) A faster algorithm for betweenness centrality. J Math Sociol 25:163–177

Brandes U, Fleischer D (2005) Centrality measures based on current flow. In: 22nd symposium theoretical aspects of computer science (STACS 05), Stuttgart. LNCS, vol 3404. Springer, pp 533–544

Chen L, Qi L (2011) Social opinion mining for supporting Buyer's complex decision making: exploratory user study and algorithm comparison. Soc Netw Anal Min J 1(4):301–320. Journal by Springer

Erdos P, Rényi A (1959) On random graphs. Publicationes Mathematicae 6:290–297

Erétéo G (2011) Semantic social network analysis. PhD thesis, Laboratoire d'Informatique, Signaux et Systèmes de Sophia-Antipolis (L3S, UMR6070 CNRS), Université de Nice Sophia-Antipolis

Fortunato S (2010) Community detection in graphs. Phys Rep 486:75–174

Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40:35–41

Freeman LC, Bloomberg W, Koff SP, Sunshine MH, Fararo TJ (1960) Local community leadership. University College of Syracuse University, Syracuse

Freeman LC, White DR, Romney AK (1989) Research methods in social network analysis. George Mason University Press, Fairfax

Galam S (2008) Sociophysics: a review of galam models. Int J Mod Phys C 19(3):409–440

Ganter B, Wille R (1998) Formal concept analysis: mathematical foundations, 1st edn. Springer, Berlin

Giugliano M (2009) Calcium waves in astrocyte networks: theory and experiments. Front Neurosci 3(2):160–161

Gruber TR (1993) A translation approach to portable ontology specifications. Knowl Acquis 5(2):199–220
Lassila O, Swick RR (1998) Resource description framework (RDF) model and syntax specification. Technical report, World Wide Web Consortium, Cambridge
Melo C, Le Grand B, Aufaure M-A (2012) A conceptual approach to characterize dynamic communities in social networks: application to business process management. In: BPMS2 2012: 5th international workshop on business process management and social software, Tallin
Miramontes O, Luque B (2002) Dynamical small-world behavior in an epidemical model of mobile individuals. Physica 168–169:379–385
Moreno J (1934) Who shall survive? – (Trad. fr) Fondements de la sociométrie. PUF
Newman MEJ (2005) A measure of betweenness centrality based on random walks. Soc Netw 27(1):39–54
Pearson M, West P (2003) Drifting smoke rings: social network analysis and Markov processes in a longitudinal study of friendship groups and risk taking. Connect Bull Int Netw Soc Netw Anal 25(2):59–76
Riadh TM, Le Grand B, Aufaure M-A, Soto M (2009) Conceptual and statistical footprints for social networks' characterization. In: SNA-KDD '09: proceedings of the 3rd workshop on social network mining and analysis, Paris. ACM, pp 1–8
Robertson SE, Sparck Jones K (1976) Relevance weighting of search terms. J Am Soc Inf Sci 27(3):129–146
Shimbel A (1953) Structural parameters of communication networks. Bull Math Biophys 15:501–507. Stress rate
Sowa JF (1976) Conceptual graphs for a data base interface. IBM J Res Dev 20(4):336–357
Tajfel H, Billig M, Bundy R, Flament C (1971) Social categorization and intergroup behavior. Eur J Soc Psychol 1:149–178
Thovex C, Trichet F (2012) Semantic social networks analysis: towards a sociophysical knowledge analysis. Soc Netw Anal Min J 2(1):1–15. Journal by Springer
Wang Z, Scaglione A, Thomas RJ (2010) Electrical centrality measures for electric power grid vulnerability analysis. In: IEEE (ed) Proceedings of the 49th IEEE conference on decision and control, CDC 2010, Atlanta, 15–17 Dec 2010, pp 5792–5797
Zhuhadar L, Nasraoui O, Wyatt R, Yang R (2011) Visual knowledge representation of conceptual semantic networks. Soc Netw Anal Min J 3:219–299. Journal by Springer

Recommended Reading

Memon N, Alhajj R (eds) (2010) From sociology to computing in social networks theory, foundations and applications. Lecture notes in social networks, vol 1. Springer, Wien/NewYork

Semantic Social Web

▸ Twitris: A System for Collective Social Intelligence

Semantic Web

▸ Sources of Network Data

Semantic Web Search

▸ Query Answering in the Semantic Social Web: An Argumentation-Based Approach

Semantic Web Service Composition

▸ Web Service Composition

Semantic Web Services

▸ Service Discovery

Semi-discrete Decomposition

Aswani Kumar Cherukuri
School of Information Technology & Engineering, VIT University, Vellore, India

Synonyms

Bump hunting technique; Matrix decomposition; Singular value decomposition

Glossary

Clustering An unsupervised data mining technique that places the data objects into different groups (clusters) such that the objects in a cluster are more similar to each other and dissimilar to objects in other clusters

Data Matrix A rectangular array with m rows and n columns, the rows representing different observations or individuals or objects and the columns representing different attributes or variables or items

Data Mining Nontrivial extraction of previously unknown, potentially useful, and reliable patterns from a set of data

Dimensionality Reduction The process of embedding a set of n points in a d-dimensional space into a k-dimensional space, where d is sufficiently large and k is much smaller than d

High-Dimensional Data Data in which the objects are described by a large number of features, where each feature factor corresponds to a dimension. While analyzing a data matrix of size m by n, we refer the matrix as n-dimensional since we consider the view of m points in an n-dimensional space

Matrix Decomposition Transformation of original data matrix into a given canonical form, as a product of new matrices. This transformation is aimed at revealing the latent structures or relations in the original data matrix. This transformation is also known as matrix factorization

Matrix Rank Reduction Given a data matrix $\boldsymbol{A}$ having rank r, the process of finding a matrix $\hat{\boldsymbol{A}}$ having rank k where $k < r$ and minimizes $||\boldsymbol{A} - \hat{\boldsymbol{A}}||$

Definition

Semi-discrete decomposition (SDD) is a matrix decomposition technique that produces low-rank approximation of original matrix as a weighted sum of outer products. With its approximation, SDD defines new axes that capture the variance in the data. Though this approximation is similar to that of singular value decomposition (SVD), the axes of the SDD transformed space are not orthonormal, and coordinates of the points in the transformed space are restricted to the set of values $\{-1,0,1\}$. With such a restriction, SDD achieves the storage economization than other decomposition techniques like SVD. Hence, the higher-rank approximations can be stored for smaller amount of storage. With an iterative procedure, SDD aims to find and extract the locations in the given dataset having extremely large magnitude values which are both positive and negative. SDD represents the data matrix as the sum of bumps and arrange the bumps such that the most significant bump appears first. Hence, SDD is generally treated as a bump hunting technique and proved to be effective in finding the outlier clusters in the data. Also SDD produces an unsupervised, hierarchical, and ternary classification by partitioning the data items having similar attribute values. Hence, SDD is applied in classification and clustering problems. SDD has a unique property of discovering more latent factors than the available features in the dataset. In addition to its primary motivation in digital image processing, SDD has successful applications in finding outliers in the data, semantic indexing, etc.

Introduction

Most of the engineering, scientific, and computer applications result in high-dimensional datasets containing large number of variables associated with each observation. Also such data is often a combination of several underlying processes coupled with noise. The dimension of a dataset is defined by the number of variables that are measured on each observation. However, all these variables are not necessary to understand the latent structure of the data. Such high-dimensional data coupled with noise poses several computational challenges. In addition to the complexity prevailing in analyzing the high-dimensional datasets, the similarities between the objects in the high-dimensional space diminish with regard to the Euclidean distance. This would negatively influence the accuracy

of the analysis. This problem is referred to as *curse of dimensionality* (Cunningham 2007). The solution to this curse is to apply dimensionality reduction techniques as a preprocessing step. This process requires identification of a suitable low-dimensional representation of original high-dimensional data. Also this preprocessing step improves the accuracy of the data analysis (Dobsa et al. 2012). Several dimension reduction techniques are available in the literature. Interested readers can refer to few authoritative references (Aswani Kumar 2009; Cunningham 2007; Fodor 2002). An optimal technique can efficiently map the original data to suitable lower dimension while preserving the properties of original data. By representing the data in the form of a matrix, we get a convenient way to store and analyze data. If a data matrix $\boldsymbol{A}$ of size $m \times n$, containing m objects and n attributes, each object can be considered as a point in the n-dimensional space spanned by the attributes.

Matrix rank reduction techniques from linear algebra are popular in data analysis and mining problems for finding low-dimensional representation of data (Elden 2006). Rank of the matrix is defined as the number of linearly independent rows or columns of the matrix and helps to measure the contents of the matrix. Redundancy in the matrix arises due to the dependent row or column vectors. This redundancy can be removed by mapping or replacing the dependent vectors with linear combination of other linearly independent vectors.

Generally decomposition of the matrix refers to the decomposition to some approximation. Decomposition of a matrix produces two or more factor matrices. The original matrix can be represented as a product of these factor matrices. The main motivation behind the matrix decomposition lies in the fact that the inner dimension value of k is much smaller than the original dimensions (m, n) of the data matrix. The matrix decomposition techniques are mainly intended to segregate the different processes that are captured by the dataset and to cluster the similar objects of the dataset in some standard understandable way. These techniques can be applied as a stand alone or in combination of other techniques. The notions like dimensionality reduction, matrix rank reduction, matrix factorization, and data compression are closely related and are based on Wedderburn rank reduction theorem (Miettinen 2009; Park and Elden 2003; Elden 2007; Skillicorn 2007).

Several matrix rank reduction techniques are available that include singular value decomposition (SVD), semi-discrete decomposition (SDD), and nonnegative matrix factorization (NMF) (Miettinen 2009). Each of these techniques differs in the way they decompose the matrix, constraints that they impose on the elements, relationship among the rows and columns, etc. Recently heuristic techniques like clustering and random projections are also used in the literature for matrix rank reduction (Aswani Kumar and Srinivas 2010; Aswani Kumar 2011).

Key Points

Semi-discrete decomposition (SDD) was originally introduced by O'Leary and Peleg (1983) for the purpose of digital image compression. Later it is extended as a storage efficient variant of SVD in latent semantic indexing (LSI)-based IR application. Based on vector space representation, LSI finds low-rank approximation of term-document collection using SVD (Aswani Kumar and Srinivas 2006; Berry et al. 1999; Deerwester et al. 1990). However, if the original matrix is sparse, the low-rank approximation achieved through SVD requires more storage than the original matrix. To overcome this difficulty Kolda and O'Leary (1998) have proposed to use SDD for LSI. An analogy can be brought between SDD and SVD, Boolean matrix decompositions. Both the factor matrices and the matrix multiplication in Boolean decompositions are binary (Miettinen 2009). Similar to SVD, SDD obtains three matrices, but elements of outer product vectors are restricted to -1, 0, and 1.

Given a data matrix $\boldsymbol{A}$ of size (m, n), with m objects and n attributes, SDD finds the approximation of $\boldsymbol{A}$ to a lower dimension k as follows:

$$\boldsymbol{A}_k = \boldsymbol{X}_k \boldsymbol{D}_k \boldsymbol{Y}_k^{\mathrm{T}}$$

where the matrix $\boldsymbol{X}_k$ is of size $m \times k$, $\boldsymbol{D}_k$ is a diagonal matrix of size $k \times k$, and $\boldsymbol{Y}_k^{\mathrm{T}}$ is a matrix of size $k \times n$. The entries of $\boldsymbol{D}_k$ matrix are nonnegative real numbers. Each $\boldsymbol{D}_i$ value ($1 \leq i \leq k$) indicates significance of the ith factor. The rows of the matrix $\boldsymbol{X}_k$ are considered as the coordinates of an object in the space defined by the new axes that are described by the rows of $\boldsymbol{Y}_k^{\mathrm{T}}$. The variation in original data is captured and is concentrated along the earlier axes defined by $\boldsymbol{Y}_k^{\mathrm{T}}$. The lower axes in which the lesser variance in the data is concentrated can be removed to achieve the approximation. The axes of the transformed space are not orthonormal. The coordinates can have different interpretations depending on the application.

The elements of the matrices $\boldsymbol{X}_k, \boldsymbol{Y}_k^{\mathrm{T}}$ are from the set $\{-1,0,1\}$. This transformation of original n-dimensional space into new k-dimensional space results in dimensionality reduction of original matrix. Generally in SVD applications the value of k will be chosen as $m \leq k \leq n$. However in SDD, the value of k can even be higher than n. Hence, SDD can identify the more latent factors than the existing features in the dataset. SDD tries to define new axis that captures larger variation in the original data. The algorithm starts by identifying the values of first column of the matrix $\boldsymbol{X}_k$, the first axis vector of $\boldsymbol{Y}_k^{\mathrm{T}}$ and multiplier in $\boldsymbol{D}_k$ that gives the least amount of error between the approximation matrix, $\boldsymbol{A}_k$ and the original data matrix $\boldsymbol{A}$. The iterative process continues by selecting successive fields of these matrices in such a manner that reduces the error in the approximation.

The SDD decomposition can have three types of interpretations, namely, factor interpretation, geometric interpretation and component interpretation, (Skillicorn 2007). By considering the rows of the matrix $\boldsymbol{Y}_k^{\mathrm{T}}$ as factors that are mixed by the rows of $\boldsymbol{X}_k$ and diagonal entries of $\boldsymbol{D}$, we can obtain the factor interpretation. This representation is useful in image processing. In geometric interpretation, the rows of the matrix $\boldsymbol{Y}_k^{\mathrm{T}}$ define the generalized quadrants, and the values of the matrix $\boldsymbol{X}_k$ can then identify whether a given object is placed in the given quadrant or not. Component interpretation can be obtained by expressing the original matrix $\boldsymbol{A}$ as sum of the outer product matrices, i.e., the ith column of matrix $\boldsymbol{X}$, the ith entry on the diagonal of the matrix $\boldsymbol{D}$, and the ith row of the matrix $\boldsymbol{Y}^{\mathrm{T}}$. Though graph interpretation for SDD can be obtained, it provides no new insights about the data.

The approximation of a matrix using SDD is achieved through an iterative and greedy algorithm, which computes a new column, a diagonal element, and a row in each step. Let $\boldsymbol{A}$ be the data matrix of size $m \times n$. Choose a value k that represents the maximum number of terms in the approximation. Let $\boldsymbol{A}_0$ be the zero matrix of size $m \times n$, x_i be the ith column of the matrix $\boldsymbol{X}_k$, d_i be the diagonal element of the matrix $\boldsymbol{D}_k$, and y_i be the ith row of $\boldsymbol{Y}_k^{\mathrm{T}}$. Let $\boldsymbol{R}_i$ be the residual matrix obtained at ith step, i.e., $\boldsymbol{R}_i = \boldsymbol{A} - \boldsymbol{A}_{i-1}$. Consider $\boldsymbol{R}_1 = \boldsymbol{A}$. In the following we present the algorithm:

1. Outer iteration, for each step of $i = 1$ to k.
2. Choose an initial y vector such that $\boldsymbol{R}_i y \neq 0$.
3. Inner iteration:
   i. Fix y and let x solve $\max_{x \in \mathrm{I}^m} \frac{(x^T \boldsymbol{R}_i y)^2}{\|x\|_2^2}$.
   ii. Fix x and let y solve $\max_{y \in \mathrm{I}^n} \frac{(y^T \boldsymbol{R}_i x)^2}{\|y\|_2^2}$.
   iii. Repeat the inner iteration until some heuristic convergence criterion is satisfied.
4. Let $x_i = x$, $y_i = y$, $d_i = \frac{x_i^T \boldsymbol{R}_i y_i}{\|x_i\|_2^2 \|y_i\|_2^2}$.
5. Calculate the ith term approximation $\boldsymbol{A}_i = \boldsymbol{A}_{i-1} + d_i x_i y_i^T$.
6. Calculate the residual matrix, $\boldsymbol{R}_{i+1} = \boldsymbol{R}_i - d_i x_i y_i^T$.
7. Repeat the outer iteration until $i = k$.

The convergence criterion for stopping the inner loop is to verify whether the residual improvement is further possible. O'Leary and Peleg (1983) have proposed a method to determine the condition for stopping the inner iterations. Computing SDD on the data matrix of size (m,n) to approximate it to a rank k, under the assumption of fixed number of inner iterations, the above heuristic algorithm has a complexity of $\mathrm{O}(k^2(m+n) + m\log m + n\log n)$. Generally it is observed that the number of inner iterations required is averaged near 10. Kolda and O'Leary (2000) have shown that the

SDD algorithm converges linearly to the original matrix. Also they have discussed strategies that can be used to initialize the y vector in outer iteration shown in step 2. Since the algorithm is a heuristic variant, the parameters need changes depending on the dataset. Implementation of this algorithm in MATLAB and C is available from http://www.cs.umd.edu/~oleary/SDDPACK.

In the SDD basic setting, the heuristic component is the selection of initial y_i. This selection does not always identify and remove the largest possible bump from the data matrix. Hence, a rearrangement of these bumps is required (Skillicorn 2007). Once the $\boldsymbol{X}_k$, $\boldsymbol{Y}_k^{\mathrm{T}}$, and $\boldsymbol{D}_k$ matrices are computed, the product of $\boldsymbol{d}_{\mathrm{i}}^{\mathrm{s}}$ with corresponding nonzero column entries of $\boldsymbol{Y}_k^{\mathrm{T}}$ is formed. The columns of $\boldsymbol{X}_k$, elements of $\boldsymbol{D}_k$, and rows of $\boldsymbol{Y}_k^{\mathrm{T}}$ are sorted into decreasing order of the products of $\boldsymbol{d}_{\mathrm{i}}^{\mathrm{s}}$ with $\boldsymbol{Y}_k^{\mathrm{T}}$. This reordering ensures that the axes with largest weight or the axes that capture large variation appear first in the ordering, and hence the strongest outlier will be placed closest to the top of the decision tree (Knight and Carosielli 2003).

Since elements of the outer product matrices $\boldsymbol{X}_k$ and $\boldsymbol{Y}_k^{\mathrm{T}}$ obtained from SDD contain the values $\{-1, 0, 1\}$, higher-rank approximations can be stored at less amount of space. For rank k approximation of matrix of size (m, n), SDD requires the storage of $k(m + n)$ values from the set $\{-1, 0, 1\}$ for the matrices $\boldsymbol{X}_k$, $\boldsymbol{Y}_k^{\mathrm{T}}$ and k scalar values for the matrix $\boldsymbol{D}_k$. To store the values from the set $\{-1, 0, 1\}$ requires $\log_2 3$ bits. The scalar values for the matrix $\boldsymbol{D}_k$ need to be only single precision values. However, the SVD is computed with double precision values and hence requires nearly 32 times more space than SDD (Kolda and O'Leary 1998).

Unlike SVD, even for value $k = n$, the SDD does not produce the approximation matrix that is equal to the original matrix, i.e., $\boldsymbol{A}_k \neq \boldsymbol{A}$ for $k = n$ (Snasel et al. 2008). When the data is organized naturally in many small and well-separated clusters, SDD and SVD tend to agree and hence produce similar results. This is the main reason for usage of SDD as a replacement of SVD in LSI, since term-document matrices usually contain several natural small clusters (Kolda and O'Leary 1998, 2000). However, SVD and SDD do not produce similar results on the datasets that are organized in the form of large clusters. The basic problem with SDD is that the approximation takes five times more time than computing SVD. However, SDD updating is much easier than the SVD updating (Kolda and O'Leary 2000). SDD can be extended as weighted SDD and tensor SDD. These extensions along with their convergence issues are discussed in Kolda and O'Leary (2000).

Objects of the data matrix $\boldsymbol{A}$ can be hierarchically classified using the columns of the matrix $\boldsymbol{X}_k$. The analysis start, with the first column of the matrix $\boldsymbol{X}_k$. Objects (i.e., rows) of the matrix $\boldsymbol{A}$ are divided into three classes according to the value $\{-1, 0, 1\}$ that appears in the first column of the matrix $\boldsymbol{X}_k$. The objects whose value is $+1$ in the first column of $\boldsymbol{X}_k$ are in one class, the objects whose value is -1 are in one class and the objects whose value is 0 are in the third class so that the classification forms a ternary decision tree structure.

From each class, the objects are further divided into three subclasses depending on the value $\{-1, 0, 1\}$, corresponding to each object of matrix $\boldsymbol{A}$, in the second column of the matrix $\boldsymbol{X}_k$ and so on. The process can be stopped when a set of objects cannot be separated by the next levels or when each object is alone at a particular level. The analysis generates a ternary, hierarchical decision tree structure of depth k. In contrast to the conventional decision trees, the decision tree induced by the SDD is an unsupervised structure. By following the same procedure on the Y_k^{T} matrix, we can obtain the hierarchical classification structure of attributes. The general notion is that the classes -1 and 1 represent the data objects that have attributes significantly different from the normal data objects represented by the class 0.

In another perspective, by treating each class as a partition, we can consider that SDD performs partitional clustering. The division of data objects into three groups using the first column of matrix $\boldsymbol{X}_k$ and further subdivision of each group based on the subsequent columns of $\boldsymbol{X}_k$ results

in a hierarchical clustering of the objects of $\boldsymbol{A}$. The clustering contains k levels. The partitions at each level are independent. Unlike standard hierarchical clustering, the result of SDD-based hierarchical clustering is a ternary tree. The branches with the groups -1 and $+1$ are equal and opposite, but not different. Similarity among the objects or attributes can be computed using a distance measurement metric in the ternary representation structure.

Another important perspective of SDD is as a bump hunting technique (McConnell and Skillicorn 2002). Let us consider the original data matrix $\boldsymbol{A}$ as a grid of entries. Each positive entry of $\boldsymbol{A}$ is considered as a bump/tower at that position in the grid, with a height proportional to the values of the entry. Similarly each negative entry of $\boldsymbol{A}$ is treated as negative bump/hole with the depth proportional to the value of the entry. SDD searches for the regions of similar height and depth. One particular component of the decomposition is identified, once such a region is found. The average height or depth of the region is computed and subtracted from all the bumps and holes involved. Then the process continues for searching such similar regions and identifying the components of the decomposition. At each iteration, the position of the region is identified using the product of x_i and y_i^T, and the height of the bump is identified using d_i. If the original matrix $\boldsymbol{A}$ is represented as sum of a set of $\boldsymbol{A}_i^{\$}$, then each $\boldsymbol{A}_i$ represents a bump. The bumps are discovered based on their volume. Since the SDD selects the bump/hole based on the height and region, it is not scale independent. However, SVD is a scale-independent technique since the scaling process does not change the decomposition result. Scaling the magnitudes by squaring, SDD first selects the smaller regions of large magnitudes. Similarly if the magnitudes are replaced by their signed square roots, then SDD first selects the larger regions of smaller magnitude. There are other bump hunting techniques like PRIM and rule-based techniques. Methods based on SVD are available that result in decision tree classification like Principle Direction Division Partitioning (PDDP) (Skillicorn 2007).

Key Applications

SDD has found several applications in the literature. SDD for outlier detection was used by McConnell and Skillicorn (2002). Based on this application, SDD is further used for counter terrorism, social network analysis, detecting deceptive communications in the e-mails, etc. (Divya et al. 2011; Keila and Skillicorn 2005; Knight and Carosielli 2003; Skillicorn 2004; Snasel et al. 2010).

In collaborative filtering applications and recommender systems, SDD can be applied to identify the groups of objects that are rated highly by the individuals (Skillicorn 2007). SDD is successfully applied for image and video compression. Pattern matches and motion vectors in video coding can be computed using SDD (Zyto et al. 2002). For compressing the large images, truncated SDD of the image matrix can be considered as approximation to the original image.

With its features, SDD is well suited for hierarchical clustering and decision tree classification problems (Skillicorn 2007). For information retrieval and text mining applications, SDD is used as an alternate method for SVD (Kolda and O'Leary 1998). In addition to the IR applications, LSI technique can be augmented with SDD in automated text categorization application (Pilato et al. 2005). SDD can also be applied for obtaining the subsymbolic representation of words (Qiang et al. 2004), topic identification (Snasel et al. 2008). In social network and link analysis, the matrix decompositions can be useful to derive the higher-order information about the relationships among the individuals in the network. Based on the relationship, the members in the network can be ranked (Skillicorn 2007).

Based on the application, SDD can be directly applied on the data matrix or on the correlation matrix of the original data matrix, or on the approximated correlation matrix. SVD and SDD can be combined so as to complement each strength. On a dataset by applying SVD, we can visualize latent clusters within the data. But SVD cannot provide a way to label these clusters.

On the other hand, SDD provides the clusters within the data and label them. However, SDD cannot produce the visualization of these clusters. While performing this combination, first SVD can be computed on the data matrix $\boldsymbol{A}$ and decompose it to an appropriate rank k matrix $\boldsymbol{A}_k$ matrix using SDD. The advantage of this computation is that the SVD performs denoising of the data, and SDD identifies and labels the clusters within the data (Aswani Kumar and Palanisamy 2010). Also this computation can effectively be applied for classification of protein sequences and exploration of minerals, galaxies, etc. (Skillicorn 2007). Also SDD can be applied on the correlation matrix obtained from truncated SVD matrix $\boldsymbol{A}_k$. In this case, the SDD is used to find the correlation structure within the denoised data.

Illustrative Examples

From the above discussion, we can understand that the SDD can be used for bump selection, hierarchical clustering, LSI-based information retrieval, etc. In the following we see some of the examples illustrating these applications. To better understand SDD as a bump hunting, let us consider the following matrix:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 5 & 5 & 5 \\ 1 & 9 & 9 & 1 & 1 \\ 1 & 9 & 9 & 1 & 1 \\ 1 & 9 & 9 & 1 & 1 \\ 1 & 1 & 5 & 5 & 5 \end{bmatrix}$$

The SDD on this data matrix produces the following factorization matrices:

$$\boldsymbol{X}_k = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}, \boldsymbol{D}_k = \begin{bmatrix} 9 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\text{and } \boldsymbol{Y}_k^{\mathrm{T}} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Now consider the first outer product $i = 1$, then $\boldsymbol{X}_i^* \boldsymbol{Y}_i^{\mathrm{T}}$ is

$$\begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} * \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

We can understand that the resultant matrix is a stencil, representing the region of the array elements having the value 9 (which is the value d_1). Similarly for $i = 2$, the second outer product produces $\boldsymbol{X}i^* \boldsymbol{Y}_i^{\mathrm{T}}$ as

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} * \begin{bmatrix} 0 & 0 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

It is clear that the second outer product result is a stencil representing the region of the array elements having the value 5 (which is the value d_2). Similarly we can obtain the other stencil regions for the values of d using corresponding outer products.

SDD derives a hierarchical clustering of the objects producing ternary tree structure. To illustrate this process, let us consider a well-explained example from Skillicorn et al. (2003). The following is the data matrix of size 9×8.

$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 3 & 4 & 4 & 5 & 5 & 6 & 7 & 9 \\ 1 & 8 & 2 & 7 & 3 & 6 & 4 & 5 \\ 9 & 8 & 7 & 6 & 5 & 4 & 3 & 2 \\ 9 & 4 & 8 & 3 & 7 & 2 & 6 & 1 \\ 2 & 3 & 2 & 4 & 2 & 5 & 2 & 6 \\ 3 & 4 & 3 & 4 & 4 & 3 & 4 & 3 \\ 3 & 2 & 4 & 3 & 2 & 4 & 3 & 2 \\ 5 & 5 & 4 & 4 & 6 & 6 & 2 & 2 \end{bmatrix}$$

We apply SDD on this data matrix for rank $k = 8$. After rearrangement of the bumps as discussed in the above section, the following are the $\boldsymbol{X}_k$, $\boldsymbol{Y}_k^{\mathrm{T}}$ and $\boldsymbol{D}_k$ matrices of sizes $9 \times 8, 8 \times 8$ and 8×8, respectively.

$$\boldsymbol{X}_k = \begin{bmatrix} -1 & -1 & 1 & -1 & 0 & 1 & 0 & 1 \\ -1 & 0 & 1 & -1 & 0 & 1 & 1 & 1 \\ -1 & 1 & 0 & 1 & 0 & -1 & 0 & 0 \\ 1 & 1 & -1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & -1 & 0 & 1 & 1 & -1 \\ -1 & -1 & -1 & -1 & 0 & 1 & -1 & -1 \\ 0 & 0 & 0 & -1 & 0 & -1 & 0 & -1 \\ 0 & -1 & -1 & -1 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 1 & -1 & 1 & -1 & -1 \end{bmatrix} \text{ and } \boldsymbol{Y}_k^{\mathrm{T}} = \begin{bmatrix} 1 & 1 & 0 & -1 & 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ -1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 1 & -1 & 0 & -1 & 1 \\ 0 & 0 & 1 & -1 & 1 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

$$\boldsymbol{D}_k = \begin{bmatrix} 4.6134 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2.685 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1.905 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.834 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.5527 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1.271 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1.1556 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1.1239 \end{bmatrix}$$

The hierarchical clustering structure can be obtained from the data objects present in the matrix $\boldsymbol{X}_k$ as shown in Fig. 1.

Starting from the first column, the objects are grouped or clustered based on their entries -1, 0, and 1. Members of each group are further divided into subgroups based on their entries in subsequent columns. For example, in the tree structure shown in Fig. 1, we can understand that based on the entries from the first column, the objects $\{1, 2, 3, 6\}$ are grouped under label -1; the objects $\{7, 8, 9\}$ are grouped under the label 0, and the objects $\{4, 5\}$ are grouped under the label $+1$. In the next level, each of these groups is subdivided into three groups based on the entries in their second column. Similarly we can obtain hierarchical clustering on the attributes of the data represented in the matrix $\boldsymbol{Y}_k^{\mathrm{T}}$, by following the procedure illustrated above.

In information retrieval applications, SDD can be used by augmenting with LSI model (Kolda and O'Leary 1998). Let us consider $\boldsymbol{A}$ is a term-document matrix of size $m \times n$, having m terms and n documents with rank r. Let q be the query vector of length m used to probe on the document collection. Column normalization will be performed on the term-document matrix. After applying SDD on the data matrix $\boldsymbol{A}$, we obtain the factor matrices $\boldsymbol{X}_k$, $\boldsymbol{D}_k$, and $\boldsymbol{Y}_k^{\mathrm{T}}$ having the sizes $m \times k$, $k \times k$, and $k \times n$, respectively, as discussed above. We apply the query on this reduced dimensional space to compute the similarity of the documents. However before processing, the query vector should be projected onto the lower-dimensional space obtained by SDD as:

$$q_k = q\boldsymbol{X}_k\boldsymbol{D}_k$$

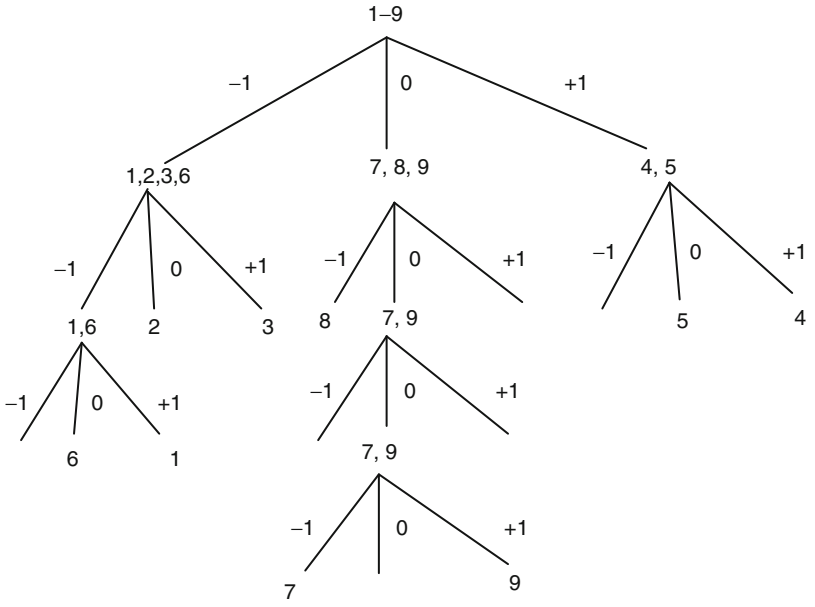
Now the similarity between the document and query vectors in the reduced dimensional space is calculated as

$$\text{sim} = q_k\boldsymbol{Y}_k^{\mathrm{T}}$$

Based on the similarity documents can be ranked and returned to the user. Consider a term-document matrix of size 9×7 having 9 index terms (T) and 7 documents (D) (Berry et al. 1999). The following are the terms and documents:

T1: Bab(y,ies,'s), T2: Child(ren's), T3: Guide, T4: Health, T5: Home, T6: Infant, T7: Proofing, T8: Safety, and T9: Toddler

D1: *Infant* & *Toddler* First Aid, D2: *Babies* & *Children's* Room, D3: *Child Safety* at Home,



Semi-discrete Decomposition, Fig. 1 Hierarchical clustering structure obtained from data objects

D4: Your *Baby's Health & Safety*: From *Infant* to *Toddler*, D5: *Baby Proofing* Basics, D6: Your *Guide* to Easy Rust *Proofing*, and D7: Beanie *Babies* Collector's *Guide*. Now the term-document for this collection is

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

Now by making the unit columns, the normalized term-document matrix is

$$\boldsymbol{A} = \begin{bmatrix} 0 & 0.5774 & 0 & 0.4472 & 0.7071 & 0 & 0.7071 \\ 0 & 0.5774 & 0.5774 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.7071 & 0.7071 \\ 0 & 0 & 0 & 0.4472 & 0 & 0 & 0 \\ 0 & 0.5774 & 0.5774 & 0 & 0 & 0 & 0 \\ 0.7071 & 0 & 0 & 0.4472 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.7071 & 0.7071 & 0 \\ 0 & 0 & 0.5774 & 0.4472 & 0 & 0 & 0 \\ 0.7071 & 0 & 0 & 0.4472 & 0 & 0 & 0 \end{bmatrix}$$

By applying SDD on the term-document matrix to approximate to rank $k = 4$, we obtain the following $\boldsymbol{X}_4$, $\boldsymbol{Y}_4^{\mathrm{T}}$, and $\boldsymbol{D}_4$ matrices:

$$\boldsymbol{X}_4 = \begin{bmatrix} 1 & 1 & 1 & -1 \\ 1 & 0 & -1 & 0 \\ 1 & -1 & 1 & 1 \\ 0 & 1 & 0 & -1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \quad \boldsymbol{Y}_4^{\mathrm{T}} = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix} \quad \boldsymbol{D}_4 = \begin{bmatrix} 0.2389 & 0 & 0 & 0 \\ 0 & 0.2412 & 0 & 0 \\ 0 & 0 & 0.2289 & 0 \\ 0 & 0 & 0 & 0.3054 \end{bmatrix}$$

The following is the rank k (k=4) approximation matrix of $\boldsymbol{A}$ obtained using SDD:

$$\boldsymbol{A}_4 = \begin{bmatrix} -0.0642 & 0.2389 & 0.0101 & 0.4801 & 0.4678 & 0.2266 & 0.4678 \\ 0 & 0.2389 & 0.4678 & 0.2389 & 0.0101 & 0.0101 & 0.0101 \\ 0.0642 & 0.2389 & 0.0101 & -0.0022 & 0.4678 & 0.7089 & 0.4678 \\ -0.0642 & 0 & 0 & 0.2412 & 0 & -0.2412 & 0 \\ 0 & 0.2389 & 0.4678 & 0.2389 & 0.0101 & 0.0101 & 0.0101 \\ 0.5465 & 0 & 0 & 0.2412 & 0 & -0.2412 & 0 \\ 0.0642 & 0.2389 & 0.0101 & -0.0022 & 0.4678 & 0.7089 & 0.4678 \\ -0.0642 & 0.2389 & 0.4678 & 0.4801 & 0.0101 & -0.2311 & 0.0101 \\ 0.5465 & 0 & 0 & 0.2412 & 0 & -0.2412 & 0 \end{bmatrix}$$

We note from this approximated term-document matrix that the elements have negative values. These values represent the linear combination of elements of original term-document matrix (Berry and Browne 2005). However, the individual term component of document vectors does not define the semantic content. The approximation automatically provides an association with relevant terms in each document. For example, along with the original terms T6 and T9 in the approximation space, the document D1 is associated with T3 and T7 also. Consider that the user wants to find the books on *Child Home Safety* from the document collection listed above. The corresponding query vector constituted from these terms is

$$q = [0\ 1\ 0\ 0\ 1\ 0\ 0\ 1\ 0]$$

Now before we process the query, we project it on the reduced dimensional space and obtain the following representation:

$$q_4 = [0.7168 \quad 0.2412 \quad -0.6866 \quad -0.3054]$$

Now we process the reduced dimensional query in the approximated space and obtain the similarity as

$$\text{sim} = [-0.0642 \;\; 0.7168 \;\; 1.4033 \;\; 0.9579 \;\; 0.0302 \;\; -0.2109 \;\; 0.0302]$$

Generally the documents whose similarity values higher than some threshold value are considered as relevant to the given query. Considering a threshold limit of 0.5, we understand from the document similarity vector that the documents D2, D3, and D4 are relevant to the query and hence returned to the user.

In order to better understand the relation between SDD and SVD, discussed above, we verify this retrieval process using SVD. The SVD on the data matrix $\boldsymbol{A}$ produces rank k approximation as shown below:

$$\boldsymbol{A}_k = \boldsymbol{U}_k \boldsymbol{S}_k \boldsymbol{V}_k^{\mathrm{T}}$$

where the unitary projection matrices $\boldsymbol{U}_k$ and $\boldsymbol{V}_k^{\mathrm{T}}$ represent truncated left and right singular vectors of the original matrix, respectively. The matrix $\boldsymbol{S}_k$ holds the first k number of singular values of the original matrix. SVD provides the best approximation of the original matrix with regard to Frobenius norm. Generally SVD is regarded as one of the powerful decomposition technique since it provides all the fundamental spaces of the original matrix $\boldsymbol{A}$, i.e., the orthogonal basis for Range space and Null space of the matrices $\boldsymbol{A}$ and $\boldsymbol{A}^{\mathrm{T}}$ (Park and Elden 2003).

By applying SVD, we obtain rank 4 approximation of the column normalized term-document matrix. After processing the query in the reduced dimensional space, we obtain the document similarity vector as shown below:

$$\text{sim} = [0.0705 \;\; 1.2360 \;\; 1.6855 \;\; 0.3747 \;\; -0.0117 \;\; 0.0128 \;\; -0.0117]$$

From this vector documents D2 and D3 are relevant to the query, and hence they will be returned to the user. From this result we can understand that the SDD and SVD have commonly identified the documents D2 and D3. Now let us consider another query aimed at retrieving the documents on *Child Proofing* from this collection. The corresponding query vector would be

$$q = [0\ 1\ 0\ 0\ 0\ 0\ 1\ 0\ 0]$$

After projecting the query in the reduced dimensional space obtained using SDD, we compute the similarity of the document vectors. The following is the similarity vector of all the documents for the given query:

$$\text{sim} = [0.0642 \;\; 0.4779 \;\; 0.4779 \;\; 0.2367 \;\; 0.4779 \;\; 0.7190 \;\; 0.4779]$$

With the chosen threshold limit of 0.5, we understand that the document D6 is the only relevant document for this query. However, let us consider posing this query in the reduced dimensional space obtained using SVD. After processing, we get the following similarity:

$$\text{sim} = [-0.0721 \;\; 0.4872 \;\; 0.6307 \;\; 0.0730 \;\; 0.3690 \;\; 0.6903 \;\; 0.3690]$$

With the threshold value of 0.5, we get the documents D3 and D6 as relevant to the given query. In this case, the SDD and SVD have commonality in D6. For equal rank values of approximation, SDD requires significantly less number of floating point operations than SVD to process the query. On standard document collections, it is proved that SDD-based LSI retrieves documents similar to SVD-based LSI with a lesser time to process the query and lesser storage. These illustrative examples provide an understanding about the usage of SDD in some of the applications. Interested readers can explore the literature cited in section "Key Applications" for more details on the applications of SDD.

Acknowledgments

Author sincerely acknowledges the support from National Board of Higher Mathematics, Dept. of Atomic Energy, Govt. of India under the grant 2/48(11)/2010-R&D II/10806.

Cross-References

- Matrix Algebra, Basics of
- Eigenvalues, Singular Value Decomposition
- Matrix Decomposition
- Principal Component Analysis
- Spectral Analysis

References

Aswani Kumar Ch (2009) Analysis of unsupervised dimensionality reduction techniques. Comput Sci Inf Syst 6(2):217–227

Aswani Kumar Ch (2011) Reducing data dimensionality using random projections and fuzzy k-means clustering. Int J Intell Comput Cybern 4(3):353–365

Aswani Kumar Ch, Palanisamy R (2010) Comparison of matrix dimensionality reduction methods in uncovering latent structures in the data. J Inf Knowl Manag 9(1):81–92

Aswani Kumar Ch, Srinivas S (2006) Latent semantic indexing using eigenvalue analysis for efficient information retrieval. Int J Appl Math Comput Sci 16(4):551–558

Aswani Kumar Ch, Srinivas S (2010) A note on weighted fuzzy k-means clustering for concept decomposition. Cybern Syst 41:455–467

Berry MW, Browne M (2005) Understanding search engines: mathematical modeling and text retrieval. SIAM, Philadelphia

Berry MW, Drmac Z, Jessup ER (1999) Matrices, vector spaces and information retrieval. SIAM Rev 41(2):335–362

Cunningham P (2007) Dimension reduction. Technical report UCD-CSI-2007-7, School of computer science and informatics, University College, Dublin

Deerwester S, Dumais S, Furnas G, Landauer T, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6):391–407

Divya R, Aswani Kumar Ch, Saijanani S, Priyadharshini M (2011) Deceiving communication links on an organization email corpus. Malays J Comput Sci 24(1):17–33

Dobsa J, Praus P, Aswani Kumar Ch, Praks P (2012) Classification of hydrochemical data in reduced dimensional space. J Inf Org Sci 36(1):27–37

Elden L (2006) Numerical linear algebra in data mining. Acta Numer 15:327–384

Elden L (2007) Matrix methods in data mining and pattern recognition. SIAM, Philadelphia

Fodor IK (2002) A survey of dimension reduction techniques. Technical report, UCRL-148494, University of California

Keila PS, Skillicorn DB (2005) Structure in the Enron email dataset. Comput Math Org Theory 11(3):183–199

Knight GS, Carosielli L (2003) Detecting malicious use with unlabelled data using clustering and outlier analysis. In: Proceedings of 18th IFIP international conference on information security, Athens, pp 205–216

Kolda TG, O'Leary DP (1998) A semi-discrete matrix decomposition for latent semantic indexing in information retrieval. ACM Trans Inf Syst 16(4):322–346

Kolda TG, O'Leary DP (2000) Computation and uses of the semidiscrete matrix decomposition. ACM Trans Math Softw 26(3):415–435

McConnell S, Skillicorn DB (2002) Semidiscrete decomposition: a bump hunting technique. In: Proceedings of the Australian data mining workshop, Canberra, pp 75–82

Miettinen P (2009) Matrix decomposition methods for data mining: computational complexity and algorithms. Academic dissertation A-2009-04, Department of Computer Science, University of Helsinki, Finland, 2009

O'Leary D, Peleg S (1983) Digital image compression by outer product expansion. IEEE Trans Commun 31:441–444

Park H, Elden L (2003) Matrix rank reduction for data analysis and feature extraction. Technical report TR 03–015, University of Minnesota, 2003

Pilato G, Vassallo G, Gaglio S (2005) Wordnet and semi-discrete decomposition for sub-symbolic representation of words. In: Apolloni B et al (eds) Biological and artificial intelligence environments. Springer, Dordrecht, pp 191–198

Qiang W, XiaoLong L, Yi G (2004) A study of semi-discrete matrix decomposition for LSI in automated text categorization. In: Proceedings of 1st International joint conference on natural language processing. Springer, Berlin/Heidelberg, China, pp 606–615

Skillicorn DB (2004) Applying matrix decompositions to counterterrorism. Technical report, Queen's University, Kingston, 2004

Skillicorn DB (2007) Understanding complex datasets – data mining with matrix decompositions. Chapman & Hall/CRC, New York

Skillicorn DB, McConnell SM, Soong EY (2003) Handbook of data mining using matrix decompositions. School of Computing, Queen's University, Kingston

Snasel V, Moravec P, Pokorny J (2008) Using semi-discrete decomposition for topic identification. In: 8th International conference on intelligent systems design and applications, Kaohsiung, pp 415–420

Snasel V, Horak Z, Abraham A (2010) Link suggestions in terrorist networks using semi discrete decomposition. In: 6th International conference on information assurance and security, Atlanta, GA, pp 23–25

Zyto S, Grama A, Szpankowski W (2002) Semi-discrete matrix transforms for image and video compression. In: Proceedings of the data compression conference, Snowbird, UT, USA, p 484

Semiring

▸ Semirings and Matrix Analysis of Networks

Semirings and Matrix Analysis of Networks

Monika Cerinšek[1] and Vladimir Batagelj[2]
[1]Hruška d.o.o., Ljubljana, Slovenia
[2]Faculty of Mathematics and Physics, Department of Mathematics, University of Ljubljana, Ljubljana, Slovenia

Synonyms

Algebraic path problem; Matrix; Multiplication of vector and matrix; Network; Network multiplication; Semiring; Simple walk; Value matrix; Walk

Glossary

Algebraic Structure A set with one or more operations defined on it

Network Analysis A study of networks as representations of relations between discrete objects

Sparse Matrix A matrix with most of entries equal to zero

Large Network A network with several thousands or millions of nodes

Complete Graph K_n A network in which every pair of nodes is linked

Introduction

Semirings are an algebraic structure with two operations that provide the basic conditions for studying matrix addition and multiplication and path problems in networks. Several results and algorithms from different fields of application turn out to be just special cases over the corresponding semirings.

Semirings

Let $\mathbb{K}$ be a set and a, b, c elements from $\mathbb{K}$. A *semiring* (Abdali and Saunders 1985; Baras and Theodorakopoulos 2010; Batagelj 1994) is an algebraic structure $(\mathbb{K}, \oplus, \odot, 0, 1)$ with two binary operations addition $\oplus$ and multiplication $\odot$ where:

- $(\mathbb{K}, \oplus)$ is an abelian monoid with neutral element 0 (zero):

  $a \oplus b = b \oplus a$ commutativity
  $(a \oplus b) \oplus c = a \oplus (b \oplus c)$ associativity
  $a \oplus 0 = a$ existence of zero

- $(\mathbb{K}, \odot)$ is a monoid with neutral element 1 (unit):

  $(a \odot b) \odot c = a \odot (b \odot c)$ associativity
  $a \odot 1 = 1 \odot a = a$ existence of unit

- Multiplication $\odot$ distributes over addition $\oplus$:

  $a \odot (b \oplus c) = a \odot b \oplus a \odot c$
  $(b \oplus c) \odot a = b \odot a \oplus c \odot a$

In the expressions we assume precedence of multiplication over addition.

A semiring $(\mathbb{K}, \oplus, \odot, 0, 1)$ is *complete* iff the addition is well defined for countable sets of elements and the commutativity, associativity, and distributivity hold in the case of countable sets. These properties are generalized in this case; for example, the distributivity takes form

$$(\oplus_i a_i) \odot (\oplus_j b_j) = \oplus_i (\oplus_j (a_i \odot b_j))$$
$$= \oplus_{i,j} (a_i \odot b_j).$$

The addition is *idempotent* iff $a \oplus a = a$ for all $a \in \mathbb{K}$. In this case the semiring over a finite set $\mathbb{K}$ is complete.

A semiring $(\mathbb{K}, \oplus, \odot, 0, 1)$ is *closed* iff for the additional (unary) *closure* operation * it holds for all $a \in \mathbb{K}$:

$$a^* = 1 \oplus a \odot a^* = 1 \oplus a^* \odot a.$$

Different closures over the same semiring can exist. A complete semiring is always closed for the closure

$$a^* = \oplus_{k \in \mathbb{N}} a^k.$$

In a closed semiring we can also define a *strict closure* $\overline{a}$ by

$$\overline{a} = a \odot a^*.$$

In a semiring $(\mathbb{K}, \oplus, \odot, 0, 1)$ the *absorption law* holds iff for all $a, b, c \in \mathbb{K}$:

$$a \odot b \oplus a \odot c \odot b = a \odot b.$$

Because of distributivity it is sufficient to check the property $1 \oplus c = 1$ for all $c \in \mathbb{K}$.

Combinatorial Semiring $(\mathbb{N}, +, \cdot, 0, 1)$

This is the most commonly used semiring. Also some other sets are used: $\mathbb{R}, \mathbb{R}_0^+, \mathbb{Q}$. For $\overline{\mathbb{N}} = \mathbb{N} \cup \{\infty\}$, the semiring is closed for $a^* = \sum_{k \in \overline{\mathbb{N}}} a^k$ because it is a complete semiring. Another possible closure for $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ is $a^* = 1/(1 - a)$ for $a \neq 1, \infty$ and $0^* = 1, 1^* = \infty$, and $\infty^* = \infty$. This semiring is commutative because it holds $a \odot b = b \odot a$ for all a and b in the set. Combinatorial semiring is not an idempotent semiring.

Reachability Semiring $(\{0, 1\}, \vee, \wedge, 0, 1)$

The logical (Boolean) semiring is useful for solving the connectivity questions in networks. The multiplication is commutative and the absorption law holds. The reachability semiring is closed for $a^* = 1 \vee a \wedge a^* = 1$.

Shortest Paths Semiring $(\overline{\mathbb{R}}_0^+, \min, +, \infty, 0)$

The commutativity of multiplication holds in this semiring. The semiring is closed for $a^* = \min(0, a + a^*) = 0$ (0 is the smallest element in the set $\mathbb{R}_0^+$). Since $\min(0, a) = 0$, the absorption law also holds. For the set $\mathbb{N} \cup \{\infty\}$, the semiring is called tropical semiring. Another set is $\mathbb{R} \cup \{\infty\}$ and in this case the semiring is isomorphic $(x \mapsto -x)$ to max-plus semiring $(\mathbb{R} \cup \{-\infty\}, \max, +, -\infty, 0)$.

Pathfinder Semiring $(\overline{\mathbb{R}}_0^+, \min, \boxed{\mathrm{r}}, \infty, 0)$

The Pathfinder semiring (Schvaneveldt et al. 1988) is a special case from the family of the semirings obtained as follows. Let $B \subseteq \overline{\mathbb{R}}$ be such that $(B, +, \cdot, 0, 1)$ or $(B, \min, +, U, 0)$ is a semiring $(U = \max(B))$. Therefore $0 \in B$ and $1 \in B$. Let $A \subseteq \mathbb{R}$ be such that $g : A \to B$ is a bijection. Let us define operations $\oplus, \nabla, \odot$ so that g is a homomorphism:

$$\begin{aligned} g(a \oplus b) &= g(a) + g(b), \\ g(a \nabla b) &= \min(g(a), g(b)), \\ g(a \odot b) &= g(a) \cdot g(b). \end{aligned}$$

This is equivalent to

$$\begin{aligned} a \oplus b &= g^{-1}(g(a) + g(b)), \\ a \nabla b &= g^{-1}(\min(g(a), g(b))), \\ a \odot b &= g^{-1}(g(a) \cdot g(b)). \end{aligned}$$

The function g^{-1} is also a homomorphism. If g is strictly increasing function, then $a \nabla b = g^{-1}(\min(g(a), g(b))) = \min(a, b)$. Since the homomorphisms preserve the algebraic properties, also the structure $(A, \oplus, \odot, g^{-1}(0), g^{-1}(1))$, $A \subseteq \overline{\mathbb{R}}$, is a semiring.

For $g(x) = x^r, g^{-1}(y) = \sqrt[r]{y}$, we get the *Pathfinder semiring* $(\overline{\mathbb{R}}_0^+, \min, \boxed{\mathrm{r}}, \infty, 0)$. The multiplicative operation is the *Minkowski operation* $a \boxed{\mathrm{r}} b = \sqrt[r]{a^r + b^r}$. This semiring is closed for $a^* = 0$ and the absorption law holds in it.

In Pathfinder algorithm the value r for the Minkowski operation is selected according to dissimilarity measure. For a value $r = 1$, the semiring is the shortest path semiring, and for a value $r = \infty$, the semiring is min–max semiring.

Several other examples of semirings can be found in Carré (1979), Burkard et al. (1984),

Gondran and Minoux (2008), Baras and Theodorakopoulos (2010), and Kepner and Gilbert (2011).

Matrices

A $m \times n$ matrix $\mathbf{A}$ over a set $\mathbb{K}$ is a rectangular array of elements from the set $\mathbb{K}$ that consists of m rows and n columns. The entry in ith row and jth column is denoted by a_{ij}. If $m = n$ the matrix $\mathbf{A}$ is called a *square* matrix. The matrix with all entry values equal to 0 is called the *zero* matrix and is denoted by $\mathbf{O}_{mn}$.

The *transpose* of matrix $\mathbf{A}$ is a matrix $\mathbf{A}^T$ in which the rows of $\mathbf{A}$ are written as the columns of $\mathbf{A}^T$: $a_{ij}^T = a_{ji}$. A square matrix $\mathbf{A}$ is *symmetric* if $\mathbf{A} = \mathbf{A}^T$.

A *diagonal matrix* is a square matrix $\mathbf{A}$ such that only diagonal elements are nonzero: $a_{ij} = 0$, $i \neq j$. If $a_{ii} = 1, i = 1, \ldots, n$, this matrix is called the *identity* matrix $\mathbf{I}_n$ of order n. The matrix $\mathbf{A}$ is *upper triangular* if $a_{ij} = 0, i > j$, and its transpose is the *lower triangular* matrix.

Let $\mathcal{M}_{mn}(\mathbb{K})$ be a set of matrices of order $m \times n$ over the semiring $(\mathbb{K}, \oplus, \odot, 0, 1)$ in which we additionally require

$$\forall s \in \mathbb{K} : s \odot 0 = 0 \odot s = 0$$

and let $\mathcal{M}(\mathbb{K})$ be a set of all matrices over the $\mathbb{K}$. The operations $\oplus$ and $\odot$ can be extended to the $\mathcal{M}(\mathbb{K})$:

$$\mathbf{A}, \mathbf{B} \in \mathcal{M}_{mn}(\mathbb{K}) : \mathbf{A} \oplus \mathbf{B} = [a_{uv} \oplus b_{uv}] \in \mathcal{M}_{mn}(\mathbb{K})$$

$$\mathbf{A} \in \mathcal{M}_{mk}(\mathbb{K}), \mathbf{B} \in \mathcal{M}_{kn}(\mathbb{K}) :$$

Then $\mathbf{A} \odot \mathbf{B} = [\oplus_{t=1}^{k} a_{ut} \odot b_{tv}] \in \mathcal{M}_{mn}(\mathbb{K})$.

- $(\mathcal{M}_{mn}(\mathbb{K}), \oplus, \mathbf{O}_{mn})$ is an abelian monoid.
- $(\mathcal{M}_{n^2}(\mathbb{K}), \odot, \mathbf{I}_n)$ is a monoid.
- $(\mathcal{M}_{n^2}(\mathbb{K}), \oplus, \odot, \mathbf{O}_n, \mathbf{I}_n)$ is a semiring.

For matrices $\mathbf{A}$ and $\mathbf{B}$, it holds

$$(\mathbf{A} \odot \mathbf{B})^T = \mathbf{B}^T \odot \mathbf{A}^T.$$

Network Multiplication

A (simple directed) network $\mathcal{N}$ is an ordered pair of sets $(\mathcal{V}, \mathcal{A})$ where $\mathcal{V}$ is the set of nodes and $\mathcal{A} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of arcs. We assume that the set of nodes is finite $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$. Let $\mathcal{N} = ((\mathcal{I}, \mathcal{J}), \mathcal{A}, w)$ be a *simple two-mode network*, where $\mathcal{I}$ and $\mathcal{J}$ are disjoint (sub)sets of nodes ($\mathcal{V} = \mathcal{I} \cup \mathcal{J}, \mathcal{I} \cap \mathcal{J} = \emptyset$), $\mathcal{A}$ is a set of arcs linking $\mathcal{I}$ and $\mathcal{J}$, and the mapping $w : \mathcal{A} \to \mathbb{K}$ is the *arcs value function* called also a *weight*. We can assign to the network its *value matrix* $\mathbf{W} = [w_{i,j}]$ with elements

$$w_{ij} = \begin{cases} w((i,j)) & (i,j) \in \mathcal{A} \\ 0 & \text{otherwise.} \end{cases}$$

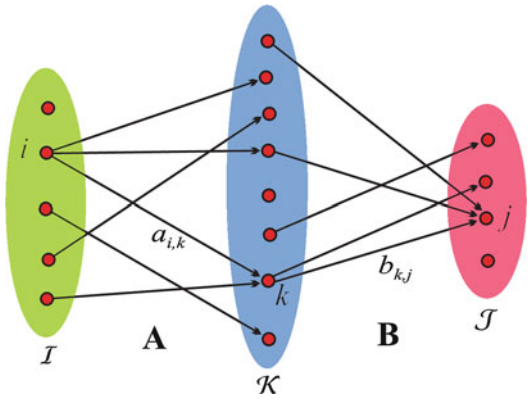
The problem with value matrices in computer applications is their size. The value matrices of large networks are sparse. There is no need to store the zero values in a matrix, and different data structures can be used for saving and working with value matrices: special dictionaries and lists.

Let $\mathcal{N}_\mathbf{A} = ((\mathcal{I}, \mathcal{K}), \mathcal{A}_\mathbf{A}, w_\mathbf{A})$ and $\mathcal{N}_\mathbf{B} = ((\mathcal{K}, \mathcal{J}), \mathcal{A}_\mathbf{B}, w_\mathbf{B})$ be a pair of networks with corresponding matrices $\mathbf{A}_{\mathcal{I}\times\mathcal{K}}$ and $\mathbf{B}_{\mathcal{K}\times\mathcal{J}}$, respectively. Assume also that $w_\mathbf{A} : \mathcal{A}_\mathbf{A} \to \mathbb{K}$, $w_\mathbf{B} : \mathcal{A}_\mathbf{B} \to \mathbb{K}$ and $(\mathbb{K}, \oplus, \odot, 0, 1)$ is a semiring. We say that such networks/matrices are *compatible*. The *product* $\mathcal{N}_\mathbf{A} \star \mathcal{N}_\mathbf{B}$ of networks $\mathcal{N}_\mathbf{A}$ and $\mathcal{N}_\mathbf{B}$ is a network $\mathcal{N}_\mathbf{C} = ((\mathcal{I}, \mathcal{J}), \mathcal{A}_\mathbf{C}, w_\mathbf{C})$ for $\mathcal{A}_\mathbf{C} = \{(i,j); i \in \mathcal{I}, j \in \mathcal{J}, c_{ij} \neq 0\}$ and $w_\mathbf{C}(i,j) = c_{ij}$ for $(i,j) \in \mathcal{A}_\mathbf{C}$, where $\mathbf{C} = [c_{ij}]_{\mathcal{I}\times\mathcal{J}} = \mathbf{A} \odot \mathbf{B}$. If all three sets of nodes are the same ($\mathcal{I} = \mathcal{K} = \mathcal{J}$), we are dealing with ordinary one-mode networks with square matrices.

When do we get an arc in the product network? Let's look at the definition of the matrix product

$$c_{ij} = \oplus_{k \in K} a_{ik} \odot b_{kj}.$$

There is an arc $(i,j) \in \mathcal{A}_\mathbf{C}$ if c_{ij} is nonzero. Therefore at least one term $a_{ik} \cdot b_{kj}$ is nonzero, but this means that both a_{ik} and b_{kj} should be nonzero, and thus $(i,k) \in \mathcal{A}_\mathbf{A}$ and $(k,j) \in \mathcal{A}_\mathbf{B}$ (see Fig. 1):



Semirings and Matrix Analysis of Networks, Fig. 1 Multiplication of networks

$$c_{ij} = \oplus_{k \in N_{\mathbf{A}}(i) \cap N_{\mathbf{B}}^{-}(j)} a_{ik} \odot b_{kj},$$

where $N_{\mathbf{A}}(i)$ are the *successors* of node i in network $\mathcal{N}_{\mathbf{A}}$ and $N_{\mathbf{B}}^{-}(j)$ are the *predecessors* of node j in network $\mathcal{N}_{\mathbf{B}}$. The value of the entry $c_{i,j}$ equals to the value of all paths (of length 2) from $i \in \mathcal{I}$ to $j \in \mathcal{J}$ passing through some node $k \in \mathcal{K}$.

The standard procedure to compute the product of matrices $\mathbf{A}_{\mathcal{I}\times\mathcal{K}}$ and $\mathbf{B}_{\mathcal{K}\times\mathcal{J}}$ has the complexity $O(|\mathcal{I}| \cdot |\mathcal{K}| \cdot |\mathcal{J}|)$ and is therefore too slow to be used for large networks. Since the matrices of large networks are usually sparse, we can compute the product of two networks much faster considering only nonzero entries (Batagelj and Mrvar 2008; Batagelj and Cerinšek 2013):

for k in $\mathcal{K}$ **do**
  for $i \in N_{\mathbf{A}}^{-}(k)$ **do**
    for $j \in N_{\mathbf{B}}(k)$ do
      if $\exists c_{ij}$ then $c_{ij} := c_{ij} \oplus a_{ik} \odot b_{kj}$
      else $c_{ij} := a_{ik} \odot b_{kj}$.

In general the multiplication of large sparse network is "dangerous" operation since the result can "explode" – it is not sparse.

From the network multiplication algorithm, we see that each intermediate node $k \in \mathcal{K}$ adds to a product network a complete two-mode subnetwork $K_{N_{\mathbf{A}}^{-}(k),N_{\mathbf{B}}(k)}$ (or, in the case $\mathbf{A} = \mathbf{B}$, a complete subnetwork $K_{N(k)}$). If both degrees $\deg_A(k) = |N_{\mathbf{A}}^{-}(k)|$ and $\deg_B(k) = |N_{\mathbf{B}}(k)|$ are large, then already the computation of this complete subnetwork has a quadratic (time and space) complexity – the result "explodes."

If for the sparse networks $\mathcal{N}_{\mathbf{A}}$ and $\mathcal{N}_{\mathbf{B}}$, there are in $\mathcal{K}$ only few nodes with large degree and no one among them with large degree in both networks, then also the resulting product network $\mathcal{N}_{\mathbf{C}}$ is sparse.

The Algebraic Path Problem

The use of special semiring and a multiplication of network can lead us to the essence of the shortest path problem (Baras and Theodorakopoulos 2010). Many other network problems can be solved by replacing the usual addition and multiplication with the corresponding operations from an appropriate semiring.

Let $\mathcal{N} = (\mathcal{V}, \mathcal{A}, w)$ be a network where $w : \mathcal{A} \to \mathbb{K}$ is the value (weight) of arcs such that $(\mathbb{K}, \oplus, \odot, 0, 1)$ is a semiring. We will denote the number of nodes as $n = |\mathcal{V}|$ and the number of arcs as $m = |\mathcal{A}|$.

A finite sequence of nodes $\sigma = (u_0, u_1, u_2, \ldots, u_{p-1}, u_p)$ is a *walk* of *length* p on $\mathcal{N}$ iff every pair of neighboring nodes is linked: $(u_{i-1}, u_i) \in \mathcal{A}, i = 1, \ldots, p$. A finite sequence σ is a *semiwalk* or chain on $\mathcal{N}$ iff every pair of neighboring nodes is linked neglecting the direction of an arc $(u_{i-1}, u_i) \in \mathcal{A} \vee (u_i, u_{i-1}) \in \mathcal{A}, i = 1, \ldots, p$. The (semi)walk is *closed* iff its end nodes coincide: $u_0 = u_p$. A walk is *simple* or a *path* iff no node repeats in it. If the ends of a simple walk coincide, it is called a *cycle*.

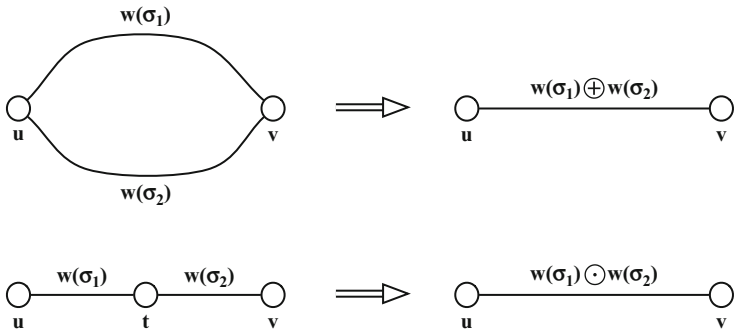
We can extend the value function w to walks and sets of walks on $\mathcal{N}$ by the following rules (see Fig. 2):

- Let $\sigma_v = (v)$ be a null walk in the node $v \in \mathcal{V}$; then $w(\sigma_v) = 1$.
- Let $\sigma = (u_0, u_1, u_2, \ldots, u_{p-1}, u_p)$ be a walk of length $p \geq 1$ on $\mathcal{N}$; then

$$w(\sigma) = \odot_{i=1}^{k} w(u_{i-1}, u_i).$$

- For empty set of walks $\emptyset$, it holds $w(\emptyset) = 0$.
- Let $\mathcal{S} = \{\sigma_1, \sigma_2, \ldots\}$ be a set of walks in $\mathcal{N}$; then

$$w(\mathcal{S}) = \oplus_{\sigma \in \mathcal{S}} w(\sigma).$$



Semirings and Matrix Analysis of Networks, Fig. 2 Semiring operations and values of walks

Let σ_1 and σ_2 be compatible walks on $\mathcal{N}$ – the end node of the walk σ_1 is equal to the start node of the walk σ_2. Such walks can be concatenated in a new walk $\sigma_1 \circ \sigma_2$ for which holds

$$w(\sigma_1 \circ \sigma_2) = \begin{cases} w(\sigma_1) \odot w(\sigma_2) & \sigma_1 \text{ and } \sigma_2 \text{ are compatible} \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be finite sets of walks; then

$$w(\mathcal{S}_1 \cup \mathcal{S}_2) \oplus w(\mathcal{S}_1 \cap \mathcal{S}_2) = w(\mathcal{S}_1) \oplus w(\mathcal{S}_2).$$

In the special case when $\mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset$, it holds $w(\mathcal{S}_1 \cup \mathcal{S}_2) = w(\mathcal{S}_1) \oplus w(\mathcal{S}_2)$. Also the concatenation of walks can be generalized to sets of walks:

$$\mathcal{S}_1 \circ \mathcal{S}_2 = \{\sigma_1 \circ \sigma_2 : \sigma_1 \in \mathcal{S}_1, \sigma_2 \in \mathcal{S}_2, \sigma_1 \text{ and } \sigma_2 \text{ are compatible}\}.$$

It also holds $\mathcal{S} \circ \emptyset = \emptyset \circ \mathcal{S} = \emptyset$.

We denote by:

- $\mathcal{S}_{uv}^k$ the set of all walks of length k from node u to node v
- $\mathcal{S}_{uv}^{(k)}$ the set of all walks of length at most k from node u to node v
- $\mathcal{S}_{uv}^*$ the set of all walks from node u to node v
- $\overline{\mathcal{S}}_{uv}$ the set of all nontrivial walks from node u to node v
- $\mathcal{E}_{uv}$ the set of all simple walks (paths) from node u to node v

The following relations hold among these sets:

$$\mathcal{S}_{uv}^k \subseteq \mathcal{S}_{uv}^{(k)} \subseteq \mathcal{S}_{uv}^*$$

$$k \neq l \Leftrightarrow \mathcal{S}_{uv}^k \cap \mathcal{S}_{uv}^l = \emptyset$$

$$\mathcal{S}_{uv}^{(k)} = \bigcup_{i=0}^{k} \mathcal{S}_{uv}^i, \mathcal{S}_{uv}^* = \bigcup_{k=0}^{\infty} \mathcal{S}_{uv}^k$$

$$k \geq |\mathcal{V}| - 1 : \mathcal{E}_{uv} \subseteq \mathcal{S}_{uv}^{(k)}$$

$$w(\mathcal{S}_{uv}^{(k)}) = \sum_{i=0}^{k} w(\mathcal{S}_{uv}^i).$$

A set of walks $\mathcal{S}$ is *uniquely factorizable* to sets of walks $\mathcal{S}_1$ and $\mathcal{S}_2$ if $\mathcal{S} = \mathcal{S}_1 \circ \mathcal{S}_2$, and for all walks $\sigma_1, \sigma_1' \in \mathcal{S}_1, \sigma_2, \sigma_2' \in \mathcal{S}_2, \sigma_1 \neq \sigma_1', \sigma_2 \neq \sigma_2'$, it holds $\sigma_1 \circ \sigma_2 \neq \sigma_1' \circ \sigma_2'$.

For example, for s, $0 < s < k$, the nonempty set $\mathcal{S}_{uv}^k$ is uniquely factorizable to sets $\mathcal{S}_{u\bullet}^s$ and $\mathcal{S}_{\bullet v}^{k-s}$, where $\mathcal{S}_{u\bullet}^s = \bigcup_{t \in \mathcal{V}} \mathcal{S}_{ut}^s$, etc.

Theorem 1 *Let the finite set $\mathcal{S}$ be uniquely factorizable for $\mathcal{S}_1$ and $\mathcal{S}_2$ or a semiring be complete. Then it holds*

$$w(\mathcal{S}_1 \circ \mathcal{S}_2) = w(\mathcal{S}_1) \odot w(\mathcal{S}_2).$$

The kth *power* $\mathbf{W}^k$ of any square matrix $\mathbf{W}$ over $\mathbb{K}$ is unique because of associativity.

Theorem 2 *The entry w_{uv}^k of kth power $\mathbf{W}^k$ of value matrix $\mathbf{W}$ is equal to the value of all walks of length k from node u to node v:*

$$w(S_{uv}^k) = \mathbf{W}^k[u, v] = w_{uv}^k.$$

Therefore if a network $\mathcal{N}$ is acyclic, then it holds for a value matrix $\mathbf{W}$:

$$\exists k_0 < n : \forall k > k_0 : \mathbf{W}^k = 0,$$

k_0 is the length of the longest walk in the network.

If $\mathbf{W}$ is the network adjacency matrix over the combinatorial semiring, the entry w_{uv}^k counts the number of different walks of length k from u to v.

Let us denote

$$\mathbf{W}^{(k)} = \sum_{i=0}^{k} \mathbf{W}^i.$$

In an idempotent semiring, it holds $\mathbf{W}^{(k)} = (1 + \mathbf{W})^k$.

Theorem 3

$$w(S_{uv}^{(k)}) = \mathbf{W}^{(k)}[u, v] = w_{uv}^{(k)}.$$

For the combinatorial semiring and $\mathbf{W}$ is the network adjacency matrix, the entry $w_{uv}^{(k)}$ counts the number of different walks of length at most k from u to v.

The matrix semiring over a complete semiring is also complete and therefore closed for $\mathbf{W}^* = \oplus_{k=0}^{\infty} \mathbf{W}^k$.

Theorem 4 *For a value matrix* $\mathbf{W}$ *over a complete semiring with closure* $\mathbf{W}^*$ *and strict closure* $\overline{\mathbf{W}}$ *hold:*

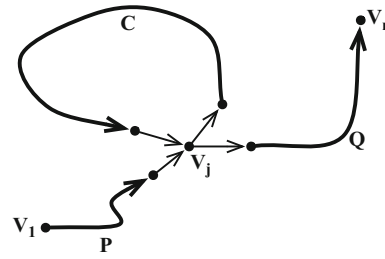
$$w(\mathcal{S}_{uv}^*) = \mathbf{W}^*[u, v] = w_{uv}^* \qquad \text{and}$$
$$w(\overline{\mathcal{S}}_{uv}) = \overline{\mathbf{W}}[u, v] = \overline{w}_{uv}.$$

For the reachability semiring and $\mathbf{W}$ is the network adjacency matrix, the matrix $\overline{\mathbf{W}}$ is its transitive closure.

For the shortest paths semiring and $\mathbf{W}$ is the network value matrix, the entry w_{uv}^* is the value of the shortest path from u to v.

The paper Quirin et al. (2008) could be essentially reduced to the observation that the structure $(\overline{\mathbb{R}}_0^+, \min, \boxed{\mathrm{r}}, \infty, 0)$ is a (Pathfinder) complete semiring.

Let $(\mathbb{K}, \oplus, \odot, 0, 1)$ be an absorptive semiring and σ be a nonsimple walk from a set $\mathcal{S}_{uv}^{(k)}$. Therefore at least one node v_j appears more than once in σ. The part of a walk between its first and last appearance is a closed walk C (see Fig. 3). The whole walk can be written as $\sigma = P \circ C \circ Q$ where P is the initial segment of σ from u to the first appearance of v_j, and Q is the terminal segment of σ from the last appearance of v_j to v. Note that $P \circ Q$ is also a walk. The value of both walks together is



Semirings and Matrix Analysis of Networks, Fig. 3 Example of a walk that is not a path

$$w(\{P \circ Q, P \circ C \circ Q\}) = w(P \circ Q).$$

We see that the walks that are not paths do not contribute to the value of walks. Therefore

$$w(\mathcal{S}_{uv}^*) = w(\mathcal{E}_{uv}).$$

Equality holds also for $\mathcal{S}_{uv}^* = \emptyset$.

Since the node set $\mathcal{V}$ is finite, also the set $\mathcal{E}_{uv}$ is finite which allows us to compute the value $w(S_{uv}^*)$. We already know that $\mathbf{W}^* = \mathbf{W}^{(k)} = (\mathbf{1} + \mathbf{W})^k$ for k large enough.

To compute the closure matrix $\mathbf{W}^*$ of a given matrix over a complete semiring $(\mathbb{K}, \oplus, \odot, 0, 1)$, we can use the Fletcher's algorithm (Fletcher 1980):

$\mathbf{C}_0 = \mathbf{W}$
for $k := 1, \ldots n$ **do**
  for $i := 1, \ldots n$ **do**
    for $j := 1, \ldots n$ **do**
      $c_k[i, j] := c_{k-1}[i, j] \oplus c_{k-1}[i, k]$
      $\odot (c_{k-1}[k, k])^* \odot c_{k-1}[k, j]$
    $c_k[k, k] := 1 \oplus c_k[k, k]$
$\mathbf{W}^* := \mathbf{W}_n$.

If we delete the statement $c_k[k,k] := 1 \oplus c_k[k,k]$, we obtain the algorithm for computing the strict closure $\overline{\mathbf{W}}$. If the addition $\oplus$ is idempotent, we can compute the closure matrix in place – we omit the subscripts in matrices $\mathbf{C}_k$.

The Fletcher's algorithm is a generalization of a sequence of algorithms (Kleene, Warshall, Floyd, Roy) for computing closures on specific semirings.

Multiplication of Matrix and Vector

Let e_i be a unit vector of length n – the only nonzero element is at the ith position and it is equal to 1. It is essentially a $1 \times n$ matrix. The product of a unit vector and a value matrix of a network can be used to calculate the value of walks from a node i to all the other nodes.

Let us denote

$$q_1^T = e_i^T \odot \mathbf{W}.$$

The values of elements of the vector q_1 are equal to the values of walks of the length 1 from a node i to all other nodes: $q_1[j] = w(\mathcal{S}_{ij}^1)$. We can calculate iteratively the values of all walks of the length s, $s = 2, 3, \ldots k$ that start in the node i:

$$q_s^T = q_{s-1}^T \odot \mathbf{W}$$

or $q_s^T = e_i^T \odot \mathbf{W}^s$ and $q_s[j] = w(\mathcal{S}_{ij}^s)$. Similarly we get ${q^{(k)}}^T = e_i^T \odot \mathbf{W}^{(k)}$, $q^{(k)}[j] = w(\mathcal{S}_{ij}^{(k)})$ and ${q^*}^T = e_i^T \odot \mathbf{W}^*$, $q^*[j] = w(\mathcal{S}_{ij}^*)$.

This can be generalized as follows. Let $\mathcal{I} \subseteq \mathcal{V}$ and $e_\mathcal{I}$ is the characteristic vector of the set $\mathcal{I}$ – it has value 1 for elements of $\mathcal{I}$ and is 0 elsewhere. Then, for example, for $q_k^T = e_\mathcal{I}^T \odot \mathbf{W}^k$, it holds $q_k[j] = w(\bigcup_{i \in \mathcal{I}} \mathcal{S}_{ij}^k)$.

Acknowledgments

The first author was financed in part by the European Union, European Social Fund. The work was supported in part by grant within the EUROCORES Programme EUROGIGA (project GReGAS) of the European Science Foundation.

Cross-References

- Eigenvalues, Singular Value Decomposition
- Iterative Methods for Eigenvalues/Eigenvectors
- Markov Chain Monte Carlo Model
- Matrix Algebra, Basics of
- Spectral Analysis

References

Abdali SK, Saunders BD (1985) Transitive closure and related semiring properties via eliminants. Theor Comput Sci 40:257–274

Baras JS, Theodorakopoulos G (2010) Path problems in networks. Morgan & Claypool, Berkeley

Batagelj V (1994) Semirings for social networks analysis. J Math Soc 19(1):53–68

Batagelj V, Cerinšek M (2013) On bibliographic networks. Scientometrics 96(3):845–864

Batagelj V, Mrvar A (2008) Analysis of kinship relations with Pajek. Soc Sci Comput Rev 26(2): 224–246

Burkard RE, Cuninghame-Greene RA, Zimmermann U (eds) (1984) Algebraic and combinatorial methods in operations research. Annals of discrete mathematics, vol 19. North Holland, Amsterdam/New York

Carré B (1979) Graphs and networks. Clarendon, Oxford

Fletcher JG (1980) A more general algorithm for computing closed semiring costs between vertices of a directed graph. Commun ACM 23(6): 350–351

Gondran M, Minoux M (2008) Graphs, dioids and semirings: new models and algorithms. Springer, New York

Kepner J, Gilbert J (2011) Graph algorithms in the language of linear algebra. SIAM, Philadelphia

Quirin A, Cordón O, Santamaria J, Vargas-Quesada B, Moya-Anegón F (2008) A new variant of the Pathfinder algorithm to generate large visual science maps in cubic time. Inf Process Manag 44(4): 1611–1623

Schvaneveldt RW, Dearholt DW, Durso FT (1988) Graph theoretic foundations of Pathfinder networks. Comput Math Appl 15(4):337–345

Sentiment Analysis

▸ Multi-classifier System for Sentiment Analysis and Opinion Mining
▸ User Sentiment and Opinion Analysis

Sentiment Analysis in Social Media

Chenghua Lin[1] and Yulan He[2]
[1]Department of Computing Science, University of Aberdeen, Aberdeen, UK
[2]School of Engineering and Applied Science, Aston University, Birmingham, UK

Synonyms

Data mining; Knowledge discovery; Opinion mining; Sentiment classification; Social media analysis

Glossary

NB Naive Bayes classifier
SVM Support vector machines
MaxEnt Maximum entropy classifier
PMI Point-wise mutual information
POS Part of speech
SO Sentiment orientation

Definition

Sentiment analysis aims to understand subjective information such as opinions, attitudes, and feelings expressed in text. Sentiment analysis tasks include, but not limited to the following:

- **Sentiment classification** which classifies a given piece of text as positive, negative, or neutral.
- **Opinion retrieval** which retrieves opinions in relevance to a specific topic or query.
- **Opinion summarization** which summarizes opinions over multiple text sources towards a certain topic.
- **Opinion holder identification** which identifies who express a specific opinion.
- **Topic/sentiment dynamics tracking** which aims to track sentiment and topic changes over time.
- **Opinion spam detection** which identifies fake/untruthful opinions.
- **Prediction** which predicts people's behaviors, market trends, political election outcomes, etc., based on opinions or sentiments expressed in online contents.

Introduction

With the explosion of people's attitudes and opinions expressed in social media such as blogs, discussion forums, and tweets, detecting sentiment or opinion from the Web is becoming an increasingly popular way of interpreting data. Sentiment analysis in social media allows business organizations to monitor their reputations, find public opinions about their products or services and those of their competitors, and provide them with insight into emerging trends and potential changes in market opinion, etc.

Customers also rely on online reviews to make more informed purchase decisions. Taking the Amazon Kindle cover reviews shown in Fig. 1 as an example, this Kindle cover receives a very high average rating of 4.5 stars from a total number of 855 reviews. Nevertheless, some reviews with high star ratings might still contain negative comments. Two example 4-star reviews shown in Fig. 1 reveal that although people think the design and quality of the cover are very good, it is overpriced. With such information, the cover would still be a good buy for price-insensitive customers. However, other customers may choose a less expensive alternative. With the sheer volume of social media data published every day on the Web and driven by the demand of gleaning insights into such great amounts user-generated data, there have been a large

Review 1

Title: **Lovely quality** (4-star)

By Technophobe, 19 April 2011

Beautiful piece of kit, protects my beloved kindle from knocks, scratches etc and looks very good at the same time. The locking mechanism that secures the kindle into the cover is very clever and looks to be safe and secure. The leather is good quality, and I love the bright apple green - very chic and smart. Only reason its not 5 stars is it wasn't exactly cheap - bring the price down a few pounds and it would perhaps represent better value for money and earn it 5 stars. No regrets about buying it though, does what its meant to and looks good at the same time!

Review 2

Title: **Very good except the price** (4-star)

By Val, 20 June 2011

The cover is very good, clips onto the Kindle easily and great protection when being transported. It makes holding and reading the Kindle so much more natural, like reading a book. I do however think the price is too high, although good quality there isn't an enormous amount of leather used.

Sentiment Analysis in Social Media, Fig. 1 Amazon Kindle cover reviews. Texts highlighted in *green* and *red* indicate the pros and cons of the product respectively

number of sentiment analysis software tools developed for alleviating users' information seeking burden.

Bazaarvoice's ratings and review platform (http://www.bazaarvoice.co.uk/) enables brands to capture customer's opinions about their services and products. Lightweight tools, such as Tweetfeel (http://www.tweetfeel.com), Twendz (http://twitter.com/Twendz), and Twitrratr (http://twitrratr.com/), scours Twitter for tweets and show how positively or negatively Twitter users feel about a particular topic. The Financial Times also introduced Newssift (http://www.newssift.com), a search tool that matches business topics to users' queries, sorts articles into positive and negative sentiment, and identifies the people, companies, places, and connections across all stories allowing for further refined search.

This article primarily focuses on sentiment classification from social media data. It describes some of the prominent approaches to sentiment classification including corpus-based approaches, lexicon-based approaches, and the incorporation of social networks into sentiment classification.

Historical Background

In the past, the majority of work in text information processing focused on mining and retrieving factual information, such as classifying documents according to their subject matter (e.g., politics vs. religion and sports vs. arts). In recent years, there has been a rapid growth of research interests in natural language processing that seeks to better understand sentiment or opinion expressed in text. One reason is that with the rise of various types of social media, communicating on the Web has become increasingly popular, where millions of people broadcast their thoughts and opinions on a great variety of topics, such as feedbacks on products and services, opinions on

political development and events, and information sharing on global disasters. Therefore, new computational tools are needed to help organize, summarize and understand this vast amount of information. Additionally, the discovery of opinions reflecting people's attitudes towards various topics enables many useful applications, which is another motivation of sentiment analysis.

Sentiment analysis can be considered as computational treatments of subjective information such as opinions and emotions expressed in text. In the simplest setting, sentiment analysis aims to automatically identify whether a given piece of text expresses positive or negative opinion. Early approaches (Pang et al. 2002; Matsumoto et al. 2005) view sentiment classification as a text classification problem where a corpus with sentiment orientation annotated is required for classifiers training. Supervised sentiment classification approaches usually perform well when the training set is large enough, where the state-of-the-art approach (Matsumoto et al. 2005) can achieve more than 90 % accuracy on the movie review data. However, there are some noticeable issues. One is that supervised classifier trained on one domain often fails to produce satisfactory performance when tested on other domains, and secondly, online content varies widely in domains and evolves rapidly over time, making corpora annotation for each domain unrealistic.

In response to the domain transfer and labeling cost problems faced by supervised approaches, there has been rising interest in exploring semi-supervised methods leveraging a large amount of unlabeled data and a small amount of labeled data for classifier training (Aue and Gamon 2005; Blitzer et al. 2007). Some representative works in this line are that of Aue and Gamon (2005) which explored various strategies for training SVM classifiers for the target domain lacking sufficient labeled data and that of Blitzer et al. (2007), which addressed the domain transfer problem with structural correspondence learning (SCL).

Unsupervised or weakly supervised approaches are mostly lexicon based which do not require labeled document for training. Instead, they assume that the sentiment orientation of a document is an averaged sum of the sentiment orientations of its words and phrases. Given the difficulties of supervised and semi-supervised sentiment analysis, it is conceivable that unsupervised or weakly supervised approaches to sentiment classification are even more challenging. Nevertheless, solutions to unsupervised or weakly supervised sentiment classification are of practical significance owing to its domain-independent nature.

The pioneer work is the point-wise mutual information (PMI) approach proposed in (Turney 2002), who calculated the sentiment orientations of phrases in documents as its PMI with a positive prototype "excellent" minus the PMI with a negative prototype "poor." The proposed approach achieved an accuracy of 84 % for automobile reviews and 66 % for movie reviews. Also work such as Read and Carroll (2009) are good examples of lexical-based approach.

Weakly supervised sentiment classification approaches are similar to unsupervised approaches in that they do not require labeled documents for training. Instead, they typically incorporate supervision information either from sentiment lexicons containing a list of words marked as positive or negative (usually much larger in size than the sentiment seed words used in unsupervised approaches) or from user feedbacks. Lin and He (2009) proposed a joint sentiment–topic (JST) model to detect document-level sentiment and extract sentiment bearing topics simultaneously from text. By incorporating a small set of domain-independent sentiment words as prior knowledge for model learning, the weakly supervised JST model is able to achieve comparable performance to semi-supervised approaches with 40 % labeled data.

Compared to the vast majority of work in sentiment analysis mainly focusing on the domains of product reviews and blogs, Twitter sentiment analysis is considered as a much harder problem than sentiment analysis on conventional text. This is mainly due to a few factors including the short length of tweet messages, the frequent use of informal and irregular words, and the rapid evolution of language in Twitter. Annotated tweets data are impractical to obtain. Previous work on twitter sentiment analysis

(Go et al. 2009; Pak and Paroubek 2010; Barbosa and Feng 2010) relies on noisy labels or distant supervision, for example, by taking emoticons as the indication of tweet sentiment to train supervised classifiers. Other work explore feature engineering in combination of machine learning methods to improve sentiment classification accuracy on tweets (Agarwal et al. 2011; Kouloumpis et al. 2011).

Prominent Methodologies

Research on sentiment classification has attracted a great deal of attention, where different classification tasks focus on various levels of granularity, e.g., from the document level (Pang et al. 2002) to the finer-grained sentence and word/phrase level (Turney and Littman 2002). In this section, we investigate the work which deals with computational treatments of sentiment using corpus-based and lexicon-based approaches, with a focus on document-level sentiment classification.

Corpus-Based Approaches

Corpus-based approaches (Pang et al. 2002; Pang and Lee 2004; Boiy et al. 2007) rely on annotated corpora where each document is annotated with a polarity label such as positive, negative, and neutral. Standard classifiers such as Naïve Bayes (NB), Maximum Entropy (MaxEnt), and Support Vector Machines (SVMs) can then be trained from such annotated corpora to detect the sentiment of text. In Twitter sentiment analysis where annotated data are impractical to obtain, noisy labels such as emoticons (“:-),” “:D,” “:(,” etc.) appeared in tweets are used to label tweets as positive or negative (Go et al. 2009).

Pioneering work on document-level sentiment classification is by Pang et al. (2002), who employed machine-learning techniques including SVMs, NB, and MaxEnt to determine whether the sentiment expressed in a movie review was *thumbs up* or *thumbs down*. They achieved the best classification accuracy with SVMs using binary features coding whether a unigram was present or not. In subsequent work, Pang and Lee (2004) further improved sentiment classification accuracy on the movie review dataset using a cascaded approach. Instead of training a classifier on the original feature space, they first filtered out the objective sentences from the dataset using a global min-cut inference algorithm and then used the remaining subjective sentences as input for sentiment classifier training. The classification improvement achieved by the cascaded approach suggests that the subjective sentences contain features which are more discriminative and informative than the full dataset for sentiment classification. The movie review dataset (also known as the polarity dataset, http://www.cs.cornell.edu/people/pabo/movie-review-data/) used in Pang et al. (2002) and Pang and Lee (2004) has later on become a benchmark for many sentiment classification studies (Whitelaw et al. 2005; Matsumoto et al. 2005). Whitelaw et al. (2005) used fine-grained semantic distinctions in features for sentiment classification, namely, the appraisal groups. Specifically, an *appraisal group* is defined as coherent groups of words that express together a particular attitude, such as *extremely boring* and *not terribly funny*. By training a SVM classifier on the combination of different types of appraisal group features and bag-of-word features, they achieved the best accuracy of 90.2 % on the movie review dataset. Matsumoto et al. (2005) proposed a method using the extracted word subsequences and dependency sub-trees as features for SVMs training and attained the state-of-the-art accuracy of 93.7 %.

A common assumption made by the aforementioned line of work (Pang et al. 2002; Pang and Lee 2004; Whitelaw et al. 2005) is that the entire document is represented as a flat feature vector (i.e., a bag-of-words format), which limits their ability to exploit sentiment or subjectivity information at a finer-grained level. In this regard, there has been work on incorporating sentence or sub-sentence level sentiment label information for document-level sentiment classification (McDonald et al. 2007; Zaidan et al. 2007).

McDonald et al. (2007) proposed a fully supervised structured model for joint sentence- and document-level sentiment classification based on sequence classification techniques using

constrained Viterbi inference. The joint model leverages both document-level and sentence-level label information and allows classification decisions from one level (e.g., the document level) to influence decisions at another level (e.g., the sentence level). It was reported that the joint model significantly outperformed both the document and sentence classifier that predict a single-level label only. Zaidan et al. (2007) used human annotators to mark the sub-sentence level text spans known as *annotator rationales*, which support the document's sentiment label. These annotator rationales were used as additional constraints for SVMs training, which ensure that the resulting classifier will be less confident in classifying the documents that do not contain the rationales. By exploiting the rationales during the classifier training, the proposed approach achieved 92.2 % accuracy on the movie review dataset and significantly outperformed the baseline SVM which only used the full text of the original documents for training.

Apart from exploiting structured information for sentiment classification, there are also works on exploring various features such as unigrams, bigrams, and part of speech (POS) tags for building sentiment classifiers. Agarwal et al. (2011) studied using the feature-based model and the tree kernel-based model for sentiment classification. They explored a total of 50 different feature types and showed that both the feature-based and tree kernel-based models perform similarly and they outperform the unigram baseline. Kouloumpis et al. (2011) compared various features including n-gram features, lexicon features based on the existence of polarity words from the MPQA subjectivity lexicon (http://www.cs.pitt.edu/mpqa/), POS features, and microblogging features capturing the presence of emoticons, abbreviations, and intensifiers. They found that microblogging features are most useful in sentiment classification.

Example: Sentiment Classification Based on Supervised Learning

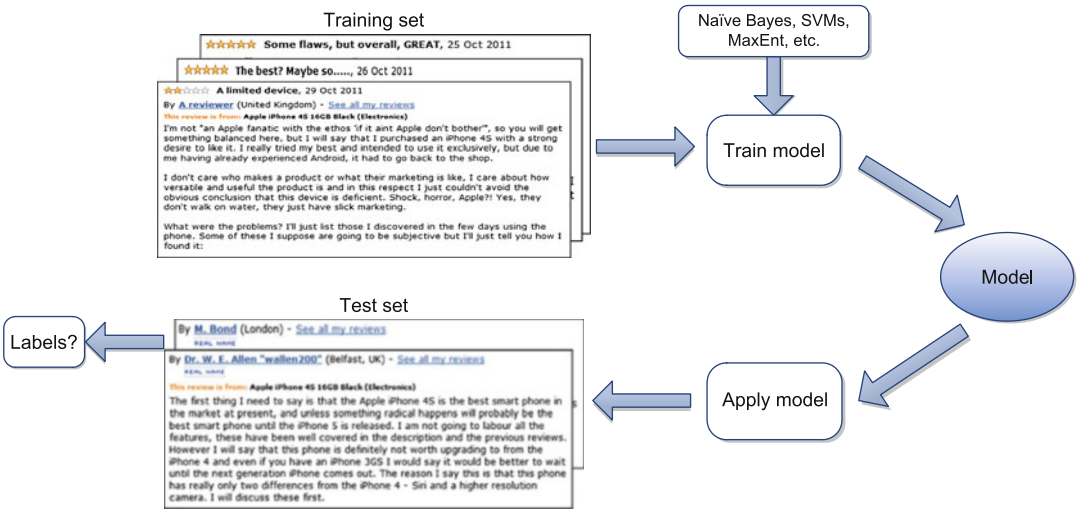
In this section, we illustrate an example of how to train a multi-variate Bernoulli naive Bayes classifier for document-level sentiment classification, i.e., to determine the sentiment orientation of a document as positive or negative. The procedures of classifier training involve three steps as depicted in Fig. 2.

Step 1 Prepare a training set: Given a set of opinionated documents $\mathcal{D} = \{d_1, d_2, \ldots, d_D\}$, each document $d \in \mathcal{D}$ needs to be annotated with a sentiment label $c \in \mathcal{C}$ as positive or negative prior to classifier training. Thus, training examples can be represented as pairs of documents and the corresponding sentiment labels as $\{\mathcal{D}, \mathcal{C}\} = \{(d_1, c_1), \ldots, (d_D, c_D)\}$. Also, using V to denote the number of distinct terms in the training set, each document can then be represented as a V-dimensional binary vector with each dimension t corresponding to term w_t. By employing the multi-variate Bernoulli naive Bayes model which only encodes the presence of words, the feature presence indicator λ_{it} (i.e., the tth dimension of document d_i) can only take two possible values, i.e., 0 indicating w_t does not appear in d_i or 1 indicating w_t has occurred in d_i at least once.

Step 2 Train a sentiment model: Given a training set $\{\mathcal{D}, \mathcal{C}\}$, the goal of model training is to calculate the optimal parameter estimates of a naive Bayes model $\mathcal{M}$. Specifically, for each term w_t in the vocabulary and each class label c_j, we need to calculate $P(w_t|c_j)$, i.e., the probability of generating w_t given class label c_j. Using the independence assumptions of NB that all attributes of data examples are independent of each other given a class label (Lewis 1998), $P(w_t|c_j)$ can be approximated from training data as

$$P(w_t|c_j) = \frac{\#\text{documents with label } c_j \text{ that contain } w_t}{\#\text{documents with label } c_j}. \quad (1)$$

We also need to compute the sentiment class probability, $P(c_j)$, which can be estimated as



Sentiment Analysis in Social Media, Fig. 2 Illustration of corpus-based approaches

the proportion of documents labeled as class c_j in the training data:

$$P(w_t|c_j) = \frac{\#\text{documents with label } c_j}{\#\text{documents in the training data}}. \quad (2)$$

Step 3 Predict sentiment label for unseen documents: Given a set of unseen documents $\mathcal{D}_u$, the final step is to predict the most probable sentiment class label $\tilde{c}$ for each unseen document $d_u \in \mathcal{D}_u$. By applying the previously trained model $\mathcal{M}$, the posterior $p(c_j|d_u)$, i.e., the probability that unseen document d_u belongs to class c_j, can be calculated as

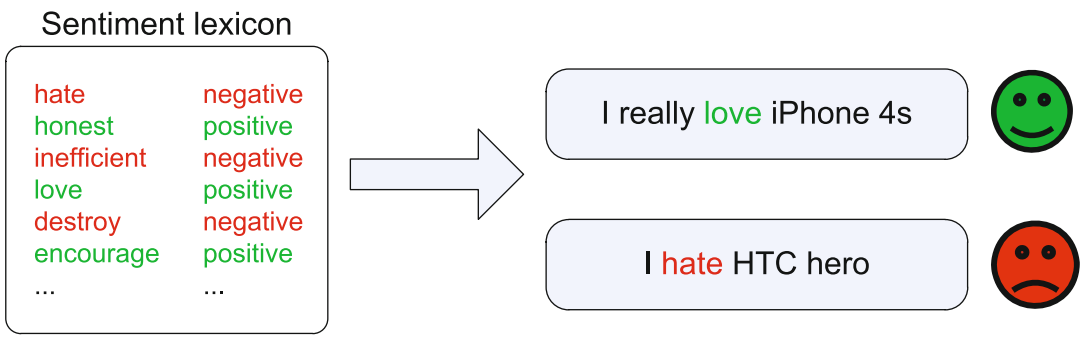
$$P(c_j|d_u) = \frac{P(c_j)P(d_u|c_j)}{P(d_u)} = \frac{P(c_j)\prod_{t=1}^{V} P(w_t|c_j)}{P(d_u)}, \quad (3)$$

where $P(d_u)$ is a normalization constant which plays no role in classification, $P(c_j)$ and $P(w_t|c_j)$ are the probabilities estimated from training data in Step 2. Finally, the class label of d_u is determined as $\hat{c}_j = \text{argmax}_{c_j} P(c_j|d_u)$.

Lexicon-Based Approaches

Lexicon-based approaches for sentiment classification are mostly unsupervised or weakly supervised. As unsupervised classifiers are usually not able to identify which features are relevant to polarity classification in the absence of annotated data, they normally resort to sentiment seed words or lexicons as a form of prior polarity knowledge for model learning as illustrated in Fig. 3. Such domain-independent sentiment lexicons can be acquired automatically or semiautomatically with much less effort compared to labeling a large training dataset.

The pioneering work in this line is that of Turney and Littman (2002), which classified a document as positive or negative by the average sentiment orientation of the phrases containing adjectives or adverbs in the document. The sentiment orientation of a phrase is calculated as the pointwise mutual information (PMI) with a positive word *excellent* minus the PMI with a negative word *poor*. The proposed approach achieved an accuracy of 84 % for automobile reviews and 66 % for movie reviews. In the same vein, Read and Carroll (2009) measured the similarity between words and polarity prototypes such as *excellent* and *good* with three different methods, namely, lexical association (using PMI),



Sentiment Analysis in Social Media, Fig. 3 Illustration of lexicon-based approaches

semantic spaces, and distributional similarity. While Turney and Littman (2002) only used one polarity prototype for each sentiment class, Read and Carroll experimented with seven polarity prototypes obtained from Roget's Thesaurus and WordNet (http://wordnet.princeton.edu/) through a selection process based on their frequency in the Gigaword corpus. The best result was achieved using PMI with 69.1 % accuracy obtained on the movie review data.

While a fixed number of sentiment seed words have been used in the aforementioned work (Turney and Littman 2002; Read and Carroll 2009), there have been attempts to incrementally enlarge the unlabeled examples with self-training based on the original seed word input (Zagibalov and Carroll 2008a, b). Starting with a single Chinese sentiment seed word meaning *good*, Zagibalov and Carroll (2008b) used iterative retraining to gradually enlarge the seed vocabulary. Those enlarged sentiment-bearing words are selected based on their relative frequency in both the positive and negative parts of the current training data. The sentiment orientation of a document is then determined by the sum of the sentiment scores of all the sentiment-bearing lexical items found in the document. Problems with this approach are that there is no principled mechanism for determining the optimal iteration number for training as well as for selecting the initial seed word, where inappropriate seed word selection may result in very poor accuracy. As such, in subsequent work, Zagibalov and Carroll (2008a) introduced a way for automatic seed word selection based on some heuristic knowledge, and an iteration control method was proposed so that iterative training stops when there is no change to the classification of any document over the previous two iterations.

Weakly supervised sentiment classification approaches are mostly lexicon based, some of which integrate with corpus-based methods as a hybrid model (Qiu et al. 2009). Compared to the seed words used in unsupervised methods, the sentiment lexicon, consisting of a list of positive and negative sentiment bearing words, is usually much larger in size and is used as reference features for sentiment classification. Analogous to the unsupervised approach that uses iterative retraining (Zagibalov and Carroll 2008b), Qiu et al. (2009) also used a lexicon-based iterative process to iteratively enlarge an initial sentiment dictionary from the first phrase. But instead of using a single seed word as Zagibalov and Carroll (2008b), they started with a much larger Chinese sentiment dictionary *HowNet* (http://www.keenage.com/download/sentiment.rar) as the initial lexicon. Documents classified from the first phase were taken as a training set to train SVMs, which were subsequently used to revise the results produced from the first phase. This self-supervised approach was tested on reviews from ten different domains and outperformed the best results of the approach by Zagibalov and Carroll (2008a) on the same data over 6 % in F-measure. In the weakly supervised joint sentiment – topic (JST) model (Lin and He 2009) can detect sentiment and topic simultaneously from text by

incorporating a small set of domain-independent sentiment lexicon (http://www.cs.pitt.edu/mpqa/databaserelease/). Unlike supervised approaches to sentiment classification which often fail to produce satisfactory performance when applied to other domains, the weakly supervised nature of JST makes it highly portable to other domains, and it is able to achieve comparable performance to the semi-supervised approaches using 40 % labeled data for training.

Example: Sentiment Classification Based on Unsupervised Learning

In this section, we show how to perform sentiment classification using pointwise mutual information (PMI) (Turney and Littman 2002) as this is one of the pioneering work of lexicon-based approach for sentiment classification. The PMI algorithm can be boiled down into three steps.

Step 1 Extract phrases containing adjectives or adverbs: The first step of the PMI algorithm is to extract two-word phrases from the document where one member of the phrase is an adjective or an adverb and the second provides context. The rationale behind is that although adjectives are generally considered good indicators for subjectivity detection from text, using an isolated adjective alone may be insufficient to determine sentiment orientation as sentiment is context dependent. For instance, the adjective "complicated" may have negative sentiment orientation as "complicated setting" in an electronic product review and conveys positive sentiment as "complicated plot" in a movie review. This phrase extraction process consists of two steps by firstly applying POS tagger to documents and then discarding the phrases with POS tags that do not conform to some predefined syntactic patterns. Readers may refer to the original paper (Turney and Littman 2002) for a full list of POS tag patterns.

Step 2 Estimate phrase sentiment orientation: In order to calculate the sentiment orientation (SO) of each extracted phrases, two sentiment polarity reference words are used, with word "excellent" indicating positive sentiment and "poor" indicating negative sentiment. So the SO of a phrase is measured by the difference of its PMI with positive word "excellent" and negative word "poor" as follows:

$$\mathrm{SO}(\text{phrase}) = \mathrm{PMI}(\text{phrase}, \text{“excellent”}) - \mathrm{PMI}(\text{phrase}, \text{“poor”}). \quad (4)$$

Formally, the PMI of words w_1 and w_2 is given by

$$\mathrm{PMI}(w_1, w_2) = \log_2\left(\frac{p(w_1 \wedge w_2)}{p(w_1)p(w_2)}\right), \quad (5)$$

where $p(w_1 \wedge w_2)$ is the joint probability of how likely that word w_1 and w_2 co-occur. If w_1 and w_2 are statistically independent, this joint probability is equivalent to $p(w_1)p(w_2)$. Thus, the ratio between $p(w_1 \wedge w_2)$ and $p(w_1)p(w_2)$ essentially measures the degree of statistical dependence between the words. In practice, the probabilities required for calculating PMI can be acquired by issuing quires to a public search engine (http://www.altavista.com/sites/search/adv), and then based on the results returned, we can approximate $p(w_1)$ with the number of hits that documents contain w_1 and approximate $p(w_1 \wedge w_2)$ with the number of hits that documents contain both w_1 and w_2 within a range of ten words. Thus, (4) can be rewritten as

$$\mathrm{SO}(\text{phrase}) = \log_2\left(\frac{\text{hits(phrase NEAR “excellent”) hits(“poor”)}}{\text{hits(phrase NEAR “poor”) hits(“excellent”)}}\right). \quad (6)$$

Step 3 Calculate document sentiment: The final step is to calculate the average SO of all extracted phrases in the document and then classify the document as positive if the average SO is positive and as negative otherwise.

Explore Social Networks for Sentiment Analysis

Recently, there have been increasing interests in employing social relations for both document-level and user-level sentiment analysis. It is based on a hypothesis that users connected with each other are likely to express similar opinions. In Twitter, social relations can be established by the following links, through retweeting using RT username or via username, or by referring to other users in one's messages using "@" mentions.

Speriosu et al. (2011) argued that using noisy sentiment labels may hinder the performance of sentiment classifiers. They proposed exploiting the Twitter follower graph to improve sentiment classification and constructed a graph that has users, tweets, word unigrams, word bigrams, hashtags, and emoticons as its nodes which are connected based on the link existence among them (e.g., users are connected to tweets they created; tweets are connected to word unigrams that they contain). They then applied a label propagation method where sentiment labels were propagated from a small set of nodes seeded with some initial label information throughout the graph. They claimed that their label propagation method outperforms MaxEnt trained from noisy labels and obtained an accuracy of 84.7 % on the subset of the Twitter sentiment test set from Go et al. (2009).

Tan et al. (2011) incorporated both textual and social relations revealed by the following links and "@" mentions in a single heterogeneous graph on a certain topic such as "Obama," where nodes correspond to either users or tweets. Starting from some seed-user nodes labelled as positive or negative, they proposed a transductive learning method to propagate sentiment label to all the users in the graph. In a similar vein, Calais Guerra et al. (2011) also proposed modeling the user opinion prediction problem as a relational learning problem over a network of users connected by endorsement (e.g., retweets in Twitter) where the goal is to classify the nodes of a partially labelled network.

Key Applications

Social media such as Twitter and Facebook has become an increasingly popular communication channel, which have enabled many useful applications by discovering opinions reflecting people's attitudes towards various topics or events from the massive user-generated data. These social media centric applications are particularly proliferous in the domains of financial marketing, brand and consumer perception, as well as anti-terrorism and violence detection.

Financial Marketing

Sentiment analysis has shown great impact on finical markets, where financial organizations are embracing new tools and techniques to help make sense of the massive amounts of unstructured data available on social media for making more informed decisions and maximizing the performance of their trading strategies. For instance, Thomson Reuters (http://thomsonreuters.com/) recently launched a sentiment analytics service for Internet news and social media, which is capable to mine expansive wealth of social media and blog content to deliver digestible analytics for algorithmic trading systems as well as risk management and human decision support processes. Social Market Analytics (SMA) (http://socialmarketanalytics.com/) tracks live stock market sentiment and offers to detect abnormal positive or negative changes in investor sentiment as it is expressed in real-time social media activity. HedgeChatter (http://www.hedgechatter.com/) also uses social media sentiment for stock market analysis, with a focus of identifying the most influential users in social media based upon their overall volume of postings, followers they have, and how accurately they predict stock price.

Brand and Consumer Perception

Engaging with consumers and gaining perceptions of brands is another active domain of applying social media analytics, where commercial products preserve similar visions such as to support brands to better understand costumer segments, what consumers value about the brands,

and how consumers perceive their products and services and those of their competitors.

IBM has developed an internally used system called Banter (http://www.research.ibm.com/social/) for monitoring and analyzing the contents of blogs and social network conversation. It answers key questions for marketers such as *How do I identify relevant blogs? Who are the key influencers?* and *What is the sentiment about these relevant topics?*. In terms of commercial products, one leading company is Bazaarvoice (http://www.bazaarvoice.co.uk/) which provides a comprehensive social media analytics platform covering a range of services, such as gathering consumer generated opinions from customer conversations on social networks as well as capturing and responding to consumer questions about products and services. Another major player in this market is PowerReviews (http://www.powerreviews.com/). In contrast to Bazaarvoice which targets big enterprises PowerReviews is more focused on small- and medium-sized business (SMB) solutions.

While the aforementioned products provide services for across industrial clients, some companies optimize their products for a dedicated domain. For instance, Musicmetric (http://www.musicmetric.com/) collects and analyzes online data globally to understand activity around artists for the entertainment industry. In addition to activity data from social media sites and peer-to-peer file sharing, they also analyze reviews to determine which artists, songs, and albums are being reviewed in an article, as well as the overall associated sentiment. Using this data, Musicmetric is able to provide aggregate sentiment statistics for an artist, song, or album over all reviews analyzed online.

Anti-terrorism and Violence Detection

Another important emerging area for sentiment analysis leveraging data from social media is violence detection and anti-terrorism. Existing work on terrorism detection from online content has been largely focusing on the study of terrorists, hate groups, and other extremists through primary sources such as terrorists' own websites, videos, chat sites, and Internet forums. For example, the University of Arizona's Dark Web Terrorism Research Programme (http://ai.arizona.edu/research/terror/) employed various data mining techniques to conduct content analysis and social network analysis of online jihadist content. A Dark Web Forum Portal (DWFP) has been developed to provide Web-enabled access to 29 important jihadist and other extremist Web forums and currently archives approximately 15 million messages. Recently, DARPA unveiled the Social Media in Strategic Communication (SMISC) (http://www.darpa.mil/) program with the goal to detect and conduct propaganda campaigns on social media. In order to help defense department to gain deep understanding of social media dynamics, particularly in the areas where the troops are deployed, SMICS can perform real-time discovering and tracking of the development and spread of ideas and concepts on social media, as well as to quickly flag rumors and emerging themes that might be considered risky.

Future Directions

This chapter gave an introduction to sentiment analysis in social media. Despite the recent successes, the field of sentiment analysis is still relatively new and many challenges remained to be tackled:

1. Topic-dependent sentiment analysis. Sentiment is domain dependent, where sentiment expressions in different domains can be quite different. Besides, even for data from the same domain, sentiment distributions may vary over time, especially for collections that span years or decades and the fast-evolving social media data such as Twitter data. Therefore, topic-sensitive sentiment analysis and detecting and tracking the dynamics in both topic and sentiment over time in time-variant datasets are promising areas for research.
2. Multilingual sentiment analysis. Most of the sentiment analysis systems are monolingual which typically process English only. However, a sentiment system with multilingual

capability is important as users such as international companies often need to gain insights into markets for more than one country, e.g., USA and China.

3. In Twitter and other social media sites such as Facebook and YouTube, short, ungrammatical utterances are commonplace. Finding an effective way of correcting these spelling mistakes is important for improving sentiment analysis system performance.
4. Sarcasm and slang. Sentiment is often embodied in subtle linguistic mechanisms such as the use of sarcasm and slang, which poses great challenges for automated sentiment analysis. For instance, without taking context into account, sarcasms expressing negative sentiment could be wrongly interpreted as extremely positive sentiment. On the other hand, understanding slang is also very difficult as it changes by geographical location. Therefore, addressing this challenge would require deeper linguistic understanding and incorporating richer background knowledge for model learning.

Cross-References

- ▸ Data Mining
- ▸ Multi-Classifier System for Sentiment Analysis and Opinion Mining
- ▸ Twitter Microblog Sentiment Analysis
- ▸ User Sentiment and Opinion Analysis

References

Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau R (2011) Sentiment analysis of twitter data. In: Proceedings of the ACL 2011 workshop on languages in social media, Portland, pp 30–38

Aue A, Gamon M (2005) Customizing sentiment classifiers to new domains: a case study. In: Proceedings of recent advances in natural language processing (RANLP), Borovets

Barbosa L, Feng J (2010) Robust sentiment detection on twitter from biased and noisy data. In: Proceedings of COLING, Beijing, pp 36–44

Blitzer J, Dredze M, Pereira F (2007) Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification. In: Proceedings of the Association for Computational Linguistics (ACL), Prague, pp 440–447

Boiy E, Hens P, Deschacht K, Moens M (2007) Automatic sentiment analysis in on-line text. In: Proceedings of the 11th international conference on electronic publishing, Vienna pp 349–360

Calais Guerra P, Veloso A, Meira W Jr, Almeida V (2011) From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, San Diego, pp 150–158

Go A, Bhayani R, Huang L (2009) Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford

Kouloumpis E, Wilson T, Moore J (2011) Twitter sentiment analysis: the good the bad and the omg! In: Proceedings of the ICWSM, Barcelona

Lewis D (1998) Naive (bayes) at forty: the independence assumption in information retrieval. In: Machine learning: ECML-98, Chemnitz. Springer, pp 4–15

Lin C, He Y (2009) Joint sentiment/topic model for sentiment analysis. In: The 18th ACM conference on information and knowledge management (CIKM), Hong Kong

Matsumoto S, Takamura H, Okumura M (2005) Sentiment classification using word sub-sequences and dependency sub-trees. In: Proceedings of the Pacific-Asia conference on knowledge discovery and data mining (PAKDD), Hanoi. Springer, pp 301–310

McDonald R, Hannan K, Neylon T, Wells M, Reynar J (2007) Structured models for fine-to-coarse sentiment analysis. In: Proceedings of the annual meeting of the Association of Computational Linguistics (ACL), Prague, pp 432–439

Pak A, Paroubek P (2010) Twitter as a corpus for sentiment analysis and opinion mining. In: Proceedings of LREC, Malta

Pang B, Lee L (2004) A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd annual meeting on Association for Computational Linguistics, Barcelona. Association for Computational Linguistics, p 271

Pang B, Lee L, Vaithyanathan S (2002) Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the conference on empirical methods in natural language processing, Philadelphia, vol 10. Association for Computational Linguistics, pp 79–86

Qiu L, Zhang W, Hu C, Zhao K (2009) SELC: a self-supervised model for sentiment classification. In: Proceeding of the 18th ACM conference on information and knowledge management (CIKM), Hong Kong, pp 929–936

Read J, Carroll J (2009) Weakly supervised techniques for domain-independent sentiment classification. In: Proceeding of the 1st international CIKM workshop on topic-sentiment analysis for mass opinion, Hong Kong, pp 45–52

Speriosu M, Sudan N, Upadhyay S, Baldridge J (2011) Twitter polarity classification with label propagation over lexical links and the follower graph. In: Proceedings of the EMNLP 2011, conference on empirical methods in natural language processing, Edinburgh, pp 53–63

Tan C, Lee L, Tang J, Jiang L, Zhou M, Li P (2011) User-level sentiment analysis incorporating social networks. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, San Diego

Turney P (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL'02), Philadelphia

Turney PD, Littman ML (2002) Unsupervised learning of semantic orientation from a hundred-billion-word corpus. ArXiv Computer Science e-prints cs.LG/0212012

Whitelaw C, Garg N, Argamon S (2005) Using appraisal groups for sentiment analysis. In: Proceedings of the ACM international conference on information and knowledge management (CIKM), Bremen, pp 625–631. doi:http://doi.acm.org/10.1145/1099554.1099714

Zagibalov T, Carroll J (2008a) Automatic seed word selection for unsupervised sentiment classification of Chinese text. In: Proceedings of the 22nd international conference on computational linguistics (COLING), Manchester, pp 1073–1080

Zagibalov T, Carroll J (2008b) Unsupervised classification of sentiment and objectivity in Chinese text. In: Proceedings of the third international joint conference on natural language processing, Hyderabad, pp 304–311

Zaidan O, Eisner J, Piatko C (2007) Using annotator rationales to improve machine learning for text categorization. In: Proceedings of NAACL-HLT, Rochester, pp 260–267

Sentiment Classification

▸ Sentiment Analysis in Social Media

Sentiment Detection and Analysis

▸ Opinion Diffusion and Analysis on Social Networks

Sentiment-Emotion-Intent Analysis

▸ Twitris: A System for Collective Social Intelligence

Server-Side Scripting Languages

Ludger Martin
Institute of Computer Science, Johannes Gutenberg-University Mainz, Mainz, Germany

Synonyms

ASP; ASP.NET; CGI; JSF; JSP; Perl; PHP; Ruby on Rails

Glossary

AJAX Asynchronous JavaScript and XML
CGI Common Gateway Interface
HTML Hypertext Markup Language
JSF Java Server Faces
JSON Java Script Object Notation
JSP Java Server Pages

Definition

Server-side scripting languages are programming languages developed especially for creating HTML pages (or Web pages) on the server side. These languages usually provide special libraries that facilitate creating HTML pages. In times of Web 2.0 and AJAX, these scripting languages can also serve as data sources (services) for AJAX.

There are two different types of scripting languages. The first variant can be embedded in HTML. The language can be embedded, for example, in places where a particular functionality is needed. The second variant is languages which can be used to create HTML tags. They provide an interface for creating HTML tags.

On the server side, a special interpreter is necessary for each scripting language. This interpreter is introduced to the Web server so that the server will be able to use it for the script execution, when required.

Introduction

There is a large number of server-side scripting languages. It is their task to dynamically build HTML pages (Web pages) on the server side. To achieve this, a Web server that is to distribute the HTML pages must be told where to find an interpreter for a particular script. Most of the server side scripts are interpreted. A small number can also be compiled.

Without server-side scripting languages, you can only create static Web pages. Then it is not possible to customize anything for single users. A customization can be something as simple as the display of search results. Because of these languages, HTML pages can be created dynamically, i.e., on request.

If we look at today's Web sites, we will find that most of them were created using servers-side scripting languages, among them Web sites which are run by a content management system. The content management system itself has been developed using a server side scripting language. There will be very few exceptions which do not use such a language.

Web 2.0 pages that use JavaScript for controlling their content also need a server side scripting language, e.g., AJAX was used to send requests for database access to a server. This can only be done using a server side scripting language.

Historical Background

It is hard to say which server side scripting language was first. It is a fact though that Perl was one of the first languages. The first version of Perl as a universal scripting language was presented in 1987 by Larry Wall. Only later, in the 1990s, did it become useful for Web pages because CGI was introduced.

In 1995, Rasmus Lerdorf developed PHP. At the beginning, PHP was based on Perl. In 1997, with version 2, the first parser for PHP was delivered. From then on PHP has been particularly suited for Web pages. PHP is a scripting language embedded in HTML. From the very start the evaluation of form variables has been important. By now, PHP has become one of the most widely used scripting languages for Web pages. But PHP has also become a universal scripting language which can be used anywhere.

Pyton is another universal scripting language, which was developed by Guido van Rossum in 1991. Today it is also commonly used for Web applications.

At the end of the 1990s, Sun Microsystems presented the language JavaServer Pages (JSP). JSP is based on the language Java, but it is embedded in HTML. Just as with Java, the JSP pages must be compiled before the byte code that was created can be executed in a virtual machine. Nowadays, JSP is considered outdated. It was superseded by JavaServer Faces (JSF) in 2004. Particularly for Web pages, JSF, as opposed to JSP, is component oriented. JSF consequently focuses on the model–view–controller pattern.

Microsoft developed the Active Server Pages (ASP) particularly for the Internet Information Service (IIS). The technology, which was presented in 1998, can be programmed among others with VBScript or JScript. The relevant programming language is also embedded in HTML. In 2002 it was superseded by ASP.NET. That is the Web-based technology which is embedded in the .NET framework. Programming languages for ASP.NET are VBScript and C#.

Another popular language for Web pages is Ruby on Rails. The programming language Ruby was presented by Yukihiro Matsumoto in 1995. At the beginning it was only known in Japan. Ruby on Rails, which was developed in 2005, is a specific library for Web applications.

Server-Side Scripting Languages

In the following, a sample server side scripting language will be described. We chose PHP because it is one of the most widely used languages.

After that, we will take a glance at Perl, which is unlike PHP a language that is not embedded in HTML.

PHP

From version 2 on, PHP has been developed for the dynamic creation and evaluation of Web pages. At first it was a procedural programming language. Version 4.0 (2000) introduced objects, which were revised in version 5.0 (2004). PHP provides a comprehensive procedural and object-oriented library.

PHP (Lerdorf et al. 2006) is a programming language that is embedded in HTML (Kessin 2011). It is always interpreted on the server. Output of a PHP script is usually an HTML page. But it is also possible to create different text formats and binary formats such as JSON, PDF, or PNG. Because the scripts are executed on the server, the user cannot see the source code. Users only get to see the output. This way, stealing the source code is not possible. If a Web server is very busy, the PHP scripts can be compiled beforehand and then only the byte-code can be executed. Without parsing and compiling, execution performance is strongly improved.

Figure 1 shows a very small PHP-file whose browser output is *hello world*. It shows clearly that the file begins with HTML source code. PHP is embedded in HTML; the actual PHP part starts only in line 8. This is marked by the string `<?php`. The command `echo` makes the browser display `<p>Hello world!</p>`. HTML tags can also be output using `echo`. The PHP part ends in line 10 with the string `?>`. You can include PHP parts anywhere and any number of times. It is not necessary to include HTML source code in a PHP-file, which is often the case with classes. In this scenario, the PHP-file starts directly with `<?php`. If the file ends with a PHP source text, you can leave out the closing `?>`. Formerly, `<?` and `?>` were used, but they had caused problems with XHTML. A PHP-filename must always finish with `.php` for the Web server to know that it is a PHP-file.

You can also include comments in the PHP sections. Introduce single-line comments using `//` or `#`. Multiline comments should be enclosed in `/*` and `*/`.

```
1  <DOCTYPE html>
2  <html>
3     <head>
4        <title>Hello World</title>
5        <meta charset="UTF-8"/>
6     <head>
7     <body>
8        <?php
9           echo "<p>Hello world!</p>";
10       ?>
11    </body>
12 </html>
```

Server-Side Scripting Languages, Fig. 1 *Hello World* PHP page

Variables and Operators

Variables are a central feature in a programming language. PHP is an untyped programming language. This means that usually you do not need to specify types. There are two exceptions, which will be explained later. As a result, you do not need to define variables, you can simply use them.

Variables always begin with `$`. Then, an arbitrary sequence of characters and numbers may follow; the first character after `$` must be a letter. PHP distinguishes between capital letters and small letters. As it is not necessary to define the variables, there is a certain danger. If you access an undefined variable, in the best case you will receive a warning. In the worst case you will only notice that the program does not work as expected.

You can assign a type to a variable by giving it a value. Figure 2 shows several examples of this. In lines 1–4, numbers are assigned. The lines 5 and 6 treat booleans, where `TRUE` is the same as 1 and `FALSE` equals an empty string. Lines 7 and 8 demonstrate how character strings are assigned. It is important to realize that there is a great difference between the opening and closing `"` and `'`.

Server-Side Scripting Languages, Fig. 2 Variables

```
1  $x = 42;
2  $x = 0xFF;
3  $x = 4.2;
4  $x = 4.2e6;
5  $x = TRUE; // or 1
6  $x = FALSE; // or ''
7  $x = "abc";
8  $x = 'abc';
9  $x = 8;
10 $y = "it's $x o'clock"; // value: it's 8 o'clock
11 $a = array('one', 'two', 'three');
12 $a = array(1 => 'one', 2 => 'two', 3 => 'three');
```

```
1  int isset (mixed var);
2  int unset (mixed var);
3  $var = NULL;
4  string gettype (mixed var);
```

Server-Side Scripting Languages, Fig. 3 Checking of variables

Only if " is specified, variables and escape sequences (e.g., \t for tabulator) in character strings are resolved. That means that as the lines 9 and 10 show $x is replaced by the numeric value 17. If ' is used, $x will remain a character string.

You can also create arrays. Line 11 shows an array with three elements. Line 12, however, shows an associative array. Using => you can separate the keys from the values. Because PHP is an untyped language, the values and the keys of the types may vary within an array.

PHP changes the variable type according to the situation, if necessary. If, for example, two variables are added as numbers and one of them is a character string, then this string will automatically be changed to a number. Sometimes, not often, an explicit type conversion may be necessary. This can be done placing the required type in round brackets in front of the variable: (int)$x.

You can check variables using the functions in Fig. 3. In the following, functions and methods will always be specified including the expected and returned types. Because PHP is an untyped language, the types will only be checked at runtime and only afterwards an error message will be displayed if the types do not match. Specification of mixed means that different types may be used. isset() checks if a variable has been specified. With lines 2 and 3 you can specify a variable as undefined. gettype() determines the current type of a variable value. This value may change during the execution of a program. The type itself will be returned as a string.

There are very few specifics for operators. As we have already seen, = is an assignment. +, -, *, /, and % are mathematical operators, whereby division and multiplication come before addition and subtraction. To concatenate character strings, use . . For comparisons, ==, !=, <>, <, >, <=, and >= are available. Variable types can be compared using === und !==. For grouping operators you can use round brackets. You can use AND, &&, OR, ||, XOR and ! as logical operators, e.g. for conditions. AND and && are equivalent and so are OR and ||.

```
1 if ($i > 0) {
2     echo '$i is grater than 0';
3 } elseif ($i = 0) {
4     echo '$i is equal to 0';
5 } else {
6     echo '$i is less than 0';
7 }
```

Server-Side Scripting Languages, Fig. 4 `if` condition

Program Control

PHP offers the usual options for program control. Figure 4 shows an `if` condition. After the keyword `if`, you must specify a condition in round brackets. Unlike in other languages, in addition to the `else` branch, there are one or more alternative conditions `elseif`. Besides the `if` condition there is additionally a `switch` statement.

Loops are also an option. A `do { ... } while (...)`–loop checks the condition after every loop run and the `while (...){...}`–loop checks the condition prior to every loop run.

The `for` loop corresponds to the C syntax in that you can specify separately first an initialization statement, followed by a condition and then an incrementation using `;`. Additionally, there is a `foreach` loop which has been developed specially for arrays. This loop executes the following statements once for each value in the array. Figure 5 gives an example. Inside the brackets the array is specified first, then follows the keyword `as`. After that, one or two variables are specified, which will be used to store the value and, optionally, the key of the array. The variable for the key and the characters `=>` can be left out if only the values are of interest.

Classes, Objects, Error Handling

PHP too allows you to define custom functions and objects. With version 5, object–orientation has been thoroughly revised.

Using the keyword `class` you can define classes. Classes' attributes and methods can be defined as *public*, *protected*, and *private* to ensure access protection. PHP supports only single class inheritance, which can be specified using the keyword `extends`. Alternatively, there are interfaces, which a new class can implement. Classes can also be abstract. This is the case as soon as at least one method of a class has been marked as abstract. These methods are not yet implemented. The class which inherits must, similar to the interfaces, implement the abstract methods.

Constructors must be named `__construct`. For downward compatibility with PHP 4, the constructor may have the same name as the class. A destructor must be named `__destruct`.

As Fig. 6 shows, you can create instances of classes using the keyword `new`. An object of the class `DateTime` is instantiated. You can access the methods or attributes of objects using the `->` operator, as demonstrated by the method `format()`.

PHP also allows creating static attributes and methods. Polymorphic methods are not supported, as PHP is an untyped language. It is only possible to specify default values for single parameters, so that the values can be omitted.

Error handling is also possible. Figure 7 shows how a `try ... catch ...` block is specified. After `try` you specify all the statements which might cause errors. One or more `catch` statements control the error handling. Only in the `catch` statement, it is necessary to specify a type within the brackets. This type determines the class of the exception that is to be treated. This enables you to react appropriately to different exceptions. Using the keyword `throw`, you can throw a new exception.

Interaction with HTML Forms

The interaction with HTML forms is a central point with server-side scripting languages. Here, you need to be particularly careful because these places are popular goals for attacks on a Web page or an application. For this very reason, PHP as well has undergone a variety of improvements in the course of time.

In Fig. 8 you can see a small HTML form which consists of two input fields for the user

Server-Side Scripting Languages, Fig. 5 foreach loop

```
1 $a = array (1 => 'one', 2 => 'two', 3 => 'three');
2 foreach ($a as $key => $value)   {
3       echo "$value has key $key";
4 }
```

```
1 $aDateTime = new DateTime();
2 echo $aDateTime->format('Y-m-d H:i:s');
```

Server-Side Scripting Languages, Fig. 6 Instantiate a class

```
1 try {
2       $date = new DateTime('2012-08-01');
3 } catch (Exception $e) {
4       echo $e->getMessage();
5 }
```

Server-Side Scripting Languages, Fig. 7 Error handling

name and the password and a send button. In the form, the send method `post` has been specified. You have two options. `get` is the simplest method. Here, data is coded and committed with the URL. An advantage is that the data is visible in the URL. The disadvantages are that the URL and with it the data amount is limited in size and anyone can see and modify the committed data. Especially for passwords this method is not recommended. Here, the `post` method can help. Data is sent to the URL separately and there is no limit in the amount of data. If you want to transfer entire files with a form, the method `post` is obligatory. It is more difficult to modify this data, but not impossible.

As destination for the form, the PHP–file `form.php` is specified. The `<input>` tags each have an attribute `name`, which determines the name of the variable as it shall be available in the destination script. In the past, these names could be used in PHP directly. Today, for security reasons they are stored in the two arrays `$_GET` and `$_POST`. The name of each `<input>` tag is its array index. Line 1 in Fig. 9 shows how to check whether the button *submit* was pressed. Because the method `post` was used, in the array `$_POST` the index `'submit'` will be searched for and a check will be done to see if the array contains the value `submit`. If so, a similar procedure can check the user name and password.

You can use cookies if a Web application is to store data on a client (or browser) permanently. Cookies may contain any data, but they have a maximum length of 4 KB. But a Web application can send up to 20 cookies to the browser. If you want to define how long data shall be kept, you can specify an individual expiry date for a cookie. Cookies can only be placed in a Web page's header. This is possible only if no character of the actual Web page has been output. It is recommended to specify a cookie for a script as early as possible. As soon as a cookie was set, the browser will send it automatically to the server with every query. In PHP cookies can be read using the array `$_COOKIE`. The problem with cookies is that users can decide whether or not their browser shall accept cookies. A Web application can only find out about this decision if it checks whether cookies are transferred back.

Web pages are generally independent of a context. When developing Web applications, this can be annoying. You can solve this problem by using sessions. Sessions provide a separate storage area on the server for each user. A session is assigned to a user through a unique identifier. This ID must always be transferred between the browser and the server. To achieve this, three variants are

Server-Side Scripting Languages, Fig. 8 Small HTML-form

```
1 <form action="form.php" method="post">
2   <p>
3     username <input type="text" name="username"/><br/>
4     password <input type="password" name="password"/><br/>
5     <input type="submit" name="submit" value="submit"/>
6   </p>
7 </form>
```

```
1 if ($_POST['submit'] == 'submit') {
2     if (($_POST['username'] == ...)&&
3         ($_POST['password'] == ...)) {
4         // secure action
5     }
6 }
```

Server-Side Scripting Languages, Fig. 9 Evaluate form

available, which have been presented earlier: get, post, cookie. Cookie is the most popular variant because it causes the least work. It is, however, the most problematic as the developer does not know whether the browser accepts cookies. If cookies are used, again the cookie must be transferred in the header.

To use a session in PHP, each file which uses the session must call the command `session_start()` (see Fig. 10). This command checks the session ID if one was transferred. If the ID is correct, an existing session will continue to be used. If it is not correct or not available, a new session will be created. The array `$_SESSION` is available for storing data (s. lines 2 and 3). The data remains stored on the server for some time and the various PHP scripts can access it. This time limit is important because sometimes it may not be clear whether

```
1 session_start(void);
2 $_SESSION['text'] = 'Hello world!';
3 echo $_SESSION['text'];
```

Server-Side Scripting Languages, Fig. 10 Session

a user is still active. If a user is inactive over a longer period of time, the session will be deleted. Sessions can also be deleted by PHP.

Any kind of data can be stored in a session. You can also store objects. But there is one condition: The object's class must be known in each PHP file before the session can be started. With respect to security (Hope and Walther 2008), sessions must be particularly protected. The usual attacks are *session hijacking* and *session riding*.

Other Data Formats

Besides HTML, PHP (Loudon 2010) allows you to create any other data format. These formats could be, for example, XML, JSON, or images. JSON, for example, is often used if it is a Web 2.0 application and you want to send back queries to a server using AJAX. For the browser to be able to recognize what kind of data it is, the HTTP-header that specifies the data format must be modified. Figure 11 shows how to set the type for JSON-data. For the function `header()`, it is important that in the body no data has yet been sent. It is the same as for cookies.

JSON is a frequently used data format. Therefore, special functions are available for creating JSON automatically from PHP-objects. In line 2 a sample object is instantiated. Using the

```
1 header('Content-type: application/json');
2 $object = new ...;
3 $json = json_encode($object);
4 echo $json;
```

Server-Side Scripting Languages, Fig. 11 Modify header

function `json_decode()`, it can be changed to a JSON character string, which can then be output with `echo`. Analogously, the function `json_decode()` can change a JSON character string into a PHP-object.

Embedded Files and Debugging

If you develop classes, usually each class is stored in a separate file. Then it is also necessary that these files can be embedded in PHP scripts. To do so, in PHP two functions are available. `require()` embeds the specified file and displays an error message if this was not successful. Processing is stopped. `include()`, however, will only output a warning and continue the script. For each one there is an alternative method `require_once()` and `include_once()`. These functions ensure that a script will be embedded only once, even if it was specified more often. This is particularly useful for, e.g., classes.

Troubleshooting is also very important for server side applications. To enable debugging of an application, there are currently two modules for the Web server Apache, *Xdebug*, and *Zend Debugger*, which allow remote debugging. Using these modules in a compatible development environment, you can execute the PHP source code step by step and examine the variables as well.

Perl

Perl (Guelich et al. 1999; Wall et al. 2000) is a universal programming language that can be used for developing server side applications, too. The module `CGI` provides an interface that can be used to create HTML elements very easily. Figure 12 shows an example of the *hello world* program in Perl.

```
1 #!/usr/local/bin/perl -w
2 use CGI;
3 $q = CGI->new;
4 print $q->header,
5       $q->start_html('hello world'),
6       $q->p('hello world'),
7       $q->end_html;
```

Server-Side Scripting Languages, Fig. 12 *Hello World* Perl page

Line 2 provides the CGI module. Then, an object whose class is `CGI` is instantiated. Then some methods are used to create the individual HTML elements. The methods `header` and `start_html` create the entire header. For each HTML element in the body, there is one method for creating that particular element. You can see this in line 6 for the `<p>`–tag. Each method returns a string, which is output using `print`.

Future Directions

Despite the development of Web 2.0 and the relocation of functionality to the client side, i.e., the browser, server-side scripting languages will still remain beneficial. Using AJAX, for example, a data source must be provided on the server side. Concerning the various libraries, a trend can be seen that the server-side scripting languages are more and more used for the automatic creation of JavaScript source code. So it will remain exciting to watch how the fast-moving world of the Web will develop.

Cross-References

▸ HTML

References

Guelich S, Gundavaram S, Birznieks G (1999) CGI programming with Perl. O'Reilly Media, Sebastopol
Hope P, Walther B (2008) Web security testing cookbook: systematic techniques to find problems fast. O'Reilly Media, Sebastopol
Kessin Z (2011) Programming HTML5 applications: building powerful cross-platform environments in javascript. O'Reilly Media, Sebastopol
Lerdorf R, Tatroe K, MacIntyre P (2006) Programming PHP, 2nd edn. O'Reilly Media, Sebastopol
Loudon K (2010) Developing large web applications. O'Reilly Media, Sebastopol
Wall L, Christiansen T, Orwant J (2000) Programming Perl: There's more than one way to do it. O'Reilly Media, Sebastopol

Service Delivery Networks

▸ Inter-organizational Networks

Service Discovery

Matthias Klusch
German Research Center for Artificial Intelligence (DFKI), Saarbruecken, Germany

Synonyms

Semantic web services; Service-oriented architectures; Service search and selection; Web services

Glossary

CAN Content-Addressable Network
DHT Distributed Hash Table
JSON JavaScript Object Notation
NAICS North American Industrial Classification System
OWL-S Ontology Web Language for Services
REST Representational State Transfer
SA-WSDL Semantically Annotated WSDL
SA-REST Semantic Annotation of Web Resources
SIC Standard Industrial Classification
SML Service Modeling Language
SOA Service-Oriented Architecture
SOAP Simple Object Access Protocol
UDDI Universal Description, Discovery, Integration
UNSPSC United Nations Standard Products and Services Code
USDL Unified Service Description Language
UUID Unique Universal Identifiers
WADL Web Application Description Language
WSDL Web Service Description Language
WSML Web Service Modeling Language

Definition

Service discovery is the process of locating existing services that are relevant for a given request based on the description of their functional and non-functional semantics. Approaches to service discovery differ in their support of service description language(s), the organization of the search, and the utilized means of service selection.

Introduction

The continuous proliferation of web services which encapsulate business software and hardware assets, e-business, or social software applications in the web 2.0 holds promise to further revolutionize the way of interaction within today's society and economy. A service can be defined as a kind of action, performance, or promise that is exchanged for value between provider and client. In other words, it is a provider-client interaction that creates and captures value for all parties involved. At present, there are tens of thousands of web services for a huge variety of applications and in many heterogeneous formats available for the common user of the web. One main challenge of web service technology is to

provide scalable and effective means for an automated discovery of relevant services with minimal human intervention in any user and application context. This paper provides an overview of service discovery in a nutshell. For a more comprehensive survey on the subject, the interested reader is referred to, for example, Crasso et al. (2011) and Klusch (2008b, 2012).

Preliminaries

Service discovery can be performed in different ways depending on how the services of the considered search space are described, how the search process is organized, and which means of service selection are used for the search.

Service Description In general, a web service can be described in terms of what it does and how it actually works. These aspects of its functional semantics (aka capability) are described in a service profile and a service process model, respectively.

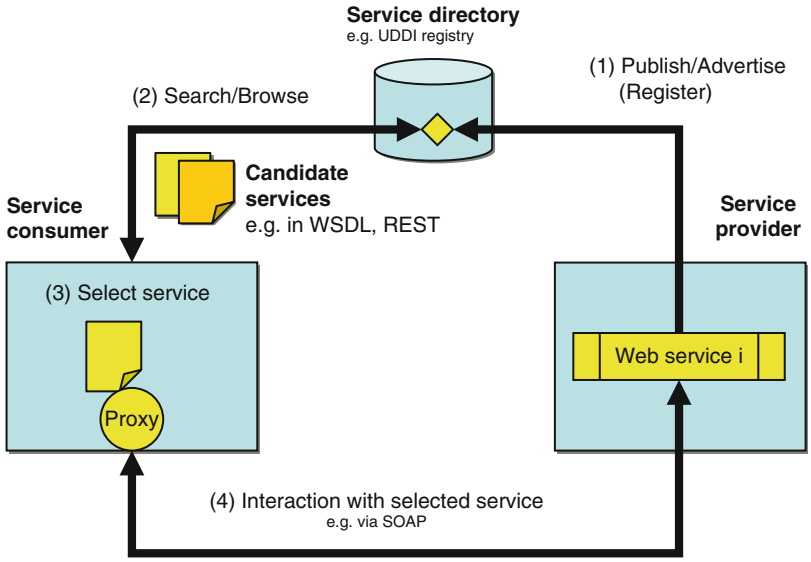
A *service profile* describes the signature of a service in terms of its input and output (I/O) parameters and the service specification, i.e., the preconditions and effects (P/E) of the service execution. The profile also describes non-functional service semantics such as information about its provenance, name, business category, pricing, delivery constraints, and quality. Prominent approaches to represent such profiles are the XML-based web service description languages WSDL (Chinnici et al. 2007), SML (Pandit et al. 2009), USDL (Oberle et al. 2013), and WADL (Hadley 2009) and the HTML microformat hREST (Kopecky et al. 2008). Other examples are the textual documentations of RESTful services (Fielding and Taylor 2002) and the ontology-based service description languages OWL-S (Martin et al. 2004), WSML (De Bruijn and Lausen 2005), SAWSDL (Farrell and Lausen 2007), SA-REST (Gomadam et al. 2010), and Linked USDL (Pedrinaci and Leidig 2011).

A *service process model* describes the operational behavior of a service in terms of its internal control and data flow. Such models are described, for example, in OWL-S, WSML, and USDL by use of standard workflow operators like sequence, split+join, and choice, while other representation approaches are adopting process algebraic languages like the pi-calculus and Petri nets for this purpose.

Discovery Architectures Approaches to organize the service search can be classified as either directory-based (aka structured) or directory-less (aka unstructured), or hybrid peer-to-peer (P2P). In the scenario of a directory-based search, service providers register their services with either one central and possibly replicated directory or multiple distributed (federated) service directories at distinguished nodes of the underlying network. Service consumers are informed about available services in the network only through these directory nodes.

Centralized directory-based service discovery can be performed by using either a contemporary web search engine or a specialized web service search engine or a dedicated and authoritative web service directory with query interface. In any case, the W3C web service interaction lifecycle for service-oriented architectures (SOA) expects a central service directory to act as an intermediary between provider and consumer (cf. Fig. 1), though it represents a potential single point of failure and performance bottleneck for dependant applications.

Decentralized directory-based service discovery relies on a structured P2P network overlay and a respective query routing protocol. In this case, services are placed and discovered by all peer nodes according to the global distribution or replication scheme and the location mechanism of the network. Classic examples of structured P2P overlays are the DHT-based Chord ring, Pastry, Tapestry, CAN, P-Grid, or a compound routing index, and a hierarchically structured federation of service directories with super-peers. In general, this type of service discovery provides a search guarantee in the sense of total recall and logarithmic complexity in the size of the network for finding popular, i.e., highly replicated, as well as rare services. On the other hand, it comes at the cost of high communication overhead for



Service Discovery, Fig. 1 W3C web service interaction life cycle

publishing and maintaining the structured overlay when peers are joining or leaving the network or the set of services which they provide changes.

Directory-less service discovery is performed in an unstructured P2P network without any given overlay structure. Each peer initially knows only about services provided by its own or its direct neighbor peers. Prominent examples of service location or query routing schemes in such networks are query flooding and k-random walks with replication and caching strategies, as well as informed probabilistic adaptive search. This type of service discovery is effective for finding popular but not rare services and provides only probabilistic search guarantees, i.e., incomplete recall.

Hybrid P2P service discovery is performed in networks with structured and unstructured overlay parts. For example, service requests can be routed to super-peers in the structured overlay part in order to find relevant rare services or processed with restricted flooding or broadcasting to peers of the unstructured network part to find relevant popular services.

Service Selection The performance of service discovery depends, in particular, on the used service selection method. The process of service selection (aka service matchmaking) encompasses (a) the pairwise semantic matching of a given service request with each service that is registered with the matchmaker and (b) the semantic relevance ranking of these services. In contrast to service brokers, a matchmaker only returns a rank list of relevant services and related provenance information to its human user or application but does not handle the interaction with selected services. In principle, a matchmaker can be used for any organizational approach to service discovery. For example, matchmakers can be part of either the query interface of one central directory or federated directories or local directories owned by peers in an unstructured P2P network (Klusch and Sycara 2001).

Types of Service Selection Current approaches to the semantic matching of web services can be classified as non logic-based, logic-based, or hybrid, depending on the nature of reasoning means used for this purpose. Non logic-based semantic matching exploits, for example, means of graph matching, schema matching, data mining, and text similarity measurement, while logic-based semantic matching performs logical reasoning on service descriptions. Hybrid semantic matching is a combination of both types of matching, while adaptive selection means learn how to best

aggregate different matching filters off or on line. In any case, it is commonly assumed that service requests and offers are given in the same format or are appropriately transformed by the considered service matchmaker.

Benchmarking Systems and tools for service discovery, in particular service matchmakers, can be evaluated according to the following five criteria: (1) the support of different service description formats and languages; (2) the usability of the tool and required amount of effort for its configuration; (3) the support of service composition planning through, for example, context-aware pruning of the search space or interactive recommendations for a step-wise forward or backward chaining of services by the user; (4) the policy to preserve user data privacy; and (5) the service retrieval performance in terms of correctness and average query response time over given service test collections. Correctness is commonly evaluated with classical information retrieval measures such as average precision and macro-averaged precision at standard recall levels for binary relevance, as well as the normalized discounted cumulative gain or Q measure for graded relevance. Current evaluation initiatives include the WS Challenge and the SWS Challenge for (semantic) web service composition and the S3 Contest for semantic web service selection (Klusch 2012; Küster et al. 2009).

Web Service Discovery

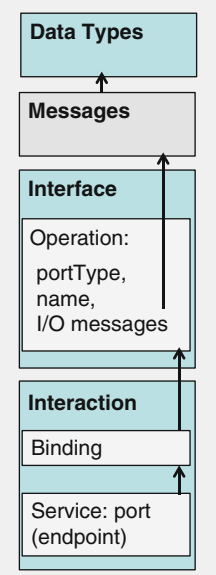
Most web services are described in the standard WSDL, USDL, or according to the REST paradigm of the web. Some service providers also publish the functional description of their services in multiple formats and languages. The number and variety of web services which are available in the public web appears tremendously high, though there are still no common and comprehensive statistics on the subject available. However, the portal seekda.com reported about 30k web services in November 2011, and the public directory programmableweb.com alone already offered about 16k single or composite RESTful web services in March 2013. In this section, we focus on the discovery of WSDL and RESTful services.

WSDL Services The W3C web services framework offers a set of technical specifications including WSDL and SOAP SOAP 2007 that codify mechanisms for XML-based interoperability between business services that are accessible in the web over stateless HTTP. A web service which profile is described in WSDL (in short WSDL service) exposes one or multiple operations which consume inputs and produce outputs both encoded in XML. Applications or other services can interact with these operations by means of XML-SOAP messaging.

Description The XML-based W3C standard language WSDL describes the functionality of a service by the set of signatures of its service operations and the set of network endpoints or ports (URIs) at which these operations can be invoked and how this can be achieved (Fig. 2). In particular, each port is associated with a respective interface which binds the operation to a given protocol for transport and messaging. The definitions of the I/O messages of each service operation include references to their data types which are defined in common XMLS namespaces. Several non functional service parameters can be added to such a WSDL service profile on demand. The description of service profiles in WSDL remains stateless, since the specification of service preconditions and effects is not part of the standard. Besides, a WSDL service description does not include any process model. In this respect, WSDL is commonly considered as weak in describing what the service actually does.

Discovery and Selection Most approaches to directory-based or directory-less discovery of WSDL services utilize means of non logic-based semantic selection, in particular, structural XML and text similarity-based matching.

Central directory-based discovery of WSDL services is the most popular. One classic example is the instantiation of the W3C service interaction lifecycle (cf. Fig. 1) with some UDDI-compliant




```
<description xmlns="http://www.w3.org/ns/wsdl">
<types> <xs:schema … xmlns:pdc="http://www.parts-
          depot.com/schemas/pdc" … >
  <xs:element name = "cid" type = "xs:string"/>…</types>
<interface name = "PartsListInterface">
 <operation name = "GetPartsList"
          pattern = "http://www.w3.org/ns/wsdl/in-out"…>
  <input messageLabel = "In" element = "xs:cid" />
  <output messageLabel="Out" element = "pdc:parts_list"/>
 </operation> </interface>
<binding name = "PListHTTPBinding"
 type = "http://www.w3.org/ns/wsdl/http"
 interface = "tns:PartsListInterface"
 <operation ref = "tns:GetPartsList" whttp:method="GET"/>
</binding>
<service name = "PartsDepot"
          interface = "tns:PartsListInterface">
 <endpoint name ="PListHTTPEndpoint
          binding = "tns:PListHTTPBinding"
          address = "http://www.parts-depot.com/parts/">
</endpoint> </description>
```

Service Discovery, Fig. 2 Example of web service description in WSDL

(Bellwood et al. 2004) registry of WSDL services and using SOAP (Mitra and Lafon 2007) for service interaction. In such an XML-based UDDI business registry (UBR), the services and their providers are categorized with standard taxonomies such as NAICS, SIC, and UNSPSC. Registration of WSDL services and their retrieval from a UBR is through its APIs PublishSOAP and InquireSOAP. In general, a UBR may provide information on the business entities of services (aka white pages), service categories (aka yellow pages), and the technical model (tModel) of services (aka green pages). Search queries to a UBR are regular expressions with identifiers and keywords for service tModels, names, and categories. Accordingly, service selection by a UBR is, in principle, based on string matching without any logical reasoning on service relationships or non functional service parameters. Thus, it requires a rather cumbersome browsing of the registry by the user to find relevant services. Since 2005, UDDI is not supported by its originally main supporters IBM and Microsoft.

Examples of non-UDDI compliant WSDL service directories are RemoteMethods.com, Xmethods.net, WebserviceX.net, webservicelist.com, service-repository.com, and wsindex.org. Most of them rely on keyword search, and service category or simple list browsing. An example of a specialized web service search engine is Woogle (Dong et al. 2004) which retrieves and indexes WSDL services from a given set of UBRs. The WSDL service selection tool WSDLAnalyzer (Zinnikus et al. 2006) returns a rank list of similar WSDL services for a given WSDL service and produces a mapping between their I/O messages. In particular, it recursively computes the XML-tree similarity of a given pair of WSDL files with integrated text matching of tree node names, using WordNet-distance and string matching, and a binary compatibility check of XMLS data types. Other approaches to WSDL service selection exploit techniques for matching software components, graphs, or schemas (Stroulia and Wang 2005), or perform a full-text matching of service names or the content of WSDL files as a whole. In addition, there are approaches to preference-, trust- or reputation-based matching of non functional parameters including quality of service, pricing, and service policies (Crasso et al. 2011; Garofalakis et al. 2006).

Decentralized directory-based discovery of WSDL services in structured P2P networks still appears in its infancies. One example is the PWSD system (Li et al. 2004) in which WSDL files and requests are distributed and located in a Chord ring of service peers. The DUDE system (Banerjee et al. 2005) enables WSDL service discovery in a hierarchical DHT-based overlay for multiple local UDDI registries. There is no approach to directory-less discovery of WSDL services in unstructured P2P networks available yet.

REST Services Web service interaction is not restricted to XML-SOAP messaging. A RESTful web service (in short REST service) represents resources which states shall be accessed only over the stateless HTTP according to the REST paradigm of the web Fielding and Taylor 2002. The call of a REST service with given input values may return output values in XML or in the text-based JSON or RSS formats. For example, the call of some REST service "books" hosted at a portal www.bookstore.com with input parameter "subject" for books on the topic Eclipse is of the form http://www.bookstore.com/books/?subject=computers/eclipse and may return book list entries like <booklist:book url = http://www.bookstore.com/books/0321288157 title="Eclipse Distilled"/> in XML.

Description At present, there is no standard for describing the functionality of REST services. Most REST service APIs are documented by their developers on dedicated, public HTML pages in more or less plain text and tables; some APIs are described in XML-based WADL files or the HTML micro-format hRESTS. This heterogeneity is a major barrier for the automated discovery of REST service APIs in the web to date.

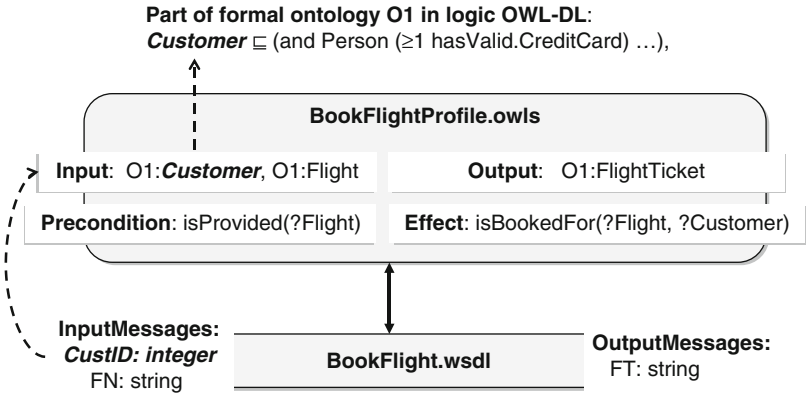
Discovery and Selection Centralized *directory-based* discovery of REST services can be performed with the prominent directory programmableweb.com. It offers about 9k REST service APIs and 7k REST service mash-ups (as of April 2013). Another open source REST API directory in the web is APIS.io. The selection of relevant REST services through their query interfaces is done by keyword search which relies on the textual description of the registered service APIs or other meta-information provided by their developers. The web services search engine seekda! identifies relevant REST service APIs based on adaptive text classification and feature extraction. An approach to automated extraction of information from REST service APIs like service operation name, description, and URI is proposed in Ly et al. (2012). It integrates means of DOM processing, information extraction, and natural language processing. An approach to structural and textual matching of REST services is proposed in Khorasgani et al. (2011). In this case, a given pair of REST service APIs is first semi automatically converted into WADL descriptions. The REST service matching score is then computed as the maximum flow in the graph of WADL service description elements.

Approaches to *directory-less* discovery of REST services in mobile ad hoc networks mostly rely on simple lookup methods based on the matching of service classes, UUID, or service attribute names (Schiele et al. 2004).

Semantic Web Service Discovery

One major challenge of automated service discovery is to make service-based applications or intelligent software agents actually "understand" the semantics of service requests and offers. From the perspective of strong AI, this requires some well-founded logic specification of service profile and process model. However, contemporary web service descriptions are lacking such formal semantics. It is well known that this problem can be addressed by exploiting semantic web technologies (Hitzler et al. 2011).

Description The key idea of encoding web service semantics not only in a machine-readable but machine-*understandable* way is as follows: The semantics of web service interface elements are described by references to appropriate concepts and rules which are formally defined in a



Service Discovery, Fig. 3 Example of semantic service profile in OWL-S

shared ontology in some W3C standard ontology language like RDFS or OWL2. Such semantically annotated web services are called semantic web services (in short semantic services). Current frameworks for semantic service description include OWL-S (Martin et al. 2004), WSML (De Bruijn and Lausen 2005), the W3C standard SAWSDL (Farrell and Lausen 2007), and Linked USDL (Pedrinaci and Leidig 2011) which is USDL modeled in RDFS. These ontology-based semantic service description languages mainly differ in their formal logic-based foundation and the possible extent of annotating services.

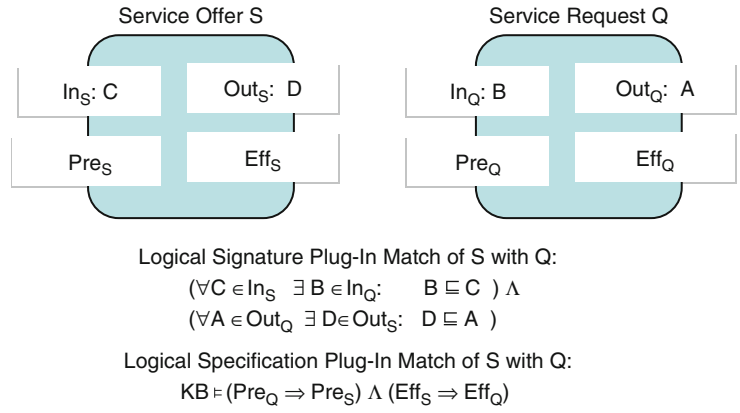
OWL-S In OWL-S the service I/O parameters are annotated with concepts which are exclusively defined in the formal logic-based W3C standard ontology language OWL2 (cf. Fig. 3). Service preconditions and effects may be specified in the formal semantic web rule language SWRL.

WSML The description of service profile semantics in one of five variants of WSML is formally grounded in the respective variant of the logic programming language F-Logic (Fensel et al. 2010). Both, WSML and OWL-S, are also providing the developer with a set of workflow operators like sequence, iterate, choice, and split+join for specifying the operational semantics of a single or composite service in its process model. The process model can be mapped to service orchestrations in BPEL as the semantic service can be grounded with a WSDL service.

SAWSDL and SA-REST The W3C standard SAWSDL allows the annotation of WSDL service elements with references to web resources of any media type such as plain text, video, picture, audio podcast, and concepts in a formal ontology. The same approach is taken in the SA-REST framework for semantically annotating REST service APIs (Gomadam et al. 2010). Both SAWSDL and SA-REST do not allow the specification of preconditions and effects, and the handling of semantic annotations is completely outside these frameworks. In this sense, unlike OWL-S and WSML, neither of both has unique formal semantics. For more details on semantic service description, the reader is referred to, for example, Klusch (2008a) and the above cited relevant technical specifications.

At present, there are no public statistics about semantic web services available. A survey conducted with the semantic service search engine Sousuo (Klusch and Xing 2008) in April 2013 reported about 3,500 semantic services in OWL-S, WSML, WSDL-S, and SAWSDL in the public web, though most of them are available only in distinguished test collections.

Discovery and Selection In the past decade, the semantic web research community has developed



Service Discovery, Fig. 4 Logic-based semantic service plugin matching

a wide range of solutions for the automated discovery and selection of semantic services. The degree of semantic correspondence between a pair of semantic web services particularly relies on the matching of the semantic annotations of their service profile and/or process model.

Types of Selection The types of semantic service selection are logic-based, non logic-based, and hybrid semantic. Classical examples of logic-based semantic matching filters are the logical I/O concept subsumption-based plugin match of service signatures and the logical specification plugin match of preconditions and effects (cf. Fig. 4). Logical and full functional (IOPE) profile matching combines the scores of logical signature (IO) and specification (PE) matching. Non logic-based semantic matching of annotated service signatures is mostly based on the textual similarity of the concept names or the text of their logical unfolding in the referenced ontology. Additional examples include the structural similarity-based matching of I/O concepts in terms of the shortest path or upward co-topic distances between them in the shared ontology.

Currently, most approaches to semantic service selection are hybrid, i.e., they combine non logic-based with logic-based semantic service matching. Besides, the majority of them support either OWL-S or SAWSDL, but only a few are devoted to WSML, or other description formats, and hardly any matchmaker is even language-agnostic (Klusch 2012). In the following, we focus on approaches to the discovery and selection of services in OWL-S and SAWSDL. More information on the subject is provided, for example, in Klusch (2008b, 2012).

Centralized Discovery and Selection There are quite a few tools and systems for central directory-based discovery of semantic services available.

Matchmakers. For example, the matchmaker iSeM (Klusch and Kapahnke 2012) performs an adaptive and hybrid semantic selection of OWL-S services. Its logic-based semantic matching of services relies on the computation of strict and approximated logical I/O concept subsumption relations and the logical specification plugin relation. Like its predecessor OWLS-MX2 (Klusch et al. 2009), it also performs non logic-based semantic matching with different classical token-based text similarity measures, as well as ontology-based structural matching of signature annotation concepts. Finally, it learns how to best aggregate the results of its matching filters by use of a binary SVM relevance classifier with an evidential coherence-based weighting scheme.

An example of a hybrid semantic and adaptive matchmaker for SAWSDL services is LOG4SWS (Schulte et al. 2010). Like iSeM it performs a logical service signature matching which is complemented with ontology-based structural matching based on the shortest path lengths between concepts. In case there are no semantic annotations of WSDL service signature elements,

it exploits the WordNet distance between the element names. LOG4SWS does not consider service preconditions and effects, but learns off line how to best aggregate the matching results by use of an ordinary least square-based classifier.

The logic-based semantic service matchmaker SPARQLent (Sbodio et al. 2010) considers the full functional profile of OWL-S services. It performs an RDF entailment rule-based matching of I/O concepts, preconditions, and effects described in SPARQL.

According to the results of the international S3 contest (Klusch 2012), iSeM and LOG4SWS are currently the best performing matchmakers for OWL-S and SAWSDL services, respectively. In fact, they provide the best trade-off between average precision and response time.

An example of a hybrid semantic matchmaker for WSML services is WSMO-MX (Klusch and Kaufer 2009): It recursively determines service matching degrees based on ontology-based signature parameter type matching, logical constraint (PE) matching, and syntactic matching with text similarity measurements.

Specialized search engines. Examples of search engines for semantic services are S3E (Giantsiou et al. 2009) and Sousuo (Klusch and Xing 2008). The latter performs a meta-search through the public web search engines Google and A9 and complements it by crawling the web with its own focused topic crawler. It also utilizes the semantic web search engine Swoogle for an inverse ontology-based search and performs a full-text search of the public scientific archive citeseer in the web. Service selection through Sousuo's query interface relies on full-text or keyword search in its XML-encoded service index.

Alternatively, the S3E engine is encoding the profiles of crawled semantic services in RDF. The selection of services from an internal RDF store with SPARQL relies, in particular, on textual matching of profile parameters. Another search engine which is restricted to a QoS-based discovery of semantic services is presented in Vu et al. (2006).

Registries. At present, there are no central and authoritative registries of semantic services available in the public web. Public collections of semantic services are, for example, the prominent OWLS-TC for OWL-S services, the SAWSDL-TC for SAWSDL services, and hREST-TC for annotated REST services; each of these collections is available at the portal semwebcentral.org. iServe (Pedrinaci et al. 2010) is a software platform that can be used to build and maintain a registry of semantic services described in SAWSDL, OWL-S, MicroWSMO, and WSMO-Lite. The services are internally represented in iServe according to a minimal service model and then exposed in HTML and RDF as linked services with a unique and resolvable HTTP URI. Any iServe registry can be queried through an SPARQL endpoint. For service selection, iServe provides means of keyword search, functional classification, and service I/O parameter matching based on RDFS reasoning.

Centralized P2P search. An example for the discovery of WSDL-S (a predecessor of SAWSDL) services in a structured P2P system is the METEOR-S system (Verma et al. 2005). It consists of a set of service-providing and service-consuming peers which may form groups on given domains or topics and one central super-peer which serves as a central service matchmaker for all peers. For this purpose, the super-peer maintains and utilizes a global registry ontology which covers the concept taxonomies of all local service registries of peers in the network. The super-peer also provides the peers with mappings between the message types and signature annotation concepts of registered services. The non logic-based semantic selection of services by the super-peer relies on structural XMLS matching and the computation of NGram-based text similarities and taxonomic relations. The super-peer can be replicated for reasons of scalability.

Decentralized Discovery and Selection A *directory-based* discovery of OWL-S services in structured P2P systems can be performed, for example, with the AGORA-P2P system (Küngas and Matskin 2006). It relies on a Chord ring for distributed storage and location of services. In particular, the service signature concept labels are hashed as literals to unique integer keys such

that peers holding the same key are offering services with equal literals in the circular key space. Service selection for multi-key queries relies on exact key matching.

Directory-less discovery of semantic services can be performed with, for example, the RS2D system (Basters and Klusch 2006). It is a solution for informed and adaptive probabilistic service search in unstructured P2P networks. In particular, each peer dynamically builds and maintains its local view of the semantic overlay of the network and uses the OWLS-MX matchmaker for hybrid semantic service selection. A peer also learns the average query-answering behavior of its direct neighbors in the network. The peer's decision to whom to forward a semantic service request is then driven by its estimated probabilistic risk of routing failure in terms of semantic loss and communication costs. Other examples are discussed, for example, in Klusch (2008b) and Staab and Stuckenschmidt (2006).

Future Directions

Despite the progress made in the field in the past decade, a major open problem is the scalable and dynamic interleaving of discovery of services with their composition, negotiation, and execution in the converging Internet of Things and Internet of Services. Examples of potential applications of solutions are intelligent condition monitoring based on large-scale, wireless, and semantic sensor service networks; the intelligent collaborative design of products in shared 3D spaces; and mobile ad hoc and context-aware business travel planning or product recommendation services.

Cross-References

- RDF
- Web Ontology Language (OWL)
- Web Service Composition
- WSDL

References

Banerjee S, Basu S, Garg Sh, Garg S, Lee SJ, Mullan P, Sharma R (2005) Scalable grid service discovery based on UDDI. In: Proceedings of the 3rd international workshop on middleware for grid computing, Grenoble. ACM

Basters U, Klusch M (2006) RS2D: fast adaptive search for semantic web services in unstructured P2P networks. In: Proceedings of the 5th international semantic web conference, Athens. Springer

Bellwood P, Capell S, Clement L, Colgrave J, Dovey MJ, Feygin D, Hately A, Kochman R, Macias P, Novotny M, Paolucci M, von Riegen C, Rogers T, Sycara K, Wenzel P, Wu Z (2004) UDDI 3.02. uddi.org/pubs/uddi_v3.htm

Chinnici R, Moreau JJ, Ryman A, Weerawarana S (2007) Web services description languages 2.0. W3C recommendation. www.w3.org/TR/wsdl20/, WSDL1.1(2001):www.w3.org/TR/wsdl

Crasso M, Zunino A, Campo M (2011) A survey of approaches to web service discovery in service-oriented architectures. J Database Manag 22(1):102–132. IGI Global

De Bruijn J, Lausen H (eds) (2005) Web service modeling language (WSML). www.w3.org/Submission/WSML/

Dong X, Halevy AY, Madhavan J, Nemes E, Zhang J (2004) Simlarity search for web services. In: Proceedings of the 30th conference on very large databases, Toronto

Farrell J, Lausen H (2007) Semantic annotations for WSDL and XML schema. www.w3.org/TR/sawsdl/

Fensel D, Lausen H, Polleres A (2010) Enabling semantic web services. Springer, Heidelberg

Fielding RT, Taylor RN (2002) Principled design of the modern web architecture. J Trans Internet Technol 2(2):115–150. ACM

Garofalakis J, Panagis Y, Sakkopoulos E, Tsakalidis A (2006) Contemporary web service discovery mechanisms. J Web Eng 5(3):265–289. Rinton Press

Giantsiou L, Loutas N, Peristeras V, Tarabanis K (2009) Semantic service search engine (S3E): an approach for finding services on the web. In: Proceedings of the 2nd world summit on the knowledge society, Crete. Springer

Gomadam K, Ranabahu A, Sheth A (2010) SA-REST: semantic annotation of web resources. www.w3.org/Submission/2010/SUBM-SA-REST-20100405/

Hadley M (2009) Web application description language. www.w3.org/Submission/wadl/

Hitzler P, Krötzsch M, Rudolph S (2011) Foundations of semantic web technologies. CRC, Boca Raton/New York/London

Khorasgani RR, Stroulia E, Zaiane OR (2011) Web service matching for RESTful web services. In: Proceedings of the 13th IEEE international symposium on web systems evalution. IEEE, Williamsburg, USA

Klusch M (2008a) Semantic web service description. In: Schumacher M, Helin H, Schuldt H (eds) CASCOM –

intelligent service coordination in the semantic web, chapter 3. Birkhäuser, Basel
Klusch M (2008b) Semantic web service coordination. In: Schumacher M, Helin H, Schuldt H (eds) CASCOM – intelligent service coordination in the semantic web, chapter 4. Birkhäuser, Basel
Klusch M (2012) Overview of the S3 contest: performance evaluation of semantic service matchmakers. In: Blake MB, Cabral L, König-Ries B, Küster U, Martin D (eds) Semantic web services, chapter 2. Springer, Berlin/London
Klusch M, Kapahnke P (2012) The iSeM matchmaker: a flexible approach for adaptive hybrid semantic service selection. J Web Semant 15:1–14. Elsevier
Klusch M, Kaufer F (2009) WSMO-MX: a hybrid semantic web service matchmaker. J Web Intell Agent Syst 7(2):1–14. IOS Press
Klusch M, Sycara K (2001) Brokering and matchmaking for coordination of agent societies: a survey. In: Omicini A et al (eds) Coordination of internet agents, chapter 8. Springer, Berlin/New York
Klusch M, Xing Z (2008) Deployed semantic services for the common user of the web: a reality check. In: Proceedings of the 2nd IEEE international conference on semantic computing (ICSC), Santa Clara. IEEE
Klusch M, Fries B, Sycara K (2009) OWLS-MX: a hybrid semantic web service matchmaker for OWL-S services. J Web Semant 7(2):121–133. Elsevier
Kopecky J, Gomadam K, Vitvar T (2008) HTML microformat for describing RESTful web services and APIs. In: Proceedings of the international conference on web intelligence, Sydney. IEEE
Küngas P, Matskin M (2006) Semantic web service composition through a P2P-based multi-agent environment. In: Proceedings of the 4th international workshop on agents and peer-to-peer computing, Utrecht. LNCS 4118. Springer
Küster U, Koenig-Ries B, Klusch M (2009) Evaluating semantic web service technologies: criteria, approaches and challenges. In: Lytras MD, Sheth A (eds) Progressive concepts for semantic web evolution: applications and developments. IGI Global, Hershey
Li Y, Zou F, Wu Z, Ma F (2004) PWSD: a scalable web service discovery architecture based on peer-to-peer overlay network. In: Proceedings of the APWeb04, Hangzhou. LNCS 3007. Springer
Ly PA, Pedrinaci C, Domingue J (2012) Automated information extraction from Web APIs documentation. In: Proceedings of the 13th international conference on web information system engineering, Paphos. Springer
Martin D, Burstein M, Hobbs J, Lassila O, McDermott D, McIlraith S, Narayanan S, Paolucci M, Parsia B, Payne T, Sirin E, Srinivasan N, Sycara K (2004) OWL-S.: semantic markup for web services. www.w3.org/Submission/OWL-S/
Mitra N, Lafon Y (2007) SOAP 1.2 Primer. www.w3.org/TR/2007/REC-soap12-part0-20070427/
Oberle D, Barros A, Kylau U, Heinzl S (2013) A unified description language for human to automated services. J Inf Syst 38(1):155–181. Elsevier
Pandit B, Popescu V, Smith V (2009) Service modeling language 1.1. www.w3.org/TR/sml/
Pedrinaci C, Leidig T (2011) Linked USDL Core. www.linked-usdl.org/ns/usdl-core
Pedrinaci C, Liu D, Maleshkova M, Lambert D, Kopecky J, Domingue J (2010) iServe: a linked services publishing platform. In: Proceedings of the 7th extended semantic web conference workshop on ontology repositories and editors for the semantic web, Heraklion. LNCS. Springer
Sbodio ML, Martin D, Moulin C (2010) Discovering semantic web services using SPARQL and intelligent agents. J Web Semant 8(4):310–328. Elsevier
Schiele G, Becker C, Rothermel K (2004) Energy-efficient cluster-based service discovery for ubiquitous computing. In: Proceedings of the 11th ACM SIGOPS European workshop, Belgium
Schulte S, Lampe U, Eckert J, Steinmetz R (2010) LOG4SWS.KOM: self-adapting semantic web service discovery for SAWSDL. In: Proceedings of the 6th world congress of services, Miami. IEEE
SOAP (2007) W3C Recommendation SOAP 1.2. www.w3.org/TR/2007/REC-soap12-part0-20070427/
Staab S, Stuckenschmidt H (eds) (2006) Semantic web and peer-to-peer. Springer, Berlin
Stroulia E, Wang Y (2005) Structural and semantic matching for assessing web-service similarity. J Coop Inf Syst 14:407–437. World Scientific
Verma K, Sivashanmugam K, Sheth A, Patil A, Oundhakar S, Miller J (2005) METEOR–S WSDI: a scalable P2P infrastructure of registries for semantic publication and discovery of web services. J Inf Technol Manag 6(1):17–39. Kluwer
Vu L-H, Hauswirth M, Porto F, Aberer K (2006) A search engine for QoS-enabled discovery of semantic web services. J Bus Process Integr Manag 1(4):244–255
Zinnikus I, Rupp H-J, Fischer K (2006) Detecting similarities between web service interfaces: the WSDL analyzer. In: Proceedings of the 2nd international workshop on web services and interoperability, Bordeaux. Wiley

Service-Oriented Architectures

▸ Service Discovery

Service Search and Selection

▸ Service Discovery

Service Systems

▸ Queueing Theory

Sex Industry

▸ Pornography Online

Sharding

▸ Weblog Analysis

Siena: Statistical Modeling of Longitudinal Network Data

Tom A. B. Snijders
Department of Statistics and Nuffield College, University of Oxford, Oxford, UK
Department of Sociology, University of Groningen, Groningen, The Netherlands

Synonyms

Coevolution of networks and behavior; Network dynamics; Network panel data; Peer influence; Statistical modeling

Glossary

Network Panel Data Longitudinal data consisting of two or more repeated observations of a network on a given set of nodes

Panel Wave The data observed for one given observation moment in a panel study

Social Actors Individuals, companies, etc., represented by the nodes in the network

Stochastic Actor-Oriented Model A probability model for network dynamics where changes may take place at arbitrary moments in continuous time and where these changes are regarded as consequences of choices made by the actors

RSiena R package implementing statistical inference according to a stochastic actor-oriented model given network panel data

Effects Model components defining the probabilities of tie changes in the stochastic actor-oriented model

Method of Moments One of the traditional methods in statistics for parameter estimation

Dependent Variable The variable defining the outcome space in a statistical model

Definition

The name "Siena" stands for *Simulation Investigation for Empirical Network Analysis*. It is a method for the statistical analysis of longitudinal network data, observed in two or more panel waves. This method was implemented in the stand-alone program *Siena*, first released in 1997, going through many versions, and superseded by the R package *RSiena* in 2009. *Siena* was programmed by Tom Snijders in Delphi, with contributions by Christian Steglich, Mark Huisman, and Michael Schweinberger. *RSiena* was originally programmed by Ruth Ripley and Krists Boitmanis, under the direction of Tom Snijders. Since 2012 it is maintained by Tom Snijders, in collaboration with Christian Steglich and Johan Koskinen; other contributors are Josh Lospinoso, Charlotte Greenan, and Paulina Preciado.

RSiena is a contributed package of the statistical software system R and as such is free, distributed under the GNU General Public License, running under Unix-like, Windows, and Mac families of operating systems. The methods used are based on Monte Carlo simulation and therefore can be time-consuming for larger data sets. The package is programmed in a combination of R and C++, the latter for the computationally intensive parts.

The orientation of the *Siena* method is primarily to the social sciences, but this of course is not exclusive.

Introduction

Statistical modeling is based on model assumptions, mostly assumptions about independence or conditional independence, but one of the main

characteristics of networks is the strong and complex dependence between network ties. If the assumptions in statistical modeling are not good representations of the data structures or the mechanisms that may have led to the observed data, then results of statistical inference can be grossly misleading. This leads to difficulties in proposing plausible statistical models for network data.

Modeling longitudinal network data can be simpler than modeling single observations of networks, because the time structure poses a constraint on the dependence structure: the present depends on the past, not on the future. The *Siena* method is based on a probability model that represents network dynamics as a Markov chain running in continuous time, called the *stochastic actor-oriented model*. The basic state space is the set of all digraphs on a given node set. The model has been expanded to allow multiple (i.e., multivariate) digraphs and also actor-based variables as components of the state space. The actor-based variables are usually referred to as "behavior," thus allowing the modeling of the coevolution, or interdependent dynamics, of networks and behavior. Thus, in the basic type of stochastic actor-oriented model, there is one dependent variable, viz., a directed network; in the extended models, there can be several dependent networks and also one or more dependent actor-based variables. Two-mode networks can also be included as dependent networks.

Key Points

The *Siena* method is defined in the usual paradigm of statistical modeling. It presupposes the availability to the user of *network panel data* or *network and behavior panel data*. This means that for a given node set, at a finite number (two or more) of observation points (also called *panel waves*), a network on this node set was observed, possibly complemented with a behavioral variable. In the ideal case the node set is constant and the data are complete; some changes in the node set (nodes entering or exiting) and some fraction of data being missing are allowed. Up to 10 % of missing tie variables are in practice not a problem; more than 20 % are not advisable.

The nodes are supposed to represent *social actors*, and the model is said to be *actor-oriented*, meaning that tie changes are regarded as the consequence of choices made by the senders of the ties. The user specifies a model by defining a set of *effects* (see below), which are model components defining the probabilities of tie changes. Given the model specification, the *RSiena* package can estimate parameters (which are coefficients indicating the strength of the effects) and test hypotheses about the parameters. With a given specification and given parameters, *RSiena* can also be used to simulate the dynamics of a network.

Historical Background

A historical overview of early work on probability models for network dynamics is given in Snijders (1995), which also was the first paper about stochastic actor-oriented models. Some important papers that are part of the general background preceding the work on this methodology are Holland and Leinhardt (1977), Wasserman (1979, 1980), Zeggelink (1994), and Leenders (1995).

The development of the stochastic actor-oriented model for digraphs was stimulated by the empirical work in van de Bunt (1999). After two precursor papers (Snijders 1996; Snijders and van Duijn 1997), the main presentation of this model was given in Snijders (2001). Methods for the coevolution of networks and behavior were developed in Snijders et al. (2007) and elaborated in Steglich et al. (2010). All these papers use the Method of Moments (one could also say the method of estimating equations) to estimate the parameters. This is the main method implemented in *RSiena*. In addition, Bayesian methods (Koskinen and Snijders 2007) and an algorithm for Maximum Likelihood estimation (Snijders et al. 2010a) were developed and implemented, but these are much more time-consuming and therefore are less used.

Statistical Model

This section outlines the basic probability model implemented in *RSiena*, for the basic case of one dependent variable, assuming this is a one-mode network. Further elaboration and details can be found in the publications mentioned above and in Snijders (2009), on which much of the explanation below is based.

The network is represented by the node set $\{1, \ldots, n\}$ with tie variables x_{ij}, where $x_{ij} = 1$ or 0 indicates whether the tie $i \to j$ is present or absent. The tie variables are collected in the $n \times n$ adjacency matrix $\mathbf{x} = (x_{ij})$. Self-ties are excluded, so that $x_{ii} = 0$ for all i. The concepts of network (directed graph) and matrix (its adjacency matrix) will be used interchangeably. Random variables will be indicated by capitals and observations, or other nonrandom variables, by lowercase. The ties are assumed to be outcomes of time-dependent random variables, denoted by $X_{ij}(t)$ and collected in the time-dependent random matrix $\mathbf{X}(t)$.

In addition to the network $\mathbf{X}(t)$, which can be regarded as the dependent variable of the model, there can be other variables, so-called covariates, regarded as independent or explanatory variables in the sense that their values are not modeled but accepted as given, and which may influence the network. Examples are the gender of actors (actor variable) and their spatial proximity (dyadic variable). For conciseness, these are disregarded in this brief overview; in practice, they are included in most data sets and of great practical importance.

Basic Model Definition

The following basic assumptions are made:

1. Time, denoted by t, is a continuous variable.
   This assumption separates time as observed (two or more moments of observation) from time that determines network dynamics (continuous).
2. $\mathbf{X}(t)$ is a Markov process.
   This means that the conditional distribution of future states depends on the past only as a function of the present. This assumption corresponds to the network ties being regarded as states rather than events.
3. At any given moment t, no more than one tie variable $X_{ij}(t)$ can change.

This set of assumptions was first proposed by Holland and Leinhardt (1977) and is very helpful because it allows representing network dynamics as a feedback process, where the actors create the network as the endogenously changing environment for themselves and each other (Zeggelink 1994) while requiring only to specify the probabilities of changes of single tie variables.

In the further model elaboration, two aspects are distinguished: the *change opportunity* process and the *change determination* model.

Opportunity for change. For each actor i, opportunities to establish one new outgoing tie $i \to j$, or dissolve one existing tie $i \to j$, occur according to a Poisson process with rate λ_i. This means that the probability that an opportunity for change occurs for actor i in the time interval from t to $t + \epsilon$, where ϵ is a small positive number, is approximated (in the limit for ϵ tending to 0) by $\lambda_i \epsilon$.

Determination of change. When actor i has an opportunity for change, she/he is permitted to choose one of the outgoing tie variables X_{ij} and change this into its opposite value, changing 0 to 1 (creating a new tie) or changing 1 to 0 (terminating an existing tie). The probabilities depend on the so-called objective function $f_i(\mathbf{x}^0, \mathbf{x})$, indicating how "attractive" it is to go to state $\mathbf{x}$ given the current state $\mathbf{x}^0$. The set of potential new network states, denoted by $\mathcal{C}(\mathbf{x}^0)$, is the set composed of $\mathbf{x}^0$ itself together with the $n - 1$ matrices which are equal to $\mathbf{x}^0$ except for exactly one non-diagonal element in line i which is replaced by its opposite, $x_{ij} = 1 - x^0_{ij}$. The probability that the new state is $\mathbf{x}$ is given by

$$\begin{aligned} &\mathrm{P}\{\mathbf{X}(t) \text{ changes to } \mathbf{x} \mid i \text{ has a change} \\ &\text{opportunity at time } t, \mathbf{X}(t) = \mathbf{x}^0\} \\ &= p_i(\mathbf{x}^0, \mathbf{x}) = \frac{\exp\left(f_i(\mathbf{x}^0, \mathbf{x})\right)}{\sum_{\mathbf{x}' \in \mathcal{C}(\mathbf{x}^0)} \exp\left(f_i(\mathbf{x}^0, \mathbf{x}')\right)} . \end{aligned} \tag{1}$$

The two model components can be put together by giving the transition rate matrix, also called Q-matrix, of which the non-diagonal elements are defined by

$$q_{\mathbf{x}^0,\mathbf{x}} = \lim_{dt \downarrow 0} \frac{\mathrm{P}\{\mathbf{X}(t+dt) = \mathbf{x} \mid \mathbf{X}(t) = \mathbf{x}^0\}}{dt} \quad (\mathbf{x} \neq \mathbf{x}^0)$$

(see textbooks on continuous-time Markov chains, such as Norris 1997). Note that the assumptions imply that

$$q_{\mathbf{x}^0,\mathbf{x}} = 0 \text{ whenever } x_{ij} \neq x^0_{ij} \text{ for more than one element } (i,j).$$

For digraphs $\mathbf{x}$ and $\mathbf{x}^0$ which differ from each other only in one element in row i, the transition rate is

$$q_{\mathbf{x}^0,\mathbf{x}} = \lambda_i(\mathbf{x}^0)\, p_i(\mathbf{x}^0, \mathbf{x}) \,. \tag{2}$$

Model Specification

The model specification consists of defining the network $\mathbf{X}$ (and other dependent variables, if any; see Steglich et al. 2010, and Snijders et al. 2013), covariates, the rate function λ_i, and the objective function f_i. If there are more than one dependent variable, each has its own rate function and objective function. The rate function may be constant between waves or depend on actor-based variables through an exponential link function. The focus of model specification is on the objective function, specified as a linear combination

$$f_i(\mathbf{x}^0, \mathbf{x}) = \sum_k \beta_k\, s_{ki}(\mathbf{x}^0, \mathbf{x}) \tag{3}$$

where the functions s_{ki} are so-called effects driving the network dynamics, while the weights β_k are parameters indicating the strength of these effects and which can be estimated from the data. The effects represent internal network dependencies as well as dependence on covariates and are discussed in the mentioned literature. The manual Ripley et al. (2013) contains the long list of implemented effects, and this list is frequently added to because of requests from applied researchers.

Parameter Estimation

The main estimation method implemented in *RSiena* is an application of the Method of Moments (or estimating equations). It makes good use of the Markov property by conditioning on the preceding observation. This enables computer simulation of the process in a straightforward way and does away with the need for an assumption of stationary marginal distributions. The moment equations, or estimating equations, define the parameter estimate θ as a function of the data $\mathbf{x} = x(t_1), \ldots, x(t_M)$ (assuming there are M waves) and are given by

$$\sum_{m=1}^{M-1} \mathrm{E}_\theta\{U\big(\mathbf{X}(t_m), \mathbf{X}(t_{m+1})\big) \mid \mathbf{X}(t_m) = \mathbf{x}(t_m)\} = \sum_{m=1}^{M-1} U\big(\mathbf{x}(t_m), \mathbf{x}(t_{m+1})\big) \tag{4}$$

for suitable functions $U\big(x(t_m), x(t_{m+1})\big)$ chosen in correspondence with the estimated parameter θ. The choice of the statistics U is discussed in Snijders (2001) and Snijders et al. (2007). The latter publication also specifies the estimating equations for the case of more than one independent variable, which are slightly more involved.

To solve the estimating equation (4), in the absence of ways to calculate analytically the expected values, stochastic approximation methods are used. Variants of the Robbins-Monro (1951) algorithm (see, e.g., Chen 2002, for a more up-to-date treatment) have been used with good success. This is a stochastic iteration method which produces a sequence of estimates $\theta^{(N)}$ which is intended to converge to the solution of (4) and which works here as follows. For a given provisional estimate $\theta^{(N)}$, the model is simulated so that for each $m = 1, \ldots, M-1$, a random draw is obtained from the conditional distribution of $\mathbf{X}(t_{m+1})$ given that $\mathbf{X}(t_m) = \mathbf{x}(t_m)$. This simulated network is denoted $\mathbf{X}^{(N)}(t_{m+1})$. Denote $U^{(N)} = \sum_{m=1}^{M-1} U\big(\mathbf{x}(t_m), \mathbf{X}^{(N)}(t_{m+1})\big)$, and let u^{obs} be the

right-hand side of (4). Then the iteration step in the Robbins-Monro algorithm for obtaining the Method of Moments estimate is given by

$$\theta^{(N+1)} = \theta^{(N)} - a_N D^{-1} \left(U^{(N)} - u^{\text{obs}} \right), \tag{5}$$

where D is a suitable matrix and a_N a sequence of positive constants tending to 0. This equation is reminiscent of the iteration step in the Newton-Raphson algorithm, but in this case the function for which the root is sought is not directly computable, and instead we simulate random variables having this function as their expected value. Tuning details of the algorithm, including the choices of D and a_N, are given in Snijders (2001). The Bayesian estimators for these models presented in Koskinen and Snijders (2007) and the Maximum Likelihood estimators of Snijders et al. (2010a) are also implemented in *RSiena*. Since Maximum Likelihood estimates can also be defined by an equation of type (4), where now U is the score function (and therefore also depends on the parameter θ), also for this purpose the Robbins-Monro algorithm is used.

A general issue for Monte Carlo-based estimation is to assess the convergence of a given run of the estimation algorithm. The output resulting from the Method of Moments as well as Maximum Likelihood estimation algorithms contains simple indications for convergence, the so-called t-ratios for convergence, which indicate the extent to which the estimates found indeed satisfy approximately the equation (4), based on independent simulations with the value of θ resulting from the estimation algorithm. For relatively simple models, it is quite usual that the first run of the algorithm produces good estimates. For more complex models or data sets, it may be necessary to iterate the algorithm, using the estimates obtained as starting values for the next run of the algorithm.

Elements of the Package

The *RSiena* package operates as all R packages by a collection of functions, and the user can mix the use of *RSiena* with using all other functions in R and its contributed packages. Also in line with the R environment, the package is totally object-oriented: data sets, model specifications, estimation results, etc.; all are defined as objects on which the user can operate and about which information can be requested.

Without going into the specifics of the model, it nevertheless may be helpful to indicate briefly the main types of functions that are available:

1. Functions to specify data objects for a specific use (as covariates, dependent variables, etc.) in the model.
2. Functions to specify the model. These create the "`sienaEffects`" objects containing the model specification and further modify such objects.
3. Functions for estimating parameters. The main workhorse here is called `siena07` – the name was given for historical reasons, because in the original *Siena* version 1 suite, this was meant to be the seventh in a sequence of executable programs. Function `siena07` can be used for estimation according to the Method of Moments as well as Maximum Likelihood estimation. It can also be used for simulating the model without parameter estimation. In combination with the estimation, it is also possible to test a hypothesized value for some of the parameters without estimating them by so-called score-type tests (Schweinberger 2012) for the Method of Moments or by regular score tests for Maximum Likelihood estimation.

   Further there are a function `sienaBayes` for Bayesian estimation, and a function `siena08` for the meta-analytic combination of the evidence produced by estimating the same model for a number of independent data sets (Snijders and Baerveldt 2003).
4. Functions for assessing the fit of the model. The main functions of this kind currently are `sienaTimeTest` for testing time homogeneity across multiple waves (Lospinoso et al. 2011) and `sienaGOF` ("goodness of fit") for assessing the adequacy of the model in reproducing a number of features of the data (Lospinoso 2012).
5. A variety of functions for summarizing results obtained.

Key Applications

The *Siena* method as implemented in the *RSiena* package has been applied in a variety of studies in sociology, psychology, political science, and other disciplines. Two special issues of the journal *Social Networks* on network dynamics, published in January 2010 and July 2012, contain a couple of examples. Many applications are listed at the Siena website, http://www.stats.ox.ac.uk/siena/. The following is a very small and somewhat arbitrary selection.

Applications to dynamics of one social network (i.e., without the inclusion of dependent behavior variables) started with van de Bunt et al. (1999), a study of friendship in a group of freshman students. This was the first publication on network dynamics that found statistical evidence for transitivity and for homophily (on gender, age, and smoking behavior) using a method that allows each tested effect to be controlled for all other effects – this being a basic purpose of the *Siena* method. Another example study on friendship development is Selfhout-Van Zalk et al. (2010), concentrating on the effects of personality characteristics (the "Big Five"); finding evidence for homophily with respect to agreeableness, extraversion, and openness to experience; and further concluding (less surprising) that individuals high on extraversion tended to select more friends and individuals high on agreeableness tended to be selected more as friends.

A large group of applications is about *peer effects* or *social influence*, i.e., the question whether individuals are being influenced in their behavior, performance, or attitudes by those to whom they have network ties. It has long been debated whether the similarity between friends with respect to smoking behavior is a consequence of homophilous selection of friends or of social influence. Mercken et al. (2009) applied *Siena* to a data set of 7,704 adolescents (aged 12–15 years) in 70 schools from 6 European countries (Denmark, Finland, the Netherlands, Portugal, the UK, and Spain). They found evidence for homophilous selection in all countries and for peer influence with respect to smoking only in Finland and the Netherlands. Obesity is another health-related variable for which the question of peer influence, and how to assess it, has recently received attention in scholarly journals. De la Haye et al. (2011) found, in a data set of two cohorts in the initial 2 years in high school, that similarities between friends with respect to their body mass index (BMI) were due mainly to processes of friend selection, and not to peer influence. Since the extent of peer influence may well depend on age, family background, cultural and contextual aspects, etc., it is quite plausible that peer influence may differ between countries and social settings. One study can therefore not give a definitive answer about questions of peer selection with respect to variables such as smoking or obesity, and further research is necessary and ongoing.

An example application in political science is Berardo and Scholz (2010), studying governance processes between organizations in 10 US estuaries and how partner selection for collaboration depended on general trust in the institutional environment as expressed by representants of the organizations. They found that partner selection is not directly dependent on trust, but trust is influenced by the trust expressed by collaboration partners.

The *Siena* method has also been applied to network data collected by other methods than self-report surveys. An example is Lewis et al. (2012), a study of Facebook friendships and cultural tastes which concluded that friendship formation is influenced by similarity in taste for music and movies, but not for books, and that there is little influence for diffusion of tastes through Facebook ties, with the exception of a taste for classical/jazz music.

Future Directions

The stochastic actor-oriented methodology and the *RSiena* package are areas of active ongoing research and development. New possibilities are the analysis of multiple dependent networks and behavioral variables, including two-mode

networks and networks with a small number of ordered values and the assessment of goodness of fit. These options have been implemented since 2011 but still require further methodological exploration and practical experience. Current work includes the development of models for the diffusion of innovations in a changing network (Greenan); models with errors in observations (Lospinoso); models for network events associated to an unobserved changing network (Lospinoso); models for continuous dependent behavior variables (Niezink); models with unobserved heterogeneity between actors (Koskinen); and random effect models for multiple groups (Koskinen and Snijders). An expected development is the so-called settings model (Preciado) which is meant to make the actor-oriented approach applicable also to larger networks (with a few hundred to a few thousand nodes); such data sets are not well suited for the current software because the basic model (like other models for network dynamics) makes assumptions of homogeneity and of accessibility of actors to each other that are less plausible for such large networks and because the time taken by the computer simulations becomes prohibitive.

Acknowledgments

Work on the *Siena* method and software has taken place at the University of Groningen (Department of Sociology) and the University of Oxford (Department of Statistics, Department of Politics and International Relations, Nuffield College). NWO (the Netherlands Organisation for Scientific Research) has supported the development of this methodology and software through the project *Statistical methods for the joint development of individual behavior and peer networks* (project number 575-28-012, researcher Mark Huisman), the integrated research program *The dynamics of networks and behavior* (project number 401-01-550, methodological researchers Christian Steglich and Michael Schweinberger), and the ERCP-Eurocores Programme *Models for the Evolution of Networks and Behavior* (project number 461-05-690, methodological researcher Christian Steglich). The US National Institutes of Health have provided funding for methodological developments (Johan Koskinen) and software development (Ruth Ripley and Krists Boitmanis) as part of the project *Adolescent Peer Social Network Dynamics and Problem Behavior* (grant number 1R01HD052887-01A2).

Cross-References

▸ Actor-Based Models for Longitudinal Networks
▸ Analysis and Visualization of Dynamic Networks
▸ Exponential Random Graph Models
▸ Human Behavior and Social Networks
▸ Social Influence Analysis
▸ Statistical Research in Networks – Looking Forward
▸ Temporal Networks
▸ Theory of Statistics, Basics, and Fundamentals

References

Berardo R, Scholz JT (2010) Self-organizing policy networks: risk, partner selection and cooperation in estuaries. Am J Pol Sci 54:632–649

Chen HF (2002) Stochastic approximation and its applications. Kluwer Academic, Dordrecht

de la Haye K, Robins G, Mohr P, Wilson C (2011) Homophily and contagion as explanations for weight similarities among adolescent friends. J Adolesc Health 49:421–427

Holland PW, Leinhardt S (1977) A dynamic model for social networks. J Math Sociol 5:5–20

Koskinen JH, Snijders TAB (2007) Bayesian inference for dynamic social network data. J Stat Plann Inference 13:3930–3938

Leenders R (1995) Models for network dynamics: a Markovian framework. J Math Sociol 20:1–21

Lewis K, Gonzalez M, Kaufman J (2012) Social selection and peer influence in an online social network. Proc Natl Acad Sci USA 109(1):68–72

Lospinoso JA (2012) Statistical models for social network dynamics. DPhil Thesis, University of Oxford

Lospinoso JA, Schweinberger M, Snijders TAB, Ripley RM (2011) Assessing and accounting for time

heterogeneity in stochastic actor oriented models. Adv Data Anal Comput 5:147–176

Mercken L, Snijders TAB, Steglich CEG, de Vries H (2009) Dynamics of adolescent friendship networks and smoking behavior: social network analyses in six European countries. Soc Sci Med 69:1506–1514

Norris JR (1997) Markov chains. Cambridge University Press, Cambridge

Ripley RM, Snijders TAB, Preciado P (2013) Manual for Siena version 4.0. Tech. rep. University of Oxford, Department of Statistics; Nuffield College, Oxford. http://www.stats.ox.ac.uk/siena/. Accessed 28 May 2013

Robbins H, Monro S (1951) A stochastic approximation method. Ann Math Stat 22(3):400–407

Schweinberger M (2012) Statistical modeling of network panel data: goodness-of-fit. Br J Stat Math Psychol 65:263–281

Selfhout-Van Zalk MHW, Burk W, Branje SJT, Denissen J, van Aken M, Meeus WHJ (2010) Emerging late adolescent friendship networks and big five personality traits: a social network approach. J Pers 78:509–538

Snijders TAB (1995) Methods for longitudinal social network data: review and Markov process models. In: Tiit E, Kollo T, Niemi H (eds) New trends in probability and statistics. Vol. 3: multivariate statistics and matrices in statistics. In: Proceedings of the 5th Tartu conference, TEV Vilnius, Lithuania, pp 211–227

Snijders TAB (1996) Stochastic actor-oriented dynamic network analysis. J Math Sociol 21:149–172

Snijders TAB (2001) The statistical evaluation of social network dynamics. In: Sobel ME, Becker MP (eds) Sociological methodology – 2001, vol 31. Basil Blackwell, Boston/London, pp 361–395

Snijders TAB (2005) Models for longitudinal network data. In: Carrington P, Scott J, Wasserman S (eds) Models and methods in social network analysis, Chapter 11. Cambridge University Press, New York, pp 215–247

Snijders TAB (2009) Longitudinal methods of network analysis. In: Meyers B (ed) Encyclopedia of complexity and system science. Springer, New York/London, pp 5998–6013

Snijders TAB, Baerveldt C (2003) A multilevel network study of the effects of delinquent behavior on friendship evolution. J Math Sociol 27:123–151

Snijders TAB, van Duijn MAJ (1997) Simulation for statistical inference in dynamic network models. In: Conte R, Hegselmann R, Terna P (eds) Simulating social phenomena. Springer, Berlin, pp 493–512

Snijders TAB, Steglich CEG, Schweinberger M (2007) Modeling the co-evolution of networks and behavior. In: van Montfort K, Oud H, Satorra A (eds) Longitudinal models in the behavioral and related sciences. Lawrence Erlbaum, Mahwah, pp 41–71

Snijders TAB, Koskinen JH, Schweinberger M (2010a) Maximum likelihood estimation for social network dynamics. Ann Appl Stat 4:567–588

Snijders TAB, van de Bunt GG, Steglich C (2010b) Introduction to actor-based models for network dynamics. Soc Netw 32:44–60

Snijders TAB, Lomi A, Torlò V (2013) A model for the multiplex dynamics of two-mode and one-mode networks, with an application to employment preference, friendship, and advice. Soc Netw 35:265–276

Steglich CEG, Snijders TAB, Pearson MA (2010) Dynamic networks and behavior: separating selection from influence. Sociol Methodol 40:329–393

van de Bunt GG (1999) Friends by choice; An actor-oriented statistical network model for friendship networks through time. Thesis Publishers, Amsterdam

van de Bunt GG, van Duijn MAJ, Snijders TAB (1999) Friendship networks through time: an actor-oriented statistical network model. Comput Math Organ Theory 5:167–192

Wasserman S (1979) A stochastic model for directed graphs with transition rates determined by reciprocity. In: Schuessler KF (ed) Sociological methodology 1980. Jossey-Bass, San Francisco

Wasserman S (1980) Analyzing social networks as stochastic processes. J Am Stat Assoc 75:280–294

Zeggelink EP (1994) Dynamics of structure: an individual oriented approach. Soc Netw 16:295–333

Recommended Reading

In addition to the help pages that are available as for all R packages, there is an extensive manual (Ripley et al. 2013) and a tutorial paper (Snijders et al. 2010b). A textbook about the *Siena* method and an edited volume with example applications are in preparation. The website http://www.stats.ox.ac.uk/siena/ is actively maintained and contains references to the basic methodology, references to applications, R scripts, example data sets, workshop announcements, and more.

For those who wish to read more about the mathematical and methodological background, a recommended sequence of readings could be Snijders (1996) as an introduction to the idea of stochastic actor-oriented models, Snijders (2001) or Snijders (2005) for the basic definition of the model for one dependent network defined as a changing digraph, and Steglich et al. (2010) for models for the dynamics of networks and behavior, which might be followed by Snijders et al. (2010a) for Maximum Likelihood estimation or Snijders et al. (2013) for models with multiple dependent networks.

Signatures

▸ Telecommunications Fraud Detection, Using Social Networks for

Signed Graphs

Krzysztof Stefaniak and Mikołaj Morzy
Institute of Computing Science, Poznań University of Technology, Poznań, Poland

Synonyms

Biased graph; Gain graph; Signed network

Glossary

Arc An ordered pair of nodes adjacent in the graph
Cycle A loop of at least three nodes in which the first and the last nodes are the same
Digraph A graph in which all relations are directed
Dyad A pair of nodes and the incidence relation between them
Edge A pair of nodes adjacent in the graph
Graph A data structure consisting of a set of entities called nodes and a set of pairs of nodes, called edges or arcs
Loop A walk in the graph in which all edges are distinct
Path A walk in the graph in which all edges and nodes are distinct
Sociomatrix Representation of the incidence relation as a two-dimensional matrix in which rows and columns represent nodes and cells represent relation values
Triad A triple of nodes and all incidence relations between them
Valence Semantic orientation of an edge in a signed graph

Definition

Given a set of nodes $N = \{n_1, \ldots, n_m\}$ and a set of edges $E = \{e_1, \ldots, e_n\}$, where each edge is a set of nodes, $e_k = \{n_i, n_j\}$. A *signed graph* is a triple $G_{\pm} = \langle N, E, S \rangle$ consisting of a set of nodes N, a set of edges E, and a mapping S which is a function $S : E \to \{+, -\}$, i.e., the mapping S associates with every edge $e_k \in E$ either a positive valence, typically denoted by $(+)$, or a negative valence, denoted by $(-)$. Positive valence of an edge usually denotes the fact that the relationship modeled by the edge (the type of association between nodes) has some positive quality, such as kindness, friendship, or trust. Likewise, the negative valence represents antagonizing feelings between nodes, such as enmity, dislike, or distrust. Edges can be lacking directional information, in such case the relationship is considered symmetrical. If edges are directional, such a graph is called a *signed digraph*. Some formulations also allow for the existence of multiedges as well as half-edges (which are edges with only one endpoint) and loose edges (which are edges without any endpoints), but half-edges and loose edges are not signed. A *complete signed graph* is a signed graph in which each unordered pair of nodes belongs to the set of edges.

Introduction

Signed graphs have been used for a long time in social network analysis to simultaneously model opposite relationships. In a signed graph, each edge is assigned either a positive or negative sign, referred to as *valence*. For instance, in a social network representing acquaintance between people, positive edges can represent friendship, while negative edges can represent animosity. If the signed graph is modeling diplomatic relations between countries, a positive edge can represent cooperation and a negative edge can represent some kind of political tension.

In general, edges can be attributed with more values, leading to the so-called *valued graphs*.

Signed graphs are a special case of valued graphs in which edges are allowed only two opposing values, and the aggregation of values along loops is performed by multiplication rather than by addition. It should be stressed that a negative edge between nodes is different from the lack of an edge between nodes. While the lack of an edge suggests the lack of interaction between nodes, a negative edge is a clear mark of an inimical relationship. Another frequent misunderstanding is that signed graphs are simply graphs with edges weighted by either $+1$ or -1 numerical values. Such a graph would be a regular graph with a constraint imposed on the set of possible values for edges. It would be very different from a signed graph because in regular valued graphs edge values are added and not multiplied. Another example where similar graphs are being used is the knot theory, where color is used to mark edges. Again, the methods and algorithms are very different because the color of an edge does not convey the intrinsic opposition of positive and negative valence.

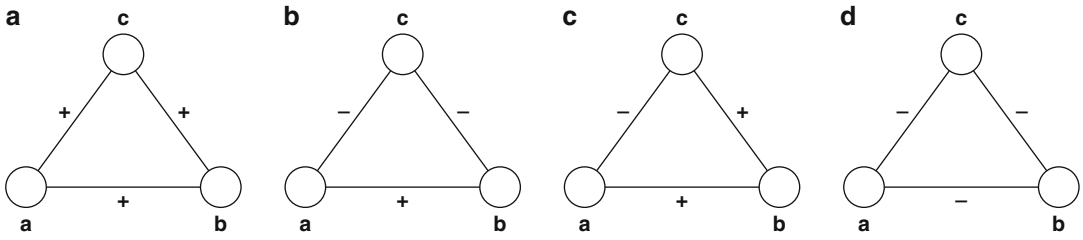
At the core of a signed graph lies a *signed relation*. It is the relation that can easily convey both positive and negative sentiments. Examples of signed relations include esteem/disesteem, like/dislike, praise/blame, and influence/negative influence, as presented by Sampson (1968). It is possible to treat these opposing sentiments as two independent relations, but in reality the two sentiments are clearly associated as one sentiment is usually an antonym of the other. Some graph theorists also require that the relation in question should satisfy the *principle of antithetical duality*, which is to say that the dual (the antonym) of a signed graph simply changes the signs of the loops. Computing the dual of the changed graph brings back the original graph. Without this property a graph cannot be used in the light of signed graph theory and balance theory. Therefore, traditional social networks, where relations usually represent some kind of social interaction, e.g., communication or interaction, cannot be modeled as signed graphs and cannot be studied using the balance theory.

Key Points

Signed graphs have several features that make them a useful tool for sociological and psychological research, but signed graphs can be also used outside of social sciences, e.g., in the field of physics or chemistry. To understand the benefit and utility of the signed graph model, we must first observe the key points that differentiate signed graphs from more general valued graphs. One of the most important methods of network analysis developed within the domain of signed graphs is the *triad analysis* that aims at capturing the dynamics of relations between very small groups of nodes. Triad analysis is described in detail in section "Triad Analysis." Triad analysis has been refined and extended in the field of social psychology under the moniker of *P-O-X triples analysis*, which we scrutinize in section "*P-O-X* Triples." Probably, the most famous concept originating from the signed graph theory is the idea of *structural balance*. In section "Structural Balance" we define the notion of structural balance and we introduce the fundamental Harary's theorem, along with its proof. We discuss the implications of a signed graph being balanced and we show a simple method for testing whether a graph is balanced. We also present several measures for the amount of imbalance. The last concept pertaining to signed graphs is the notion of *frustration*, discussed in section "Frustration."

Historical Background

Signed graphs have been studied since 1950s. They were first introduced by Cartwright and Harary (1956) and Harary (1953) in a structural balance theory – graph generalization of Heider's theory (1946) from sociology. Heider's theory of social balance can be described as a balance of sentiments between people, i.e., in subgroups of people certain relationships tend to be more socially plausible. For example, in group of two individuals (a dyad), there is only one relationship, positive or negative, but when we look at



Signed Graphs, Fig. 1 Possible signed triad configurations

a complete graph of relationships between three people (a triad), we can distinguish four different principles: "a friend of my friend is my friend," "an enemy of my enemy is my friend," "a friend of my friend is my enemy," and "an enemy of my enemy is my enemy," with the two latter ones clearly causing cognitive dissonance and thus making the whole graph unbalanced.

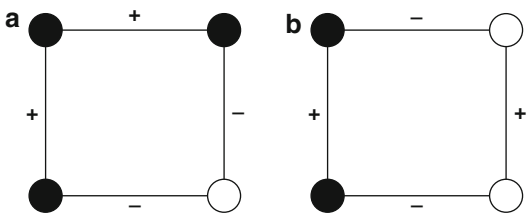
There is a tendency to avoid unbalanced structures and increase balance of the graph even if it makes sign shifts necessary. Sign changes include enemies becoming friends (positive edge) or friends becoming enemies (negative edge). According to these changes, Davis questioned the significance of the last principle (Davis 1967) arguing that it is rather difficult to make any of three mutual enemies friendly towards each other; thereupon, he proposed *weakly balanced graphs*, which rule out only the structure with one negative edge reflecting the principle "a friend of my friend is my enemy." In consonance with the original structural balance theory, balanced graph can be divided into two groups (bipartite graph) (Harary 1953), but in case of weakly balanced graphs, it is possible to have multiple clusters with positive edges inside the group and negative edges between the subgroups.

Signed Graphs

Triad Analysis

Most of the analysis of signed graphs is concerned with the analysis of dyads and triads. Each dyad can be in one of three states: a positive relationship, a negative relationship, and no relationship between the nodes in the dyad. For complete signed graphs each dyad is either positive or negative. For triads (for the sake of brevity, we are considering only complete triads) each triad can be in one of four states depending on the number of negative relationships between the nodes in the triad (zero, one, two, or three).

Consider possible triad configurations shown in Fig. 1. The configuration Fig. 1a is the simplest and most obvious, all actors have positive feelings about all other actors in the triad, so there is no room for a conflict. Similarly, the configuration Fig. 1b is stable, since actors a and b like each other and share the same negative feeling towards actor c. This configuration is stable in the sense that it is coherent and no actor has to choose between any other actor. Now compare previous configurations with the configuration presented in Fig. 1c. This configuration is unstable, because actor b is torn in his allegiance to actors a and c, who dislike each other. In order to maintain social ties, the actor b has to choose one of his friends, and the remaining relationship will probably become broken. Finally, the configuration presented in Fig. 1d is also considered unstable. What is characteristic about this configuration is that the "enemy of my enemy is my friend" rule of thumb does not apply here. This fundamental difference in triad configurations can be very easily expressed by the number of negative signs along the loop. Triads with an even number of negative edges tend to be stable, whereas triads with an odd number of negative edges tend to be unstable and eventually break down. This observation can be extended to loops of the length greater than 3, as depicted in Fig. 2. Experimental research suggests that this type of stability is quite often encountered in real networks because unstable configurations appear far less often in real networks than stable configurations.



Signed Graphs, Fig. 2 Stable configurations of loops of the length 4

P-O-X Triples

Another usage for signed graphs comes from the field of social psychology, where signed graphs were used to model the cognition of social relationships. A well-known example of this line of research was the analysis of the so-called *P-O-X* triples. According to this model, P denotes a person, O denotes another individual (the other), and X denotes an entity or object. The task is to find how the positive or negative attitude of the primary person P towards the object X is consistent with the attitude of the other O. This analysis is fairly similar to the discussion of the basic triple model presented above but with some slight differences that we will underline next.

To make our discussion as general as possible, we will assume that the relationships depicted in Fig. 3 represent the attitudes of liking (positive valence denoted with the (+) sign) and disliking (negative valence denoted with the (−) sign). Both P and O are allowed to express their attitudes towards object X; furthermore, they can express their sentiment towards each other. All relationships under discussion are assumed to be symmetrical. Scenarios Fig. 3a through Fig. 3d presented in the upper row depict balanced situations, where either both actors like each other and agree in their assessment of the object X (scenarios Fig. 3a, b), or both actors disagree in their assessment of the object X, but this difference in opinions can be explained by their mutual dislike (scenarios Fig. 3c, d). Now compare these to scenarios depicted in the lower row of the Fig. 3. Scenarios Fig. 3e, f represent the situation where the actors agree in their sentiment towards the object X despite having negative feelings about each other. Even more awkward situation is depicted in scenarios Fig. 3g, h, where actors P and O apparently like each other but cannot reach a consensus about the attitudes towards object X. Such disagreements, as shown by sociological research, can quickly undermine the general positive relationship between the actors.

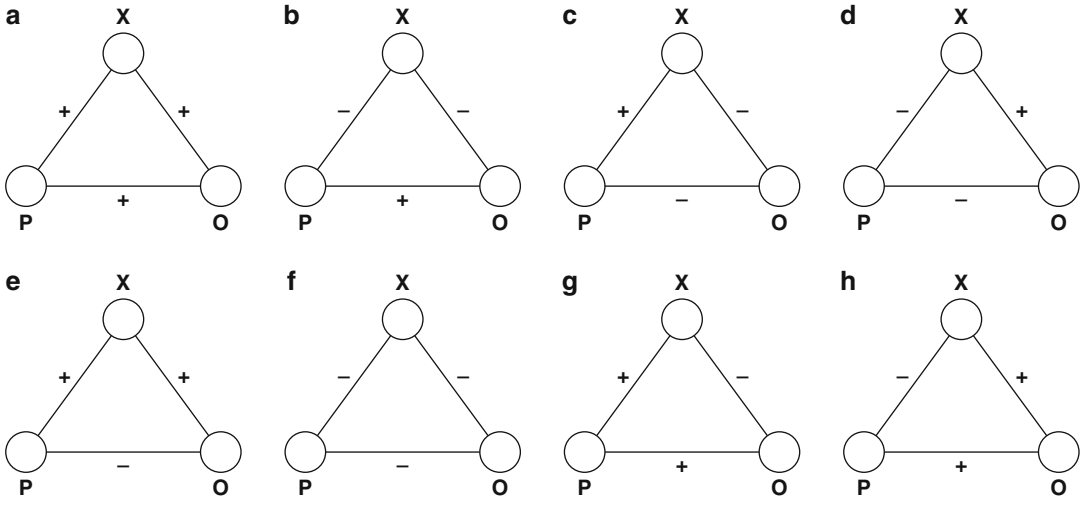
Structural Balance

Most of the analysis of signed graphs depends on the notion of loops. In particular, one is often interested in the sign of particular loops in the signed graph. We will use the term *loop* to describe any closed walk in the graph in which all nodes (except the first and the last) are distinct. The *sign of a loop* is the product of all edges contained in the loop. Since only negative edges change the sign of a loop and two negative edges cancel each out, an even number of negative edges on the loop will produce a positive loop, and an odd number of negative edges will produce a negative loop. The idea of a signed loop can be further extended to semicycles. A semicycle is a closed sequence of nodes in which every pair of consecutive nodes forming a semicycle is adjacent (in other words, a semicycle is a cycle in which arcs can point in any direction). The *sign of a semicycle* is defined also as a product of signs of arcs.

A signed loop is called *stable* if it contains an even number of edges. A graph is called *stable*, or, in other words, is said to show *structural balance*, if all loops in the graph are stable. Harary (1953) presents an important finding pertaining to signed graphs:

Harary's Theorem *A balanced graph can be divided into connected groups of nodes such that all connections between members of the same group are positive and all connections between members of different groups are negative.*

According to Harary, each group can contain an arbitrary number of nodes and there can be many groups of nodes. A graph is *clusterable* if its nodes can be divided into separate groups such that all positive relationships are happening only within the group and all relationships between groups are only negative. Harary's theorem states that all balanced graphs are indeed clusterable.



Signed Graphs, Fig. 3 *P-O-X* triples

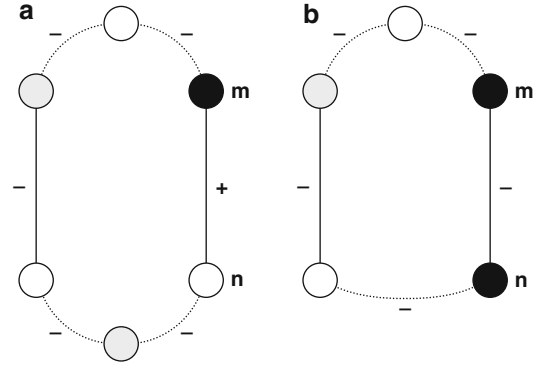
The opposite does not work, i.e., not all clusterable graphs are balanced. The term *structural balance* is used by sociologists and psychologists to refer to groups that are coherent and lack inner tensions between members.

Since structural balance is the key concept in many applications of the signed graph theory, we will provide a simple constructive proof of Harary's theorem. For the sake of simplicity, we will consider a graph with a single-connected component.

Proof We select a random node in the graph and we color this node with white color. Then, we iterate over all remaining uncolored nodes in the graph and we color them in according to two simple rules:

1. A node n connected by a positive edge to a node m that has already been colored receives the same color as m.
2. A node n connected by a negative edge to a node m that has already been colored receives the opposite color to m.

If, at any moment, we arrive at a node n that has already been colored, but according to the above rules it should be colored with an opposite color (i.e., a conflict arises), then the entire graph is not balanced. The reasoning behind this simple procedure is the following. While iterating over the nodes in the graph if we stumble upon a node n that has already been colored, this means that there must be an alternative path leading to the node n from the starting point. According to Harary's theorem, for the graph to be balanced, each loop in the graph has to have an even number of negative edges. Let us examine in detail the situation in which a conflict in coloring arises (see Fig. 4). There are only two such situations: Fig. 4a either we want to assign the node n the same color as the node m to which n is connected (i.e., the edge between m and n is positive) but n is already colored with an opposite color or



Signed Graphs, Fig. 4 Clustering of a balanced graph

Fig. 4b we want to assign the node n the opposite color as the node m to which n is connected (i.e., the edge between m and n is negative) but n is already colored with the same color as m. In the first situation the fact that n is colored differently from m means that there is a loop between m and n with an odd number of negative edges (because the color changes between m and n an odd number of times) and thus m and n should be placed in opposite groups. However, the existence of a direct positive edge between m and n contradicts this, thus, the graph is not balanced. A similar reasoning applies to the second situation. If n has the same color as m, then there is a loop in the network between m and n with an even number of negative edges (this is why color alternates an even number of times on the loop). But m and n cannot be placed in the same group because of the direct negative edge between them. Again, a contradiction proves that the graph cannot be balanced. The generalization of this proof to the graph consisting of several connected components is trivial since it requires simply to repeat the above procedure to all components sequentially. □

An interesting question arises of how to check efficiently if a given graph is balanced. Since a single loop with a negative sign makes the entire graph unbalanced, one needs to consider loops of length $l = 2, \ldots, n-1$ sequentially looking for a loop with a negative sign. In order to find the sign of a loop of a given length l, it is sufficient to check the main diagonal of the graph's sociomatrix raised to the power of l. If M is the sociomatrix of the signed graph $G_{\pm}$, then the main diagonal of M^l represents all loops of length l starting and ending at a given node. Consider a simple signed graph $G_{\pm}$ depicted in Fig. 5.

The sociomatrix M for the graph $G_{\pm}$ is given as:

M

	A	B	C	D	E
A	0	−1	1	0	0
B	−1	0	1	1	−1
C	1	1	0	0	1
D	0	1	0	0	0
E	0	−1	1	0	0

Below we show the powers of the sociomatrix M. Please observe the values on the main diagonal of the matrix as they contain the signs of all loops of the length $l = 2, 3, 4$ starting and ending in respective nodes.

As we can see, the main diagonal of each power of the sociomatrix M is nonnegative, thus we may conclude that the graph $G_{\pm}$ is balanced. The method presented here applies only to undirected signed graphs. The extension of the method to signed digraphs is not trivial. The interested reader will find the detailed description of the method in Harary et al. (1965, pp. 352–355).

One may ask if it is possible to somehow quantify the amount of imbalance in the graph. In other words, one may wonder how many changes would have to occur in order to make the graph balanced. Several indexes have been proposed that aim at measuring the degree of imbalance. A *cycle index for balance* has been used to collectively refer to such indexes. The general idea is to find the number of cycles in the graph that have a negative sign (i.e., the number of cycles that violate the balance condition) and to compare this number to the total number of cycles present in the graph. The simplest index (Cartwright and Gleason 1966) simply divides the number of

M^2

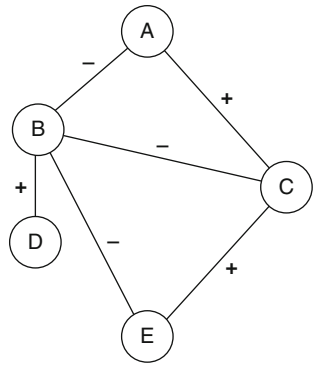
	A	B	C	D	E
A	2	−1	1	−1	2
B	−1	4	−2	0	−1
C	1	−2	3	−1	1
D	−1	0	−1	1	−1
E	2	−1	1	−1	2

M^3

	A	B	C	D	E
A	2	−6	5	−1	2
B	−6	4	−6	4	−6
C	5	−6	4	−2	5
D	−1	4	−2	0	−1
E	2	−6	5	−1	2

M^4

	A	B	C	D	E
A	11	−10	10	−6	11
B	−10	22	−16	4	−10
C	10	−16	16	−6	10
D	−6	4	−6	4	−6
E	11	−10	10	−6	11



Signed Graphs, Fig. 5 Checking for balance in the network

positive cycles by the total number of cycles in the graph. More elaborate indexing schemes propose to weight each cycle (positive or negative) by the length of the cycle (Harary 1959; Norman and Roberts 1972). Several other measures of structural balance are discussed in Taylor (1970).

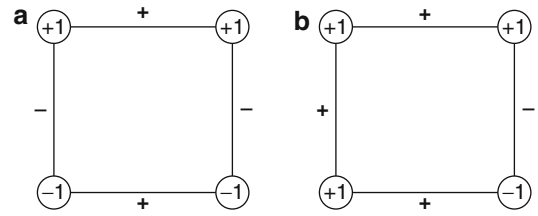
Frustration

Signed graphs can be extended by adding positive and negative values to nodes (in addition to edges). Let each node n has a state $S(n)$, where the state of a node can be either positive $(+1)$ or negative (-1). An edge e is said to be *satisfied* if and only if:

- Edge e is positive and both its endpoints are in the same state.
- Edge e is negative and both its endpoints are in opposite states.

If an edge is not satisfied, it is called *frustrated*. The minimum number of frustrated edges in any state of the graph is called the *line index for balance* (Harary 1960) or a *frustration index*.

The concept of frustration and frustration index is used in physics, in particular when modeling the ferromagnetism of spin glasses using Ising models (Barahona 1982). In a two-dimensional Ising model, a square lattice graph is used in the study of minimum energy configurations called *ground states*. In this graph each vertex, representing a molecule, can have a spin-up $(+1)$ or a spin-down (-1) orientation. Positive and negative edges correspond to



Signed Graphs, Fig. 6 Examples of (**a**) unfrustrated and (**b**) frustrated plaquettes

ferromagnetic and antiferromagnetic bonds between molecules, respectively. Elementary square segments of such lattice are called *plaquettes*; two examples are presented in Fig. 6.

If the number of negative bonds in a plaquette is even (Fig. 6a), then the perfect spin configuration that satisfies all the edges exists, and, therefore, there is no frustration in the plaquette. In other words, it is possible to set spins in such a configuration that the ground state has a minimum energy. On the other hand, an odd number of negative bonds (Fig. 6b) always causes a conflict, because to satisfy one bond, one has to break another bond instead, leaving at least one frustrated edge in the graph.

If an edge is frustrated, its molecules are unable to minimize their energy; thus, the whole graph gains an extra ground state degeneracy. For each configuration of molecule spins, there is a particular number of frustrated edges unable to minimize the energy. The total energy of the entire graph is proportional to the number of frustrated edges, and the preferred ground state of the minimum energy is the state with the lowest frustration index.

Unfortunately, finding the frustration index is computationally hard. Consider a signed graph $G_{\pm}$ consisting of only negative edges. Computing the frustration index of $G_{\pm}$ can be reduced to the maximum cut problem of a graph, which is known to be NP-hard.

Key Applications

The structural balance theory and signed graphs were initially invented to solve the subgrouping

problem in social psychology, but they found applications in other areas as well. Signed graphs are adequate to model opposite relations between objects. The first feasible entities are obviously humans with our complex psychological, social, and anthropological relations. Beyond that, it is possible to model natural phenomena as signed graphs too. Most notable adaptations can be found in chemistry (Trinajstic 1983) and physics (Mezard et al. 1987). Frustration index presented in section "Frustration" is the most significant adaptation of signed graphs in physics.

Nevertheless, social sciences are the major application areas for signed graphs. First of all they are used to describe dynamics of human sentiments. In addition to simply predicting and explaining friendship and animosity changes in groups of people (Antal et al. 2006), signed graphs proved to be helpful in anthropology and politics, i.e., structural balance was used in analysis of enmity in tribal wars, political conflicts (Hage and Harary 1983), or international relations (Harary 1961; Moore 1979).

On the other hand, current development of web-based social networks revealed new possibilities for signed graphs in social network analysis. Thanks to the global networks, social scientists have gained access to massive datasets. Recent research in this area include modeling trust and distrust (Guha et al. 2004), finding friends and foes (Brzozowski et al. 2008), community structure mining (Yang et al. 2007), or link prediction (Kunegis et al. 2009; Leskovec et al. 2010) in large signed networks. The last application is particularly interesting, since it makes recommendations of new acquaintances possible.

Cross-References

- Social Networks and Politics
- Structural Holes

References

Antal T, Krapivsky P, Redner S (2006) Social balance on networks: the dynamics of friendship and enmity. Dynamics on complex networks and applications. Phys D: Nonlinear Phenom 224(1–2):130–136

Barahona F (1982) On the computational complexity of ising spin glass models. J Phys A: Math Gen 15(10):3241

Bondy AJ, Murty USR (2002) Graph theory with applications. Wiley, New York

Brzozowski MJ, Hogg T, Szabo G (2008) Friends and foes: ideological social networking. In: Proceedings of the twenty-sixth annual SIGCHI conference on human factors in computing systems, CHI '08, Florence. ACM, New York, pp 817–820

Cartwright D, Gleason T (1966) The number of paths and cycles in a digraph. Psychometrika 31(2):179–199

Cartwright D, Harary F (1956) Structural balance: a generalization of Heider's theory. Psychol Rev 63(5): 277–293

Davis JA (1967) Clustering and structural balance in graphs. Hum Relat 20(2):181–187

Guha R, Kumar R, Raghavan P, Tomkins A (2004) Propagation of trust and distrust. In: Proceedings of the 13th international conference on world wide web, WWW'04, Manhattan. ACM, New York, pp 403–412

Hage P, Harary F (1983) Structural models in anthropology. Cambridge University Press, Cambridge/New York

Harary F (1953) On the notion of balance of a signed graph. Mich Math J 2(2):143–146

Harary F, Norman RZ, Dorwin C (1965) Structural models: an introduction to the theory of directed graphs. Wiley, New York, pp 352–355

Harary F (1959) On the measurement of structural balance. Behav Sci 4(4):316–323

Harary F (1960) A matrix criterion for structural balance. Nav Res Logis Q 7(2):195–199

Harary F (1961) A structural analysis of the situation in the Middle East in 1956. J Confl Resolut 5(2):167–178

Harary F (1969) Graph theory. Addison-Wesley, Reading

Heider F (1946) Attitudes and cognitive organization. J Psychol 21(2):107–112

Holland PW, Leinhardt S (1971) Transitivity in structural models of small groups. Small Group Res 2(2):107–124

Kunegis J, Lommatzsch A, Bauckhage C (2009) The slashdot zoo: mining a social network with negative edges. In: Proceedings of the 18th international conference on world wide web, WWW'09, Madrid. ACM, New York, pp 741–750

Leskovec J, Huttenlocher D, Kleinberg J (2010) Predicting positive and negative links in online social networks. In: Proceedings of the 19th international conference on world wide web, WWW'10, Raleigh. ACM, New York, pp 641–650

Mezard M, Parisi G, Virasoro MA (1987) Spin glass theory and beyond. World scientific lecture notes in physics, vol 9. World Scientific, Singapore
Moore M (1979) Structural balance and international relations. Eur J Soc Psychol 9(3):323–326
Newman M (2010) Networks: an introduction. Oxford University Press, New York
Norman R, Roberts F (1972) A derivation of a measure of relative balance for social structures and a characterization of extensive ratio systems. J Math Psychol 9(1):66–91
Sampson S (1968) A novitiate in a period of change: an experimental and case study of relationship. PhD thesis, Cornell University
Taylor HF (1970) Balance in small groups. Van Nostrand Reinhold Co., New York
Trinajstic N (1983) Chemical graph theory. CRC, Boca Raton
Wasserman S, Faust K (1994) Social network analysis: methods and applications, 1st edn. Structural analysis in the social sciences. Cambridge University Press, Cambridge/New York
Yang B, Cheung W, Liu J (2007) Community mining from signed social networks. IEEE Trans Knowl Data Eng 19(10):1333–1348
Zaslavsky T (1981) Characterizations of signed graphs. J Graph Theory 5(4):401–406
Zaslavsky T (1998) A mathematical bibliography of signed and gain graphs and allied areas. Electron J Comb 8

Recommended Reading

Signed graphs are covered thoroughly in the literature, both from the theoretic and application angles. A good starting point is a general book on graph theory, such as excellent text by Harary (1969) or Bondy and Murty (2002). An approach focusing more on the social aspects of networks is presented by famous textbooks by Wasserman and Faust (1994) and Newman (2010).

A very detailed summary of social balance research covering over 200 different papers is presented by Taylor (1970). Readers interested in a more anthropological approach to the study of social structure and balance should consult (Hage and Harary 1983). Our discussion on balance in social structures can be further extended to the notion of clusterability. Concepts of clusterability, ranked clusterability, and transitive tournaments are discussed at length by Holland and Leinhardt (1971).

If readers desire to investigate mathematical properties of signed graphs, they are advised to follow the work of Zaslavsky (1981). For the most comprehensive analysis of signed graphs literature, the reader is encouraged to study the bibliography compiled by Zaslavsky (1998).

Signed Network

▸ Signed Graphs

Similarity Metrics on Social Networks

Cuneyt Gurcan Akcora and Elena Ferrari
DISTA, Università degli Studi dell'Insubria, Varese, Italy

Synonyms

Categorical data similarity; Network similarity; Profile similarity

Glossary

Homophily Tendency to create friendships with similar people
Undirected Network Network where relationships are created by mutual consent of the two involved users
Profile Data User-uploaded text-based personal information on social networks

Definition

In the last decade, online social networks have gained millions of users who are daily creating terabytes of personal data (Ellison et al. 2007). With this big amount of data, it quickly becomes impractical to analyze all of the network for solving user-specific problems, such as finding

communities of users or classifying them according to a specific criteria. At the basis of most of these computations, there is the need of computing similarity between social network users. In this entry, we show how similarity computation can be done by using a family of metrics which provide fast, local, and efficient solutions to the question of computing user similarity on social networks. We classify user-generated social network data into network and profile data and discuss metrics for each type. We give particular importance to define what are the benefits and shortcomings of the considered metrics and how they are used in current research work.

Introduction

In the literature, the term *similarity* has been used in different meanings (e.g., the short distance between two users, shared features, or shared actions) to quantize similarity for different application fields. Some works have attempted to define similarity rigorously (Richter 2007; Ha and Haddawy 2003; Lin 1998). Among these, the four principles by Lin (1998) are widely used to implement similarity metrics on social networks. We will explain these four principles with commonality, differences, maximum similarity, and minimum similarity. In *commonality*, shared commonalities (e.g., race, gender, sex of users) increase the similarity of two users. On the other hand, more *differences* lead to smaller similarity. Regardless of the number of features, two users are said to have the *maximum similarity* when they are identical in every feature. Similarly, regardless of the number of features, two users are the least similar when they are different in every feature. In current studies, the maximum and minimum similarity values are given as 1 and 0, respectively.

A more rigorous set of properties for similarity metrics can be adopted from distance metrics by considering $similarity = 1 - distance$. These properties are (1) symmetry, (2) identity, (3) non-negativity, and (4) triangle equality. In the identity property, $distance(a, b) = 0$, when $a = b$. On social networks, this property can have different explanations; on the graph structure, $distance(a, b) = 0$ is assumed to be true when the two nodes have the same set of friends, whereas if profile information are considered, two users must have the same values for every profile item. In setting a lower bound for distance, the non-negativity property defines $distance(x, y) \geq 0$ for any user pair. The symmetry condition assures that $distance(a, b) = distance(b, a)$, while in the triangle equality $distance(a, c) \leq distance(a, b) + distance(b, c)$.

The absence of some of these properties can be used to classify different formulas. For example, *quasi-metrics* do not provide the symmetry property, whereas *semi-metrics* do not have the triangle property. Traditionally, the symmetry property is not applicable in directed networks, because directions of edges can lead to different similarity values for a pair of users. In this case, two different values are computed; $sim(a, b)$ denotes the similarity value according to a user a, and $sim(b, a)$ denotes a potentially different value from the perspective of user b.

Similarly, the triangle equality is difficult to achieve for similarity of user profiles because profiles can consist of more than one dimension (i.e., profile item). As a result, although the triangle property holds for one dimension, similarities for three user profiles with multiple dimensions might not adhere to the triangle property. Ideally, any formula that does not carry all these four properties should be called a measure, but researchers still prefer to use the word metric interchangeably with the term measure.

Founded upon these theoretical definitions, similarity metrics that have been proved efficient and practical on social networks are those that exploit a locality principle in similarity computations. The basic idea underlying such metrics is that, given two social network users, their similarity is computed by observing only a subset of vertices (e.g., friends of the two users) in the social network. This approach restricts required information about the social network to a minimum and reduces the required time of calculations. Even though global measures (e.g., the shortest path between users or the commu-

nity membership of users) can be used in the same context, they are more costly in time and computational power because they require too much information about the social network. For example, shortest path calculation might require observing friendship links of many users. Moreover, even though the costs can be undertaken, researchers and companies cannot have access to the whole social network data because of privacy issues. Only owners of social networking services can have the complete data that is required to compute global similarity measures. Therefore, local similarity metrics, which are the focus of this entry, provide a simple alternative in the face of these costly issues.

Historical Background

In the 1970s, early attempts at defining similarity metrics involved finding similarities among text-based documents (McGill 1979) that were modeled as a collection of words. Similarity metrics were used to discover relations between documents or rank the documents according to their similarity to a given query. These efforts have resulted in several well-known metrics, such as the cosine and Jaccard similarities (Han et al. 2006). With the advent of the Internet, researchers have applied document-based similarity metrics to user-generated web items, such as friendships and status posts, to discover relationships between web users. Specifically, similarity research on user-generated data has focused on predicting links (relationships) among users and mining past user behavior to predict future actions.

In the link prediction problem (Liben-Nowell and Kleinberg 2007), similarity of social network users has been exploited to predict new friendships. In this context, high similarity between two social network users is assumed to increase the probability of them creating a new friendship (Spertus et al. 2005). With generalization, this idea has been explored in the homophily theory which states that people tend to be friends with other people who are similar to them along personal attributes, such as gender, race, and religion (McPherson et al. 2001).

In addition to user characteristics, actions of a user are exploited to predict actions of similar users. This idea has been studied in recommender systems to observe existing item ratings (for movies, books, songs, etc.) of users and predict ratings of unseen items (Melville and Sindhwani 2010). Similar users are assumed to give similar ratings to similar items. From this assumption, similarity of ratings are predicted by finding either similar users (i.e., user based) or similar items (i.e., item based).

Methods

On social networks, user-generated data are classified into two types: profile data, which refers to user entered textual information, such as personal information, and network data that is information on created relationships, such as friendships with other users on the social network. Depending on the type of user data, similarity metrics differ in how they model a social network user. Furthermore, some similarity metrics can be used only on one type of user-generated data. By taking this into account, we will first explain how similarity metrics work on network data and then continue to explain metrics for profile data.

Network Similarity

In network similarity metrics, existing user relationships are exploited to find similarity. For example, a big number of shared friends between two users can be assumed to imply their high similarity. Network data that represent relationships can be modeled as a graph $\mathcal{G} = (V, E)$, where each user a in the network is considered a vertex $v_a \in V$, and a relationship between users a and b is an edge $e_{ab} \in E$ on the graph $\mathcal{G}$. If relationships are established by mutual consent of two users, an edge between them is said to be undirected (the edge has no start and end points), and it can be called a *friendship*. Friendship relations on social networks, such as Facebook, Orkut, and LinkedIn, are modeled with undirected graphs. On an undirected graph, first-level neighbors of a vertex v_a are a set of vertices $\Gamma(v_a)$ who share an edge with v_a (e.g., friends of a). Similarly, second-level

neighbors $\Gamma(\Gamma(v_a))$ (e.g., friends of friends of a) share an edge with first-level neighbors of a. If relationships can be created without mutual consent, an edge is said to be directed; it starts from the user who initiated the relationship (i.e., source vertex) and ends at the user with whom the relationship was established (i.e., target vertex). For example, the popular social networks Twitter and Google+ are directed social networks; when user a starts following user b, an edge is created, starting from vertex v_a and ending at vertex v_b. On directed graphs, the *friendship* term from undirected graphs is replaced with *in-neighbors* and *out-neighbors*. In-neighbors $\Gamma^-(v_a)$ and out-neighbors $\Gamma^+(v_a)$ of a vertex v_a are defined as target and source vertices of all edges that have vertex v_a as their source and target vertex, respectively.

Although the similarity metrics that we will discuss are designed to work with the *neighborhood* notion of undirected graphs, they can be applied to directed networks by small modifications. For example, direction of edges can be removed to make the graph undirected. Another approach is to consider only one type of edges (out-neighbors or in-neighbors) as neighbors while using the metrics. As these modifications can be used to define neighbors of a user on directed graphs, we will give metric definitions in terms of user neighbors. Assume that a similarity function $sim(v_a, v_b)$ computes the similarity of users v_a and v_b by considering their neighbors $\Gamma(v_a)$ and $\Gamma(v_b)$, respectively. We will denote one of their common neighbors with v_c, i.e., $v_c \in (\Gamma(v_a) \cap \Gamma(v_b))$. High similarity between two users who do not have a relationship will be assumed to increase the probability of them creating a relationship edge. With these definitions, metric formulas that we will explain are given in Table 1. Next we will discuss these network similarity metrics in more details.

Overlap: The overlap measure (Tan et al. 2006) counts the number of common neighbors of v_a and v_b to compute similarity.

Preferential attachment: For relationship creation on social networks, the preferential attachment metric reflects the "rich gets richer" notion from sociology (Barabási and Albert 1999). The metric assumes that highly connected vertices (i.e., users who have many neighbors) are more likely to create relationships with each other.

Jaccard ($\mathcal{L}1$ norm): The Jaccard metric (Tan et al. 2006) counts the number of common neighbors as in the overlap metric, but it normalizes this value by using the total number of neighbors of users v_a and v_b.

Cosine ($\mathcal{L}2$ norm): Cosine similarity was originally devised to find the similarity of two documents by computing the cosine of the angle between their feature vectors (Tan et al. 2006). When the angle between the two vectors is 0, they are considered identical, and the cosine of the angle equals the maximum similarity value 1.

Adamic and Adar: Like the overlap metric, Adamic and Adar considers common neighbors, but each neighbor's impact on the similarity value depends on the number of its neighbors (Adamic and Adar 2003). If a common neighbor has few neighbors, its impact on the similarity is assumed to be higher. For example, if users v_a and v_b are the only neighbors of v_c, $1/log(2)$ is added to the overall similarity. The similarity is computed by summing values from all common neighbors.

Point-wise mutual information (positive correlations): Point-wise mutual information (Bouma 2009) is computed in probabilistic terms where joint probability distribution function $P(\Gamma(v_a), \Gamma(v_b))$ computes the probability of a graph vertex $v_x \in V | x \neq a \neq b$ sharing edges with both v_a and v_b, whereas marginal probability distribution functions $P(\Gamma(v_a))$ and $P(\Gamma(v_b))$ are probabilities of a graph vertex sharing an edge with v_a and v_b, respectively. If edges are assumed to represent friendships, $P(\Gamma(v_a), \Gamma(v_b))$ is equal to $\frac{\#mutual\ friends}{\#users\ in\ the\ social\ network}$. Similarly, $P(\Gamma(v_a))$ is equal to $\frac{\#friends\ of\ v_a}{\#users\ in\ the\ social\ network}$. In other words, point-wise mutual information shows whether two users share mutual friends due to randomness. Note that due to computing probability with the total number of users in the social network, point-wise mutual information

Similarity Metrics on Social Networks, Table 1 Network similarity metrics

Measure	Formula	Description
Overlap	$\lvert\Gamma(v_a) \cap \Gamma(v_b)\rvert$	The number of common neighbors
Preferential attachment	$\lvert\Gamma(v_a)\rvert \times \lvert\Gamma(v_b)\rvert$	Multiplied neighbor counts of both users
Jaccard	$(\lvert\Gamma(v_a) \cap \Gamma(v_b)\rvert)/(\lvert\Gamma(v_a) \cup \Gamma(v_b)\rvert)$	The percentage of shared neighbors over all neighbors
Cosine	$(\lvert\Gamma(v_a) \cap \Gamma(v_b)\rvert)/(\sqrt{(\lvert\Gamma(v_a)\rvert \times \lvert\Gamma(v_b)\rvert)})$	The number of common neighbors normalized by multiplied neighbor counts
Adamic and Adar	$\sum_{v_c \in \{\Gamma(v_a) \cap \Gamma(v_b)\}} \frac{1}{Log(\lvert\Gamma(v_c)\rvert)}$	Common neighbors who have very few neighbors are given more importance
Point-wise mutual information	$P(\Gamma(v_a), \Gamma(v_b)) \times log\left(\frac{P(\Gamma(v_a), \Gamma(v_b))}{P(\Gamma(v_a)) \times P(\Gamma(v_b))}\right)$	How much the probability of having the current set of common neighbors differs from the case where neighbors would be added by users on the graph randomly
Katz's measure	$\sum_{p=1}^{+\infty} \beta^p \times NumOfPath(v_a, v_b, p)$	Similarity is implied by the number and length of paths that connect two users on the graph. Each path $0 < p < +\infty$, whereas $0 < \beta < 1$

produces very low average similarity values, because *#users in the social network* can be in millions.

Katz's measure: Katz's measure (1953) was designed in the 1950s to find the status of a vertex on a graph. The vertex which had the biggest number of shortest paths to the other vertices was considered a central vertex with high status. To compute $sim(v_a, v_b)$, Katz's measure finds the number of paths that connect v_a and v_b for path length p, $1 < p < +\infty$. The number of paths (i.e., the value of $NumOfPath(v_a, v_b, p)$) is dampened by a β^p value, where $0 < \beta < 1$. In practice, the β value is chosen as small as 0.005 (Liben-Nowell and Kleinberg 2007). Although p values can be increased to cover a big portion of the graph, usually $p = 2$ or $p = 3$ values are chosen to find similarity, because computations for $p > 3$ contribute very less to the overall value. Note that although the Katz's measure can be used as a global measure with big p values, small p values (e.g., 2 or 3) make it a practical measure for fast, local similarity computations.

In research work, performance of similarity metrics has been compared by making predictions based on similarity values and validating the results (Liben-Nowell and Kleinberg 2007; Spertus et al. 2005). Typically, metrics are used to predict top-k relationships (e.g., k most probable future friendships that will be created between users) on graphs at a time t_1, and these predictions are validated at a time $t_2 > t_1$. The performance of a metric can be computed by counting the number of correct predictions.

In Liben-Nowell and Kleinberg (2007), Adamic and Adar has been shown to perform better than preferential attachment, Jaccard, overlap, and Katz's measure on a scientific coauthorship network. On another social network, Orkut.com, Spertus et al. (2005) have found that cosine similarity performed better than Jaccard and point-wise mutual information metrics.

Despite these comparisons, it is important to understand that each metric has its weaknesses in different application fields. Preferential attachment is widely used in social networks to predict friendships, but unlike Jaccard or cosine similarity, its computed value does not reside within [0,1]. In fact, its max value is only bounded by the total number of users in a social network, because there are no theoretical limits to prevent a user from having every other social network user as a neighbor. However, some social networks may choose to limit the number of neighbors;

for example, an undirected network, Facebook, allows up to 5,000 neighbors, whereas a directed network, Twitter, does not have such a limit. Because of this, preferential attachment cannot be used to quantify how much *percent* two users are similar. When the graph has many vertices (i.e., the probability of an edge between two users are very small), computed value of point-wise mutual information can be very small, and it cannot be used to define how much *percent* two users are similar. Point-wise mutual information and preferential attachment can be best used in ranking a set of users according to their similarity to a specific user. There are some limitations in using Katz's measure too. In popular social networks, discovering edge counts for paths of length 2 or more can be restricted because social networking services do not allow access to social network data. Furthermore, Katz's measure can be costly to compute when the social network is large. Considering these limitations, research work (Jin et al. 2011; Zheleva et al. 2010) have mostly used Jaccard, cosine, or Adamic and Adar in their user similarity computations, because these measures are fast and easier to interpret.

Profile Similarity

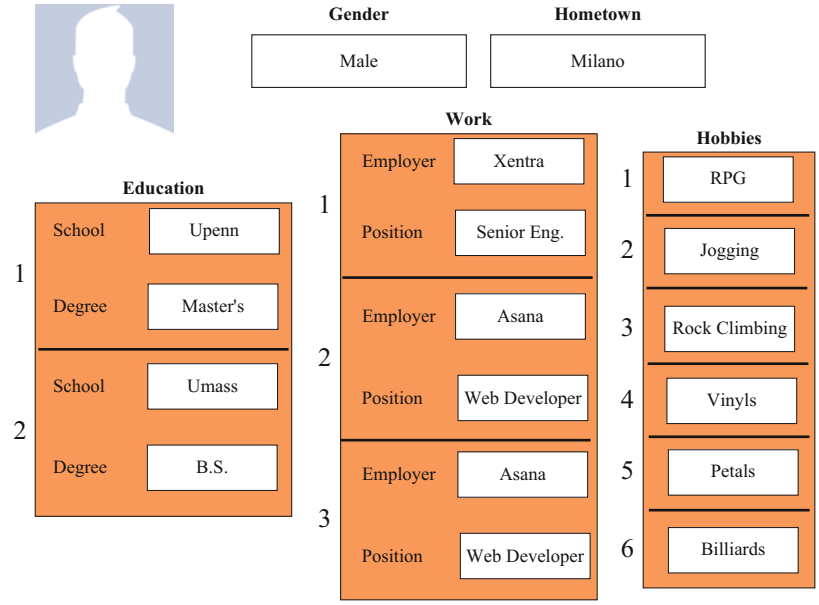
Along with network data, profile data constitute the second type of user-generated data on social networks. We will call similarity metrics which work with profile data as profile similarity metrics. On social networks, we will consider profile information as a set of unique items (e.g., hometown, location, education of a user), which can have one or multiple subfields for each value. Figure 1 shows an example of user profile with two education values, where an education value can have more than one subfields (i.e., school and degree). The figure also shows some items which can have only one value, such as gender:male.

After modeling profile data with a set of items, similarity between two users is computed by first finding the similarity of individual item values on the two profiles. For example, similarity of two users according to the gender item compares gender values on the two profiles. If the considered item has many values, or each value has multiple subfields (e.g., the education item on Fig. 1), an aggregation function is required to find item similarity. An overall profile similarity value is determined by aggregating (e.g., by weight averaging) all item similarity values. However in practice, most of the research work on profile similarity consider a user profile to be a set of unstructured keywords. For example, in Bhattacharyya et al. (2010), Facebook user profiles only consist of user values from the "hobbies" item. With this simplification, profile similarity of two users is found by computing item similarity of hobbies on two profiles.

In a more detailed study, similarity of items which have multiple values or subfields has been weight averaged to model the importance of similarity for some items or subfields (Akcora et al. 2011). For example, when user profiles consist of hometown and hobbies fields, hometown similarity can be weighted with a bigger coefficient to show that hometown similarity is more important than hobbies similarity.

When similarity computations are reduced to finding item similarities, the type of item values determines the way similarity is computed. Although some items have numerical values (e.g., age, zip code), most of the item values on social networks are text-based categorical data which cannot be ordered on an axis to find similarity/distance of two data points. Using simple approaches, such as string matching, is not efficient because the text represents an identity (e.g., hometown:Barcelona) and partial (n-gram) similarities (e.g., b**AR**celona:p**AR**is) are trivial.

Two main approaches are used to find similarity of item values: ontology based (Jung and Euzenat 2007; Mika 2005) and social graph based (Akcora et al. 2011). In ontology-based approaches, a graph of entities is created to define their relationships or distance (Cristani and Cuel 2005). For example, considering hometown similarity of three social network users with values Barcelona, Madrid, and New York, an ontology can classify Barcelona and Madrid as Spanish cities, whereas New York is classified as an American city, and compute a higher similarity for users from Barcelona and Madrid. The main disadvantage of this approach is that it requires a reliable ontology which can be difficult



Similarity Metrics on Social Networks, Fig. 1 An exemplary profile for a social network user. The profile consists of single-valued gender and hometown items, as well as multiple-valued education and work items

to create. Furthermore, as social networks are dynamic, new item values are added in time and the ontology must be updated frequently.

The social graph-based approach assumes that neighbors (e.g., friends/coauthors) of a user v_a are similar to v_a along profile attributes. For example, if hometown of v_a is Barcelona, we can expect many of its neighbors to be from Barcelona and other Spanish cities. When hometown values of neighbors are observed, another city (e.g., Madrid) can be found similar to Barcelona without explicitly creating an ontology. With this intuition, a user v_b is said to be similar to v_a if v_b's hometown is similar to the hometown values of v_a's neighbors. Furthermore, even when v_a has a blank profile, using its neighbors in such a way allows one to compute its similarity with other social network users. The social graph approach has also been found effective for network similarity metrics (Cukierski et al. 2011). The disadvantage of this approach is that neighbors are assumed to be similar to users. This assumption is more applicable in undirected social networks where mutual consent is required to create a relationship edge. However, if the network is undirected, neighbors can have very different characteristics from users, and performance of social graph-based approaches can deteriorate.

As ontology-based approaches have been extensively studied in semantic web communities (Cristani and Cuel 2005), in the rest of this section, we will detail social graph-based approaches. To this end, we will discuss relevant categorical data similarity measures (Boriah et al. 2008) which can be used with the assumptions of social graph-based approaches. Before doing that, we need to introduce some notations and definitions.

Assume that for an item i, we are given a pair of item values i_a and i_b from profiles of v_a and v_b, respectively.

From the set of v_a's neighbors, we create a collection of values $values(i) = \{\forall i_c | v_c \in \Gamma(v_a)\}$. We will define three functions over $values(i)$. Function $distinct(i)$ finds the number of distinct values in $values(i)$, and $sup(i_x)$ finds the count of value i_x in $values(i)$, whereas $freq(i_x) = sup(i_x)/|\Gamma(v_a)|$. With these functions, we will explain the categorical similarity functions reported in Table 2. Variations of Lin and Eskin measures are excluded for brevity.

Similarity Metrics on Social Networks, Table 2 Item similarity measures with their formulas. The two items are identical when $i_a = i_b$. Each measure uses a different formula for identical and nonidentical values pairs

Measure	Formula
Eskin	1, if $i_a = i_b$ $\frac{distinct(i)^2}{(2+distinct(i)^2)}$, if $i_a \neq i_b$
Occurrence frequency	1, if $i_a = i_b$ $\frac{1}{1+(log(\frac{\lvert\Gamma(v_a)\rvert}{sup(i_a)})\times log(\frac{\lvert\Gamma(v_a)\rvert}{sup(i_b)}))}$, if $i_a \neq i_b$
Lin	$\frac{2\times log(freq(i_a))}{log(freq(i_a))+log(freq(i_b))}$, if $i_a = i_b$ $\frac{2\times log(freq(i_a)+freq(i_b))}{log(freq(i_a))+log(freq(i_b))}$, if $i_a \neq i_b$

Eskin: Eskin's measure (Boriah et al. 2008) assigns 1 in identical cases ($i_a = i_b$), and it penalizes users when their values do not match while there are very few distinct values in $values(i)$. For example, it punishes users more for mismatches in the gender item (2 values with male and female) than it does in the hometown item because hometown can have many more values.

Occurrence frequency: The occurrence frequency measure assigns 1 to identical value pairs, and it favors mismatches with highly frequent values. If $values(i)$ has two distinct values i_x and i_y, with $i_x = i_a$ and $i_y = i_b$, $sim(i_a, i_b)$ reaches its maximum value.

Lin: Unlike others, Lin's measure (1998) does not assign 1 to two identical item values. Instead, it assigns high similarity when the two values are highly frequent in $values(i)$. In other words, if item value of v_b is very frequent among friends of v_a, v_b is considered very similar to v_a. For mismatches (i.e., $i_a \neq i_b$), the measure gives less importance to infrequent values.

A comparative evaluation of these measures have been carried out in Akcora et al. (2011), where the occurrence frequency has been found superior in performance. Overall, choosing social graph-based approaches over ontology-based approaches improves profile similarity results because social graph-based approaches can use profiles of neighbors to infer blank profile items of a user. By doing so, scarcity of data on user profiles can be compensated, and similarity can be computed for more social network users. This gives an edge to social graph-based approaches, because analysis of real-life social networks indicates that a big portion of user profiles (up to 60 % for a popular network, Facebook.com (Akcora et al. 2011)) are indeed missing.

Comparing Network and Profile Data for Similarity Purposes

A comparison between metrics for profile and network data can be done according to two dimensions: interpretation and availability. In interpreting results, profile data is richer than the network data because it covers more relations between users, and similarity values can be interpreted in terms of items. For example, high similarity between two users can be pointed to their common hometown, education, or religion values. On the other hand, network similarity offers only graph edges as its data, and network similarity results can only be interpreted in terms of being connected on a graph. For example, in Jaccard similarity, two users' similarity can be due to many shared friends, but the metric cannot explain why they share these friends at the first place. In such a case, profile similarity could point out that common friends are due to the shared hometown values.

In availability, network similarity is easier to use because network edges are structured and easier to discover. In comparison, profile data is more scarce and polluted; users might not enter any profile data, or any data they enter might be unstructured. Profile data is also more difficult to find in research data sets because of privacy issues.

Future Research

So far, usage of similarity metrics has been confined to well-studied problems such as link prediction and malicious clone detection on social networks. Recently, some work have used similarity metrics as an auxiliary method in link

prediction where users have not yet generated any actions on social networks (i.e., in the face of cold start) (Leroy et al. 2010) or to predict the risk of interacting with a user in terms of disclosure of personal information (Akcora et al. 2012).

Despite these work, similarity metrics are still used as black box models without considering their descriptive powers. If this aspect is dully considered, similarity metrics can be used to explain *why* users are similar and what types of users interact with each other. In this vein, we expect similarity metrics to be used in understanding phenomena such as homophily (McPherson et al. 2001) on a global scale. With this approach, the problem of predicting a link among users can be broaden to the issue of predicting links among general types of users, and user interactions models can be found by aggregating similarity among users.

Cross-References

- ▸ Centrality Measures
- ▸ Components of the Network Around an Actor
- ▸ Link Prediction: A Primer

References

Adamic L, Adar E (2003) Friends and neighbors on the web. Soc Netw 25(3):211–230

Akcora C, Carminati B, Ferrari E (2011) Network and profile based measures for user similarities on social networks. In: 2011 IEEE international conference on information reuse and integration (IRI), Las Vegas. IEEE, pp 292–298

Akcora C, Carminati B, Ferrari E (2012) Privacy in social networks: how risky is your social graph? In: 2012 IEEE 28th international conference on data engineering (ICDE), Washington, DC. IEEE, pp 9–19

Anderson A, Huttenlocher D, Kleinberg J, Leskovec J (2012) Effects of user similarity in social media. In: Proceedings of the fifth ACM international conference on web search and data mining, Seattle. ACM, pp 703–712

Barabási A, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Bhattacharyya P, Garg A, Wu S (2010) Analysis of user keyword similarity in online social networks. Soc Netw Anal Min 1:1–16

Boriah S, Chandola V, Kumar V (2008) Similarity measures for categorical data: a comparative evaluation. SIAM 30(2):243–254

Bouma G (2009) Normalized (pointwise) mutual information in collocation extraction. In: Proceedings of GSCL conference, Potsdam, Germany, pp 31–40

Cristani M, Cuel R (2005) A survey on ontology creation methodologies. Int J Semant Web Inf Syst 1(2): 49–69

Cukierski W, Hamner B, Yang B (2011) Graph-based features for supervised link prediction. In: The 2011 international joint conference on neural networks (IJCNN), San Jose. IEEE, pp 1237–1244

De Meo P, Ferrara E, Fiumara G (2011) Finding similar users in facebook. In: Social networking and community behavior modeling: qualitative and quantitative measurement, Igi Publishing. Hershey, Pennsylvania (USA). vol 4, pp 1–26

Ellison N et al (2007) Social network sites: definition, history, and scholarship. J Comput-Mediat Commun 13(1):210–230

Ha V, Haddawy P (2003) Similarity of personal preferences: theoretical foundations and empirical analysis. Artif Intell 146(2):149–173

Han J, Kamber M, Pei J (2006) Data mining: concepts and techniques. Morgan kaufmann, San Francisco

Huang Z, Li X, Chen H (2005) Link prediction approach to collaborative filtering. In: Proceedings of the 5th ACM/IEEE-CS joint conference on digital libraries, Denver. ACM, pp 141–142

Jin L, Takabi H, Joshi J (2011) Towards active detection of identity clone attacks on online social networks. In: Proceedings of the first ACM conference on data and application security and privacy, San Antonio. ACM, pp 27–38

Jung J, Euzenat J (2007) Towards semantic social networks. Semant Web: Res Appl 1:267–280

Katz L (1953) A new status index derived from sociometric analysis. Psychometrika 18(1):39–43

Leroy V, Cambazoglu B, Bonchi F (2010) Cold start link prediction. In: Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining, Washington, DC. ACM, pp 393–402

Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. J Am Soc Inf Sci Technol 58(7):1019–1031

Lin D (1998) An information-theoretic definition of similarity. In: Proceedings of the 15th international conference on machine learning, San Francisco, vol 1, pp 296–304

McGill M (1979) An evaluation of factors affecting document ranking by information retrieval systems. Syracuse Univ., NY. School of Information Studies.

McPherson M, Smith-Lovin L, Cook J (2001) Birds of a feather: homophily in social networks. Ann Rev Sociol 27:415–444

Melville P, Sindhwani V (2010) Recommender systems. In: Encyclopedia of machine learning, Springer

Science+Business Media. LLC 2011. New York, vol 1, pp 829–837
Mika P (2005) Ontologies are us: a unified model of social networks and semantics. In: The semantic web–ISWC 2005, Galway, vol 4, pp 522–536
Richter M (2007) Foundations of similarity and utility. In: The 20th international FLAIRS conference, Key West
Spertus E, Sahami M, Buyukkokten O (2005) Evaluating similarity measures: a large-scale study in the orkut social network. In: Proceedings of the 11th SIGKDD, Chicago. ACM, pp 678–684
Tan P, Steinbach M, Kumar V et al (2006) Introduction to data mining. Pearson/Addison Wesley, Boston
Zheleva E, Getoor L, Golbeck J, Kuter U (2010) Using friendship ties and family circles for link prediction. Advances in social network mining and analysis, Springer Berlin Heidelberg. pp 97–113

Recommended Reading

Birds of a feather: Homophily in social networks (McPherson et al. 2001)
Effects of user similarity in social media (Anderson et al. 2012)
Evaluating similarity measures: a large-scale study in the orkut social network (Spertus et al. 2005)
Finding similar users in facebook (De Meo et al. 2011)
Link prediction approach to collaborative filtering (Huang et al. 2005)

Simple Walk

▸ Semirings and Matrix Analysis of Networks

Simulated Datasets

Fahimah Al-Awadhi
Department of Statistics and Operations Research, Kuwait University, Kuwait City, Kuwait

Synonyms

Gibbs sampler; Markov chain Monte carlo algorithms; Metropolis Hastings; Monte Carlo methods; Statistical simulation

Glossary

MC Monte Carlo
MCM Monte Carlo methods
MCS Monte Carlo simulation
MCMC Markov chain monte carlo
IID Independent identically distributed
MHA Metropolis-hastings algorithm
GS Gibbs sampler

Introduction

Simulation is the imitation of the operation of a real-world process or system over time. Simulation has appeared at the very early stages of the development of statistics as a field. Francis Galton invented mechanical devices in 1873 to compute estimators and distributions by means of simulation. His well-known quincunx (Stigler 1986) is a derivation of the Central Limit Theorem for Bernoulli experiments. The randomized experiments of Ronald Fisher (1935) and the bootstrap revolution started by Brad Efron (Efron and Tibshirani 1993) are intrinsically connected with calculator and computer simulations, respectively.

Simulations were used to test a previously understood deterministic problem that has no random variables and no degree of randomness. Statistical sampling was used to estimate uncertainties in the simulations. Monte Carlo simulation (MCS) inverts this approach, solving deterministic problems using a probabilistic analog. An early variant of the Monte Carlo Methods (MCM) can be seen in Buffon's needle experiment, in which π can be estimated by dropping needles on a floor made of parallel strips of wood.

In the 1930s, Enrico Fermi first experimented the MC methods while studying neutron diffusion. In the early 1940s, it was applied in research into nuclear fission. The scientists working on the Manhattan project, making the atomic bomb, had intractably difficult equations to solve in order to calculate the probability with which a neutron from one fissioning uranium atom would cause another to fission. The equations were complicated because they had to mirror the com-

plicated geometry of the actual atomic bomb. The answer had to be right because, if the first test failed, it would be months before there was enough uranium for another attempt. Despite having most of the necessary data, researchers were unable to solve the uncertainty problems using conventional, deterministic methods.

Stanislaw Ulam suggested MCM for evaluating complicated mathematical integrals that arise in the theory of nuclear chain reactions. Von Neumann carried this suggestion to the more systematic development of MC. with the primitive facilities available at the time, Ulam and von Neumann did carry out numerical computations that led to a satisfactory design.

In the 1950s MCM were used at Los Alamos for early work relating to the development of the hydrogen bomb and became popularized in the fields of physics, physical chemistry, and operations research. The RAND Corporation and the US Air Force were two of the major organizations responsible for funding and disseminating information on MCM during this time, and it began to find large applications in many different fields.

Widespread Applications of Simulation

MCM are especially useful for simulating phenomena with significant uncertainty in inputs and systems with a large number of coupled degrees of freedom. Areas of application include:

- Statistics: MCM are generally used for comparing competing statistics for small samples under realistic data conditions and are also used to provide implementations in various fields such as image analysis, signal processing, point processes, econometrics, and surveys.
- Mathematics: To evaluate multidimensional definite integrals with complicated boundary conditions, it is an alternative for the deterministic numerical integration algorithms.
- Physical sciences: MCM are used in computational physics, physical chemistry, quantum systems, and related applied fields. MC molecular modeling is an alternative to computational molecular dynamics.
- Astrophysics: MCM are used in the ensemble models that form the basis of modern weather forecasting. They are also used to model both the evolution of galaxies and the transmission of microwave radiation through a rough planetary surface.
- Engineering: MCM are used for sensitivity analysis and quantitative probabilistic analysis in process design. For example, MCM are applied to analyze correlated and uncorrelated variations in analog and digital integrated circuits in microelectronics engineering.
- Geostatistics: MCM underpin the design of mineral processing flow sheets and contribute to quantitative risk analysis.
- Computational biology: MCM are used in Bayesian inference in phylogeny.
- Finance: To calculate the value of companies and to evaluate investments in projects at a business unit or corporate level, they are also used to calculate the risk and to evaluate financial derivatives and to model project schedules.

Simulation Techniques

Let $X = (X_1, \ldots, X_n)$ be IID random variable defined on a suitable sample space $\mathbb{E}$, and assume that each X_i has a known density function $\pi_{X_i}(x_i)$ defined on $\mathbb{E}$. In many problems, X is high dimensional and evaluation of a function $g(X)$ using π_X is a challenging problem. The function under interest will be given by

$$E_{\pi_X}(g(x)) = \int_{x \in \mathbb{E}} g(x)\pi_x dx.$$

Analytical calculation of the above integral is not possible because of the complexity of the distribution function π_{X_i}. Simulation inference can be used instead. Suppose, for example, that we have a way to obtain independent samples $x^{(1)}, x^{(2)}, \ldots$ from π_X, we could then approximate the expectation of $g(X)$ by the *empirical estimate*

$$\overline{g}_n = \frac{1}{n}\sum_{i=1}^{n} g(x^{(i)}),.$$

By the strong law of large numbers as $n \to \infty$,

$$\overline{g}_n \xrightarrow{a.s.} E_{\pi_X}(g).$$

So, if we were able to find an explicit, easily codable function $h(U_1, U_2, \cdots)$ of uniform $(0; 1)$ variates with values in $\mathbb{E}$ and probability distribution the same as π_X, then we would evaluate the desired integral as

$$\frac{\sum_{j=1}^{n} g(h(U_{j1}, U_{j2}))}{n},$$

where U_{jk} is a double indexed array of IID uniform variables. It is allowed for h to depend on unboundedly many U variables, as long as the number of such variables required is a random variable with finite expectation. For example, in a one-dimensional integral, we would use random numbers to select points $\{x_i, i = 1, \cdots, n\}$ in the interval $a \le x \le b$ and then use the approximation

$$\int_a^b \pi(x)dx \approx \frac{(b-a)\sum_{i=1}^{n} \pi(x_i)}{n}.$$

MCM all follow a similar pattern:

1. Define some domain of inputs $\mathbb{E}$. This just means we have some set of variables in the model and we want to know the range of the values they can take on.
2. Generate inputs randomly, governed by some probability distribution π.
3. Perform some computation on these inputs.
4. Repeat 2 and 3 over and over a very large number of times.
5. Aggregate the results from the previous step into some final computation.

The result is an approximation to some true but unknown quantity.

Examples

Here are some simple examples on MCS.

Example 1 A simple MCS to approximate the value of π could involve randomly selecting points $(x_i, y_i), i = 1, \cdots, n$ in the unit square and determining the ratio $\rho = \frac{m}{n}$, where m is number of points that satisfy $x_i^2 + y_i^2 \le 1$. Consider a circle inscribed in a unit square. Given that the circle and the square have a ratio of areas that is $\frac{\pi}{4}$, the value of π can be approximated using an MCM as follows:

1. Draw a square, then inscribe a circle within it.
2. Uniformly scatter some objects of uniform size, over the square.
3. Count the number of objects inside the circle and the total number of objects.
4. The ratio of the 2 counts is an estimate of the ratio of the 2 areas, which is $\frac{\pi}{4}$. Multiply the result by 4 to estimate π.

In a typical simulation of $n = 1{,}000$ sample size there were 787 points satisfying. Using this data, we obtain $\rho = \frac{787}{1{,}000} = 0.787$ and

$$\pi \approx \rho \times 4 = 0.0787 * 4 = 3.418.$$

If points are purposefully dropped into only the center of the circle, they are not uniformly distributed, so our approximation is poor. The approximation is generally poor if only a few points are randomly dropped into the whole square. The approximation improves as more points are dropped.

Example 2 Suppose we want to find out the probability that, out of a group of 50 people, 2 people or more people share birthdays. The probability of having at least 2 people in the group having the same birthday is equal to $1 - \frac{365!}{365^{50}(365-50)!} = 0.970$, where ! is the factorial operator. Using the MC approach:

1. Pick 50 random numbers in the range [1,365]. Each number represents 1 day of the year.
2. Check to see if any of the 50 are equal.
3. Go back to step 1 and repeat 1,000 times.
4. Report the fraction of trials that have matching birthdays.

Using 1,000 iterations, the probability is approximately 0.864. Obviously, the more times we repeat the experiment, the more precise our result would be. Better than repeating the experiment a 1,000 times, we can easily use a computer to simulate the experiment 10,000 times (or more). Using 100,000 iterations, the simulated result was 0.969.

Example 3 Consider calculating the probability of a particular sum of the throw of two dice. There are 36 combinations of dice rolls. We can manually compute the probability of a particular outcome. For example, there are six different ways that the dice could sum to seven. Hence, the probability of rolling 7 is equal to 6 divided by $36 = 0.167$. Using MC approximation:

1. Throw the two dice and record the sum of the output.
2. Go back to step 1 and repeat 10,000 times.
3. Report the fraction of trials for each of the 11 different sum.

If the dice totaled 1,813 times out of 10,000 rolls, we would conclude that the probability of rolling 7 is approximately 0.1813.

The accuracy of an MCS is a function of the number of realizations. That is, the confidence bounds on the results can be readily computed based on the number of realizations. Every time a Monte Carlo simulation is made using the same sample size, it will come up with a slightly different value. The values converge very slowly of the order $O(n^{-\frac{1}{2}})$. This property is a consequence of the law of large numbers.

Markov Chain Monte Carlo Methods

Typically, the distribution π_X is too complex for direct simulations. Thereupon, the indirect approach of MCMC must be applied. This approach will simulate correlated samples $\{x^{(i)}\}$ from π_X. As the iterations depart more from independence, the number of iterations required for a given degree of accuracy increases. Other algorithms for constructing such transition kernel have been proposed such as importance sampling which involves sampling the points randomly, but more frequently where the integrand is large. One can approximate the integral by an integral of a similar function or use adaptive routines such as stratified sampling, or adaptive umbrella sampling, or the quasi-MCM which uses low-discrepancy sequences.

MCMC methods are widely advocated in a variety of situations where the complexity of the distribution of interest is an issue. In these situations usually the direct sampling from such complicated models is not possible. The key idea of the MCMC methods is to generate an iterative sequence of samples in such a way that it converges in distribution to the model of interest. To implement this strategy, many attempts were made to define algorithms for constructing chains with specified equilibrium distributions. The most common, well-known algorithms for constructing chains with specified equilibrium distributions were defined by Metropolis et al. (1953) and Hastings (1970). A wide range of discussion papers on MCMC theory and application can be referred to, for example, Smith and Roberts (1993), Gilks et al. (1996), Robert and Casella (2004) and Suess and Trumbo (2010). In this section, we shall briefly discuss in an appropriate framework the theory of the MCMC technique.

To sample from a specified distribution π on $\mathbb{E}$, we construct an MC transition kernel $\mathbb{P}(x, A)$. Let $X_1, X_2, \cdots, X_n$ be random variables. We say that X satisfies a Markov condition if

$$P(X_{n+1} = x | X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = P(X_{n+1} = x | X_n = x_n).$$

The transition kernel $\mathbb{P}$ is a map, $\mathbb{P} : \mathbb{E} \times \mathbb{E} \to [0, 1]$, that implies the target distribution π is a stationary distribution of the chain. The distribution of $X^{(t+1)}$ given $X^{(t)}$ satisfies

$$P(X^{(t+1)} \in A | X^{(0)} = x^{(0)}, \ldots, X^{(t-1)} = x^{(t-1)}, X^{(t)} = x) = \mathbb{P}(x, A).$$

We say that π is the invariant measure (hence equilibrium) of the MC if it satisfies the *general balance* equation

$$\int_{x\in\mathbb{E}} \pi(dx)\mathbb{P}(x, A) = \pi(A),$$

for all measurable sets A $\subset \mathbb{E}$.

General balance, $\pi\mathbb{P} = \pi$, is also referred to as the *global balance*.

The conditional distribution of $X^{(t)}$ given $X^{(0)} = x^{(0)}$ is

$$P(X^{(t)} \in A | X^{(0)} = x^{(0)}) = \mathbb{P}^t(x^{(0)}, A),$$

where $\mathbb{P}^t$ denotes the kernel $\mathbb{P}$ after iterating it t times. $\mathbb{P}$ should π irreducible, aperiodic, positive recurrent and $\pi\mathbb{P} = \pi$ (Nummelin 1984). A chain is π *irreducible* if starting at any initial state $x \in \mathbb{E}$, then for all measurable sets A $\subset \mathbb{E}$ with $\pi(A) > 0$ there exists $t > 0$ such that $P(X^{(t)} \in A | x^0 = x) > 0$. The chain is *aperiodic* if the chain does not oscillate between different sets of spaces in a regular periodic movement. The term *positive recurrent* is defined as follows: let τ_A be the first return time to state $A \subset \mathbb{E}$ where $\pi(A) > 0$, then we say the π irreducible chain $X^{(t)}$ is recurrent if $P(\tau_A < \infty) = 1$ and is positive recurrent if $E(\tau_A) < \infty$. If the chain is $\pi-$ irreducible, aperiodic, positive recurrent and if the initial value of $X^{(0)}$ is sampled from π, then all subsequent iterations using MCMC will also be distributed according to π. For example, drawing a number from $\{1, 2, 3\}$ with replacement where X_t is last number seen at time t is an MC, but if we draw a number without replacement, then it is not MC. If the initial value of $X^{(0)}$ is sampled from π, then all subsequent iterations using MCMC will also be distributed according to π.

Such methods include the MHA, GS, and the Wang and Landau algorithm. We recall now the most commonly used.

The Metropolis-Hastings Algorithm

This algorithm was first proposed by Metropolis et al. (1953) and extended by Hastings (1970). The algorithm is designed to give samples from a distribution π. It defines a proposal kernel $q(x, \cdot)$ to produce a potential new state $x' \in \mathbb{E}$. The proposed candidate x' is accepted with probability α where

$$\alpha(x, x') = \min\left\{1, \frac{\pi(x')q(x', x)}{\pi(x)q(x, x')}\right\}.$$

If we are currently at time t and x' is accepted, then $X^{(t)} = x'$ otherwise the chain does not move, i.e., $X^{(t)} = x^{(t-1)}$.

Formally, the target distribution π is defined with respect to a $\sigma-$finite measure. The proposal density q could be defined with respect to a different $\sigma-$finite measure from that for π. The transition kernel $\mathbb{P}(x, x')$ using the MHA can be written as

$$q(x, x')\alpha(x, x') \text{ if } x' \neq x.$$

The choice of the distribution $q(\cdot, \cdot)$ is arbitrary provided that $q(x, x') > 0$ if and only if $q(x', x) > 0$. It is convenient to choose a q that is simple and fast to sample from and for which it is easy to evaluate the acceptance probability. However, the relation between q and $\pi(\cdot)$ will affect the rate of convergence.

The Gibbs Sampler

The GS was given its name by Geman and Geman (1984) who used it for analyzing Gibbs distributions on a lattice. The algorithm constructs the transition kernel $\mathbb{P}$ using the full conditional densities of each component X_i, $i = 1, \ldots, n$, given the values of the other components $X_{-i} = \{X_j; j \neq i, j = 1, \ldots n\}$. We denote this density by $\pi_{X_i|X_{-i}}(x_i|x_{-i})$. Suppose we are at time t and want to update the chain, then (as it is with the MHA) we use either a random sampler or a systematic scan sampler. At each iteration, the random sampler picks a random component say, X_i, $i \in \{1 \ldots n\}$ to update, then the conditional density for $X_i^{(t)}$ becomes $\pi_{X_i|X_{-i}}(x_i|X_{-i} = x_{-i}^{(t-1)})$. In the systematic scan, we update all the components in turn during one iteration using the marginal conditional densities of the components. In progressing from $X^{(t-1)}$ to $X^{(t)}$, the value of X_i is obtained by sampling from

$$\pi_{X_i|X_{-i}}(x_i|x_1^{(t)},\ldots,x_{i-1}^{(t)},x_{i+1}^{(t-1)},\ldots,x_n^{(t-1)}).$$

Hence, to update X we make random draw from these full conditional densities for each of its components. The iteration is completed when all the components are updated. Hence, the transition probability from $x^{(t-1)}$ to $x^{(t)}$ is given by

$$\mathbb{P}(x^{(t-1)},x^{(t)}) = \prod_{l=1}^{n}\pi_{X_l|X_{-l}}(x_l^{(t)}|x_1^{(t)},x_2^{(t)},\ldots, x_{l-1}^{(t)},x_{l+1}^{(t-1)},\ldots,x_n^{(t-1)}).$$

The GS can be regarded as a special case of MHA in which the acceptance rate α is one, meaning that the candidate x' is always accepted.

The Knapsack Example

Given a set of items, each with a weight and a value, determine the number of each item to include in a collection so that the total weight is less than or equal to a given limit and the total value is as large as possible. It derives its name from the problem faced by someone who is constrained by a fixed-size knapsack and must fill it with the most valuable items. To find the most valuable subset of n items that will fit into the knapsack given their weight w_i and value v_i, and subject to knapsack weight limit b.

$z = (z_1,\cdots,z_n) \in \{0,1\}^n, z_i$ means whether we take item i feasible solutions $\mathbb{E} = \{z \in \{0,1\}^n; \sum_i w_i z_i \le b\}$. We want to maximize $\sum_i v_i z_i$ subject to $z \in \mathbb{E}$.

1. Let the current state be $X_t = (z_1,\cdots,z_n)$, we choose $j \in \{1,\ldots,n\}$ uniformly at random.
2. Flip z_j so $Y = (z_1,\cdot\cdot,1-z_j,\cdot\cdot,z_n)$.
3. If Y is feasible; that is, the acceptance probability α is high, then set $X_{t+1} = Y$, else $X_{t+1} = X_t$.

Given a state space $\mathbb{E}$ and a target distribution $\pi = C_b^{-1}\exp(b\sum_i v_i z_i)$, where C_b is constant. We apply Metropolis algorithm and choose $Y \in \mathbb{E}$ randomly using the proposal distribution $Q = P[Y = j|X_t = i] = q_{ij}$. If Y is feasible, it will be accepted with acceptance probability $\alpha = \min\{1,\exp(b\sum_i v_i(y_i - z_i))\}$.

Notice again that this process is an MC because the state we visit next depends only on the state we are currently at and no other state. The n objects z are candidates for inclusion into our random sample. But we must select the members of this set according to some probability Q.

Conclusion

MCS is a very useful mathematical technique for analyzing uncertain scenarios and providing probabilistic analysis of different situations. The basic principle for applying MC analysis is simple and easy to grasp. Various softwares have accelerated the adoption of MCS in different domains including mathematics, engineering, and finance. Various options are available to use MCS in computers. One can use any high-level programming language like C, C++, and Java. R and WinBUGS are free statistical softwares that implement MCMC methods.

Cross-References

- Gibbs Sampling
- Simulated Datasets
- Theory of Probability, Basics and Fundamentals
- Theory of Statistics, Basics, and Fundamentals

References

Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman and Hall, New York

Fisher RA (1935) The design of experiments. Oliver & Boyd, Edinburgh. RE, RI

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell PAMI-6 6: 721–741

Gilks W, Richardson S, Speigelhalter D (1996) Markov chain Monte Carlo in practice. Chapman and Hall, London

Hastings W (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109

Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E (1953) Equations of state calculations by fast computing machines. J Chem Phys 21:1087–1092
Nummelin E (1984) General irreducible Markov chains and non-negative operators. Cambridge University Press, Cambridge
Robert CP, Casella G (2004) Monte Carlo statistical methods, 2nd edn. Springer, New York. ISBN 0-387-21239-6
Smith AFM, Roberts GO (1993) Bayesian computation via the Gibbs sampler and related MCMC methods. J R Stat Soc B 55:3–23
Stigler SM (1986) The history of statistics: the measurement of uncertainty before 1900. The Belknap Press of Harvard University Press, Cambridge
Suess EA, Trumbo BE (2010) Introduction to probability simulation and Gibbs sampling with R. Springer, New York. ISBN 038740273X

Simulation

▸ Markov Chain Monte Carlo Model

Simulation Modeling

▸ Modeling of Business Processes and Crisis Management

Singular Value Decomposition

▸ Principal Component Analysis
▸ Semi-discrete Decomposition

Singular Value Decomposition = Principal Component Analysis

▸ Eigenvalues, Singular Value Decomposition

Situational Web Applications

▸ Web Mash-ups

Small-World Networks

▸ Networks in Geography

Smartphone Proximity Networks

▸ Extracting Individual and Group Behavior from Mobility Data

SMS

▸ Microtext Processing

SNA

▸ NodeXL: Simple Network Analysis for Social Media

SNA Software

▸ Tools for Networks

SNS

▸ Social Networking on the World Wide Web

SNSs

▸ Privacy and Disclosure in a Social Networking Community

Social Analysis

▸ Social Web Search

Social Anthropology

▸ Social Network Analysis in a Digital Age

Social Bookmarking

Johann Stan[1] and Pierre Maret[2]
[1]ISCOD - Institut Henri Fayol, Ecole Nationale Supérieure des Mines, Saint-Étienne, France
[2]Laboratoire Hubert Curien, Université Jean-Monnet, Saint-Étienne, France

Synonyms

Annotations; Social tags

Glossary

Tag A descriptive keyword entered by a human individual with the objective to describe a resource (e.g., a photo, a web page). It is also called an annotation or user-generated content

Resource In the context of this work, a multimedia content (e.g., text file, photos, videos, web page) available on the Internet. A resource is generally identified by an URI (Unique Resource Identifier) which enables its access using the REST protocol

Social Bookmarking System Web-based systems allowing users to describe resources with tags

Social Bookmark Tag in the form of a link to a resource (e.g., web page) that is intentionally stored, and possibly shared, by an identified individual on a social bookmarking system, on which individuals can attach tags

Folksonomy Whole set of tags that constitutes an unstructured collaborative knowledge classification scheme in a social tagging system

Definition

Social Bookmarking Systems (SBS). Web-based systems allowing users to describe resources with annotations, also called tags. The fundamental unit of information in a social bookmarking system consists of three elements in a triplet, represented as (user, resource, and tag) (Cattuto et al. 2006). This triplet is also called a tag application (instance of a user applying a tag to a resource; this is also referred to as a tag post) (Sen et al. 2006). The combination of elements in a tag application is unique. For example, if a user (also known as tagger) tags a paper twice with the same tag, it would only count as one tag application. Resources can mean different things for different social bookmarking systems. In the case of del.icio.us, a resource is a web site, and in the case of CiteULike, it is an academic paper.

Overview

Social Bookmarking: A Means for Describing Resources

Social bookmarks are tags attached to a resource, with the main objective to describe said resource. They describe the context or the meaning of such artifacts. Social bookmarks can be of multiple forms, depending on the semantic structure they rely on.

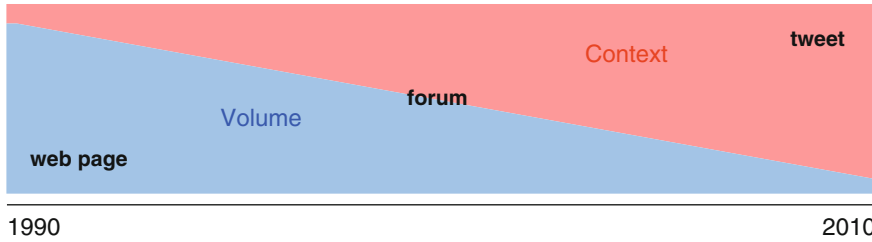
The manipulation of web resources involves tasks such as description, retrieval, reuse, presentation, and search. All these tasks need a layer of prior knowledge, which is represented by the social bookmarks, which can be composed of different types of annotations.

Such annotations may be either structured, semi-structured, or unstructured:

1. *Structured Annotations.* In this case, the terms employed in the annotation are regulated by a common domain vocabulary that must be used by the members of the system. These types of annotations are currently not used in the majority of social platforms because a domain vocabulary containing the necessary terms for the annotations is needed. Although such an approach has many advantages (e.g., absence of synonyms, absence of differences in pronunciation), this is not the natural way to describe resources in Web 2.0 platforms, as the domain is not well-defined and, therefore, it is very difficult to build such vocabularies and to

establish a consensus for each term used. At the same time, the use of semantic annotations would be cumbersome for people, as it is time-consuming and requires additional cognitive effort to select concepts from existing domain ontologies. In addition, semantic annotations work well in systems where the domain is well-defined (e.g., a system for sharing knowledge about human genes Yeh et al. 2003), but in social bookmarking systems, this is not the case, as the shared content is generally very heterogeneous, as people can discuss without limits (i.e., covers multiple domains with no regularities and relations).

2. *Semi-Structured Annotations.* In contrast, semi-structured annotations, such as social tags, are widely used or photo tagging and bookmarking (e.g., the annotation of a web page). These annotations are generally freely selected keywords without a vocabulary in the background. However, we consider them to be semi-structured, as they represent an intermediate approach between semantic annotations (i.e., annotations that are based on concepts from domain ontologies) and free-text annotations. Besides, such collections of tags converge to a structured data organization, called a folksonomy (Gruber 2007). This consists of a set of users, a set of free-form keywords (called tags), a set of resources, and connections between them. As folksonomies are large-scale bodies of lightweight annotations provided by humans, they are becoming more and more interesting for research communities which focus on extracting machine-processable semantic structures from them. These underlying data clouds of collaborative tagging systems enable Internet users to annotate or search for resources using custom labels instead of being restricted by predefined navigational or conceptual hierarchies (e.g., ontologies).
3. *Unstructured Annotations.* Finally, a more recent form of annotations is represented by free-text annotations, also called social awareness streams, composed of status updates or microposts (Naaman et al. 2010). This can be found in the majority of social networks and microblogging systems and primarily consists of free texts in the form of short messages describing a resource, a finding, an impression, a feeling, a recent activity, mood, or future plan. The limitations of this practice from the viewpoint of information retrieval and knowledge management are similar to that of social tagging, as users have complete freedom in the formulation of these messages. It is important to mention that in social awareness streams, the produced content often contains the described resource itself, in the form of an integrated hyperlink. A common practice is either to express an opinion about the resource (e.g., web page) or to provide its short summary for the community. Since Internet took over Usenet as the main computer-based means of communication, it has gone through several stages: read-only web, with large pieces of information close to magazine article size; read-write Web, or web 2.0, with forums mimicking Usenet, exchanging pieces of information up to half a page in size; blogging, close to the web page model but with a shift in authorship towards the general public; and microblogging, based on very short messages (e.g., 140 characters on Twitter) (Fig. 1). This shift from large, authoritative information to very short and amateur information is contemporary with the mobility evolution, with the more user-friendly web-enabled devices (e.g., the iPhone) emphasizing a particular factor: the context in which information is written. This has blurred the distinction between information and messaging, as all information on Twitter is in fact a message to followers, and all messages may be shared, thus creating information. Events and documentation on the contrary are becoming more distinct: in the traditional newspaper information model, documentation is delivered with events in a single article; in the Twitter-driven model, events are tweets, and the user is meant to seek information in more reliable and static sources, such as Wikipedia. An example of such as shift is the growing use of Twitter in the scientific community, contrasting strongly with the process of peer-reviewed publication.



Social Bookmarking, Fig. 1 Evolution of content production on the Internet: from structured documents to microposts

An interesting issue about free-form posts in social platforms is their short size, which emerged as a simple, convenient way to communicate about activities or share findings. The size limitation of such posts, defined by the majority of such platforms, is mostly due to the fact that users can in this way follow hundreds of friends in real time, without an important time investment. Also, this lightweight form of communication enables users to easily broadcast opinions, activities, and status (Java et al. 2007; Naaman et al. 2010).

Common practices emerged to reduce the length of messages and to help users to rapidly identify the messages relevant for them. Thus, hashtags are used to identify posts relevant to a specific event or a specific topic. Also, common ways are used to synthesize information: include the source web page or reduce the amount of stop words in order to gain place for the informative terms (keywords and named entities). These practices largely depend also on the targeted user community, which can vary from a small family to the world at large.

The same applies to the composition of such posts, where common practices emerged as new means to better identify posts relevant to a specific event, also called "hashtags," or common ways to synthesize an information, such as including the source web page or reducing the amount of stop words in order to gain place for the informative terms (keywords and named entities). These practices largely depend also on the targeted user community, which can vary from a small family to the world at large. Microposts are often called "social signals" (Mendes et al. 2010), and users of such systems "social sensors" (Sakaki et al. 2010), as they can be useful to detect important events in a given location, such as an earthquake.

An Overview of Web Repositories of User-Generated Content

Launched in 2001, Wikipedia (2001) was one of the first public crowd-sourced web site. This free encyclopedia has been allowing anyone to edit the content of any article. Whereas this openness has implied many disputes on pages related to controversial subjects (e.g., facts about presidential candidates just before election, about historical events, companies), it has grown to become a major and useful reference, covering many languages. This encyclopedia has been translated to a semantic database called DBpedia (Bizer et al. 2009) since 2007, enabling its user-generated content to be machine-readable, so that computer programs (and mashups) can leverage knowledge facts by formulating precise queries.

Even though Wikipedia has been opened to any voluntary contributions, contributors are still few, compared to the number of readers. Participating in social bookmarking sites, like Delicious (2003), has become more popular, as the contribution process was quicker, simpler, and more personal. After creating a (free) account on the site, users can immediately bookmark web pages that they want to keep, because they enjoy them, they want to be able to easily find them later, and they (often) want to share them with other people. In order to make bookmarked web pages more easy to find later, users are invited to annotate them with "tags," unconstrained words

(in any language, without even spell-checking) that subjectively reflect the apparent nature, function, category, and context of those web pages (Golder and Huberman 2006). Web pages bookmarked (and tagged) by several people are thus described by a "tag cloud," a displayed set of tags. The size of a tag depends on the number of people who used this tag to describe this page.

As any URL-located resource can be bookmarked on social bookmarking sites, these descriptions can apply on various types of entities represented by those resources. For example, tags given to a page that presents a car are most probably associated to the car, than to the page/site itself. Now that web pages exist for almost anything on earth (e.g., people, objects, places, events), social bookmarking is a promising paradigm for gathering crowd-sourced descriptions and classifications of virtual and real entities. More specific repositories also exist to represent and describe real-world entities and discover their involvement with people's activities. Concerning music, Musicbrainz (2001) can identify the name and interpreter of a song from a sampled audio (e.g., recorded with a microphone), and tags given by people to songs and artists are gathered on web sites like Last.fm (2002), which also maintains a history of the last songs that users listened to. Image-sharing web sites like Flickr (2004) can be considered as social bookmarking applied to photographs, as it is possible to tag one's own and other people's photographs, including the time and geographical location where the picture was taken.

Additionally, real-world places are described, reviewed by people, and geographically located on various web sites (and their mobile applications) such as Yelp (2004) and Qype (2006).

Rattenbury et al. (2007) have proven that names of places and events can also emerge by analyzing the frequency and temporal distribution of tags associated to geolocated pictures. Most web sites cited above expose public feeds that one can subscribe for being aware of last updates and/or APIs that allow computer programs to query information, given specific criteria (e.g., information about a place, a topic, at a given time range).

Thousands of other APIs are referenced on sites like ProgrammableWeb. Also note that tags are not directly available on all the web sites cited above, but keywords can be identified from the user-generated content they feature. It is also possible that pages from those sites are tagged on Delicious.

Knowledge Management in Social Bookmarking Systems

A category of annotations in Social Platforms are semi-structured, also called social tags. Social bookmarking systems have become extremely popular in recent years. Their underlying data structures, known as folksonomies (Mathes 2004), consists of user-tag-resource triples.

Folksonomies contain peoples' structural knowledge about documents. A person's structural knowledge has been defined as the knowledge of how concepts in a domain are interrelated from the individual's point of view. According to (Mathes 2004), an important aspect of a folksonomy is that it is comprised of terms in a flat namespace: that is, there is no hierarchy and no directly specified parent-child or sibling relationships between these terms. There are, however, automatically generated "related" tags, which cluster tags based on common URLs. This is unlike formal taxonomies and classification schemes where there are multiple kinds of explicit relationships between terms. These relationships include functions like broader, narrower, as well as related terms. These folksonomies are simply the set of terms that a group of users tagged content with; they are not a predetermined set of classification terms or labels.

Folksonomies claim to have many advantages over controlled vocabularies or formal taxonomies. Tagging has lower costs because there is no complicated, hierarchically organized vocabulary to learn and adapt to its own one. Users simply create and apply tags. According to Wu et al. "Folksonomies are inherently open-ended and therefore respond quickly to changes and innovations in the way users

categorize content" (2006). Collaborative tagging is regarded as democratic metadata generation where metadata is generated by both the creators and consumers of the content. Folksonomies can be divided into broad folksonomies, which allow different users to assign the same tag to the same resource, and narrow folksonomies, in which the same tag can be assigned to a resource only once.

The question of why folksonomies are successful has been the subject of several studies in the literature. An important argument for this is the fact that the feedback loop is tight (Mathes 2004) i.e., once the user assigns a tag to an item, the cluster of items with identical or similar tags can be immediately retrieved. This can help the user decide whether to keep the tag or change it to a similar or different one. The scope of such a cluster can be expanded by showing all items from all users in the system which are tagged with the same tag. By viewing the result set, the user can decide how to better adapt the tag to the group norm or to have better visibility in the community for the tagged resource. The issue of how to influence the group norm was also studied by Udell. This tight feedback loop leads to a form of asymmetrical communication between users through metadata. The users of a system are negotiating the meaning of the terms in the folksonomy, whether purposefully or not, through their individual choices of tags to describe documents for themselves.

A folksonomy eases collaboration. Groups of users do not have to agree on a hierarchy of tags or detailed taxonomy; they only need to agree, in a general sense, on the "meaning" of a tag enough to label similar material with terms for there to be cooperation and shared value. Although this may require a change in vocabulary for some users, it is never forced, and as Udell discussed, the tight feedback loop provides incentives for this cooperation.

The main problems of social tagging systems include ambiguity, lack of synonyms, and discrepancies in granularity (Golder and Huberman 2006). An ambiguous word, e.g., apple, may refer to the fruit or the computer company, and this in practice can make the user retrieve undesired results for a given query. Synonyms like lorry and truck or the lack of consistency among users in choosing tags for similar resources, e.g., *nyc* and *new york city*, makes it impossible for the user to retrieve all the desired resources unless he/she knows all the possible variants of the tags that may have been used. Different levels of granularity in the tags may also be a problem: documents tagged "java" may be too specific for some users, but documents tagged "programming" may be too general for others.

Several attempts have been made to uncover the structure of this kind of data organization. Basic formal models of folksonomies include that of Mika (2007) and Hotho et al. (2006). Mika proposes a model based on *tripartite hypergraphs*, while Hotho et al. on *triadic context* (term used in formal concept analysis). We present in the following the formal model of Mika, one of the most cited models in the literature for the representation of these structures.

As said before, a folksonomy is an association of users, annotations, and resources. The corresponding three disjoint set of vertices are considered by Mika in the formal model: the set of actors (users) -A-, the set of concepts (tags) -C-, and the set of resources -O- (e.g., photos, videos, or web resources, like bookmarks, web sites). Since in a social tagging system, users tag objects with concepts, ternary relations are created between the user, the concept, and the object.

This resulting tripartite hypergraph can be transformed into several bipartite graphs, each having a very specific meaning, like AC, the graph that associates actors and concepts; CO, the graph that associates concepts and objects; and AO, the graph that associates actors and resources.

Abel (2008) investigates the benefits of additional semantics in folksonomy systems. Additional context can be provided to the tagging activity with an extension of the tripartite model, i.e., an association of the user, the tag, and the tagged resource, that describes more precisely the particular tagging activity. For example, time stamp helps to categorize tags in a temporal manner; the mood the user had when tagging

the resource would allow to qualify opinions expressed in a tag. Other information, like background knowledge about the user, would allow to have information about the reliability of the tagger. The GroupMe! folksonomy system is proposed, which is a new kind of resource sharing system for multimedia web resources. A first extension of previous models is the introduction of the term group, which is a finite set of related resources. The folksonomy model in GroupMe! can be thus formalized in the following manner (we note with F the folksonomy model): $F = (U, T, IR, G, Y)$, where U, T, R, G are finite sets that contain instances of users, tags, resources, and groups. $IR = R \cup G$ is the union set of resources and the set of groups.

Wu et al. (2006) identify the key challenges in collaborative tagging systems. The three identified challenges are the following: (i) the identification of communities (i.e., groups of users with similar interests), (ii) preventing information overload by filtering out high-quality documents and users (e.g., experts in a domain), and (iii) how to create scalable, navigable structures from folksonomies. Folksonomies are criticized to have flaws that formal classification systems are designed to eliminate, including polysemy, words having multiple related meanings, and synonymy, multiple words having the same or similar meanings.

Information Retrieval from Folksonomies: Social Information Retrieval

In the previous section, we have seen the general definition and structure of folksonomies, the data organization in social tagging systems. In this section, we go further and review existing techniques of information retrieval in folksonomies.

The biggest challenge in folksonomies is information retrieval, i.e., the question of how to efficiently rank items (e.g., tags, resources, users) for a given user query. In traditional Internet applications, the search and navigation process serves two vital functions: retrieval and discovery. Retrieval incorporates the notion of navigating to a particular resource or a resource containing particular content. Discovery incorporates the notion of finding resources or content interesting but therefore unknown to the user. The success of collaborative tagging is due in part to its ability to facilitate both these functions within a single user-centric environment. Reclaiming previously annotated resources is both simple and intuitive, as most collaborative tagging applications often present the user's tag in the interface. Selecting a tag displays all resources annotated by the user with that tag. Users searching for particular resources they have yet to annotate may select a relevant tag and browse resources annotated by other users. However, the discovery process can be much more complex. A user may browse the folksonomy, navigating through tags, resources, or even other users. Furthermore, the user may select one of the results of a query (i.e., tag, resource, or user) as the next query itself. This ability to navigate through the folksonomy is one reason for the popularity of collaborative tagging.

In order to provide efficient retrieval mechanisms, a formal model of folksonomies is required. There are several models in the literature, e.g., that of Mika (2007) and Hotho et al. (2006). Mika proposes a model based on *tripartite hypergraphs*, while Hotho et al. on *triadic context* (term used in formal concept analysis).

Hotho et al. adapt the well-known PageRank algorithm in order to apply it on folksonomies, called *FolkRank*. The impossibility of applying *PageRank* has its origins in the fact that a folksonomy is different from the web graph (undirected triadic hyperedges instead of directed binary edges). By modifying the weights for a given tag, FolkRank can compute a ranked list of relevant tags.

The original formulation of PageRank (Brin and Page 1998) reflects the idea that a page is important if there are many pages linking to it and if those pages are important themselves (recursive aspect of importance). The distribution of weights can thus be described as the fixed point of a weight passing scheme on the web graph. This idea was extended in a similar fashion to bipartite subgraphs of

the web in HITS (Kleinberg 1999) and to n-ary directed graphs (Xi et al. 2004). The same underlying principle is employed for the ranking scheme in folksonomies. The basic notion is that a resource which is tagged with important tags by important users becomes important itself. The same holds, symmetrically, for tags and users. Such a ranking schema can help the emergence of a common vocabulary in collaborative tagging systems, by recommending to the user tags that have a bigger visibility in the community and that is also semantically close to the user-defined tag.

Abel et al. (2008) perform an in-depth analysis of ranking algorithms specially designed for folksonomies – FolkRank, SocialSimRank (Bao et al. 2007), and SocialPageRank – and adapt them to the GroupMe! social bookmarking system, where an additional dimension is added to folksonomies, i.e., groups of resources.

Gemmell et al. (2008) propose a method to personalize a user's experience within a folksonomy using unsupervised clustering of social tags as intermediaries between a query and a set of items. Terms in the query are weighted based upon their affinities to particular clusters to help disambiguate queries.

Bao et al. (2007) propose different algorithms, such as SocialSimRank and SocialPageRank to optimize web search using social annotations. The underlying hypotheses of the proposed algorithms are the following: (i) social annotations about web pages are good summarizations of the given web page and can be used for efficient computation of similarity between a search query and a web page, and (ii) the amount of annotations assigned to a web page is a good indication of its popularity.

Vocabulary Construction and Emergence of Semantics

In this section, we present different approaches for extracting and constructing a hierarchical structure of tags in collaborative tagging systems. Recently, several papers proposed different approaches to construct conceptual hierarchies from tags collected from social web sites. Mika (2007) uses a graph-based approach to construct a network of related tags, projected from either a user-tag or object-tag association graphs. Although there is no evaluation of the induced broader/-narrower relations, the work provides a good suggestion to infer them by using betweenness centrality and set theory. Other works apply clustering techniques to keywords expressed in tags and use their co-occurrence statistics to produce conceptual hierarchies (Brooks and Montanez 2006; Zhou et al. 2007).

Brooks and Montanez (2006) argue that hierarchical structures which seem to match that created by humans can in fact be inferred from existing tags and articles in collaborative tagging systems. This may imply that folksonomies and traditional structured representations are not so opposed after all, rather, tags are a first step in helping an author or reader to annotate her information. Automated techniques can then be applied to better categorize specific articles and relate them more effectively to other articles. The method used is agglomerative clustering and consists of the following steps: the comparison of each tag cluster to every other tag cluster, using the pairwise cosine similarity metric. Each article in cluster one is compared to each article in cluster two, and the average of all measurements is computed. The two closest-similarity clusters from the list of tag clusters are removed and replaced with a new abstract tag cluster, which contains all of the articles in each original cluster. This cluster is annotated with an abstract tag, which is the conjunction of the tags for each cluster.

This procedure is followed until there is a single global cluster that contains all of the articles. By recording the order in which clusters are grouped into progressively more abstract clusters, a tree that shows the similarity of tags can be constructed. Plangprasopchok and Lerman (2009) propose a different approach for constructing folksonomies from user-specified relations on Flickr by statistically aggregating tags from different collections. This approach uses the shallow hierarchies created through the collection-set relations on Flickr. Authors argue that partial hierarchies are a good source information for generating folksonomies and propose a

simple statistical approach to resolve hierarchical relation conflicts in the aggregation process.

Another approach for the extraction of hierarchical semantics from social annotations is proposed by Zhou et al. (2007). A probabilistic unsupervised method is proposed, called Deterministic Annealing. This method performs a top-down approach on the flat tag space, beginning with the root node containing all annotations and splitting it to obtain clusters with narrower semantics.

Cattuto et al. (2008) perform an analysis on a large-scale snapshot of the popular social bookmarking system Delicious. To provide a semantic grounding of the folksonomy-based measures, tags of Delicious are mapped to synsets of WordNet (Markines et al. 2009) and use the semantic relations of WordNet to infer corresponding semantic relations in the folksonomy. In WordNet, the similarity is measured by using both the taxonomic path length and a similarity measure by Jiang and Conrath (1997) that has been validated through user studies and applications (Budanitsky and Hirst 2006). The use of taxonomic path lengths, in particular, allows to inspect the edge composition of paths leading from one tag to the corresponding related tags, and such a characterization proves to be especially insightful. Co-occurrence is a measure that extracts from the folksonomy a graph for tags, where edges are weighted with the number of times they co-occur (i.e., tags on the same resource).

The results can be taken as indicators that the choice of an appropriate relatedness measure is able to yield valuable input for learning semantic term relationships from folksonomies, i.e., (i) synonym discovery, (ii) concept hierarchy extraction, and (iii) the discovery of multiword lexemes. The cosine similarity is clearly the measure to choose when one would like to discover synonyms. Cosine similarity delivers not only spelling variants but also terms that belong to the same WordNet synset. Both FolkRank and co-occurrence relatedness yield more general tags. This could be a proof that these measures provide valuable input for algorithms to extract taxonomic relationships between tags.

Conclusion

The objective of this entry was to present the main constituents of social bookmarking, a very popular activity on the web nowadays. We depicted the main pillars of such systems and highlighted reference scientific work related to the manipulation of the knowledge in such systems for information retrieval and classification.

Cross-References

- ▸ Analysis and Mining of Tags, (Micro)Blogs, and Virtual Communities
- ▸ Collective Intelligence, Overview
- ▸ Folksonomies
- ▸ Tag Clouds

References

Abel F (2008) The benefit of additional semantics in folksonomy systems. In: Proceedings of the 2nd PhD workshop on Information and knowledge management, PIKM'08, Napa Valley. ACM, New York, pp 49–56. doi:10.1145/1458550.1458560, http://doi.acm.org/10.1145/1458550.1458560

Abel F, Henze N, Krause D (2008) Ranking in folksonomy systems: can context help? In: Proceedings of the 17th ACM conference on information and knowledge management, CIKM'08, Napa Valley. ACM, New York, pp 1429–1430

Bao S, Xue G, Wu X, Yu Y, Fei B, Su Z (2007) Optimizing web search using social annotations. In: Proceedings of the 16th international conference on World Wide Web, WWW'07, Banff. ACM, New York, pp 501–510

Bizer C, Lehmann J, Kobilarov G, Auer S, Becker C, Cyganiak R, Hellmann S (2009) Dbpedia – a crystallization point for the web of data. J Web Sem 7(3): 154–165

Brin S, Page L (1998) The anatomy of a large-scale hypertextual web search engine. Comput Netw ISDN Syst 30(1–7):107–117. doi:10.1016/S0169-7552(98)00110-X

Brooks CH, Montanez N (2006) Improved annotation of the blogosphere via autotagging and hierarchical clustering. In: Proceedings of the 15th international conference on World Wide Web (WWW2006), Edinburgh, pp 625–632

Budanitsky A, Hirst G (2006) Evaluating wordnet-based measures of lexical semantic relatedness. Comput

Linguist 32(1):13–47. doi:10.1162/coli.2006.32.1.13, http://dx.doi.org/10.1162/coli.2006.32.1.13
Cattuto C, Loreto V, Pietronero L (2006) Semiotic dynamics in online social communities. Eur Phys J 46(S2):33–37
Cattuto C, Benz D, Hotho A, Stumme G (2008) Semantic grounding of tag relatedness in social bookmarking systems. In: Proceedings of the 7th international conference on the semantic web, ISWC'08, Karlsruhe. Springer, Berlin/Heidelberg, pp 615–631. doi:10.1007/978-3-540-88564-1_39, http://dx.doi.org/10.1007/978-3-540-88564-1_39
Delicious (2003) Delicious social bookmarking system. http://delicious.com
Flickr (2004) Flickr – image and video hosting website. http://flickr.com
Gemmell J, Shepitsen A, Mobasher M, Burke R (2008) Personalization in folksonomies based on tag clustering. In: Proceedings of the 6th workshop on intelligent techniques for web personalization and recommender systems, Chicago
Golder SA, Huberman BA (2006b) Usage patterns of collaborative tagging systems. J Inf Sci 32(2):198–208. doi:10.1177/0165551506062337, http://dx.doi.org/10.1177/0165551506062337
Gruber T (2007) Ontology of folksonomy: a mash-up of apples and oranges. Int J Semant Web Inf Syst 3(2):1–11
Hotho A, Jaschke R, Schmitz C, Stumme G (2006) Information retrieval in folksonomies: search and ranking. In: Proceedings of the 3rd European semantic web conference, Budva, Montenegro. Lecture notes in computer science, vol 4011. Springer, Berlin, pp 411–426
Java A, Song X, Finin T, Tseng B (2007) Why we Twitter: understanding microblogging usage and communities. In: Proceedings of the joint 9th WEBKDD and 1st SNA-KDD workshop 2007, San Jose
Jiang J, Conrath D (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the international conference on research in computational linguistics, Taipei
Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632. doi:10.1145/324133.324140, http://doi.acm.org/10.1145/324133.324140
Last.fm (2002) Last.fm – music website. http://last.fm
Markines B, Cattuto C, Menczer F, Benz D, Hotho A, Stumme G (2009) Evaluating similarity measures for emergent semantics of social tagging. In: Proceedings of the 18th international conference on World Wide Web, WWW'09, Madrid. ACM, New York, pp 641–650. doi:10.1145/1526709.1526796, http://doi.acm.org/10.1145/1526709.1526796
Mathes A (2004) Folksonomies – cooperative classification and communication through shared metadata. In: Computer mediated communication – LIS590CMC. http://www.adammathes.com/academic/computer-mediated-communication/folksonomiesl.html. Accessed 25 Feb 2013
Mendes PN, Passant A, Kapanipathi P, Sheth AP (2010) Linked open social signals. In: Proceedings of the 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology – volume 01, WI-IAT'10, Los Alamitos. IEEE Computer Society, Washington, pp 224–231. doi:10.1109/WI-IAT.2010.314, http://dx.doi.org/10.1109/WI-IAT.2010.314
Mika P (2007) Ontologies are us: a unified model of social networks and semantics. Web Semant 5(1):5–15. doi:10.1016/j.websem.2006.11.002, http://dx.doi.org/10.1016/j.websem.2006.11.002
Musicbrainz (2001) Musicbrainz – the open music encyclopedia. http://musicbrainz.org
Naaman M, Boase J, Lai CH (2010) Is it really about me?: message content in social awareness streams. In: Proceedings of the 2010 ACM conference on computer supported cooperative work, CSCW'10, Savannah. ACM, New York, pp 189–192. doi:10.1145/1718918.1718953, http://doi.acm.org/10.1145/1718918.1718953
Plangprasopchok A, Lerman K (2009) Constructing folksonomies from user-specified relations on flickr. In: Proceedings of the 18th international conference on World wide web (WWW '09), Madrid. ACM, New York, pp 781–790. doi:10.1145/1526709.1526814, http://doi.acm.org/10.1145/1526709.1526814
Qype (2006) Qype – local directory service with social networking and user reviews. http://qype.com
Rattenbury T, Good N, Naaman M (2007) Towards automatic extraction of event and place semantics from flickr tags. In: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR'07, Amsterdam. ACM, New York, pp 103–110. doi:10.1145/1277741.1277762
Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World Wide Web, WWW'10, Raleigh. ACM, New York, pp 851–860. doi:http://doi.acm.org/10.1145/1772690.1772777, http://doi.acm.org/10.1145/1772690.1772777
Sen S, Lam SK, Rashid AM, Cosley D, Frankowski D, Osterhouse J, Harper FM, Riedl J (2006) Tagging, communities, vocabulary, evolution. In: Proceedings of the 20th anniversary conference on computer supported cooperative work, CSCW'06, Banff. ACM, New York, pp 181–190. doi:10.1145/1180875.1180904
Wikipedia (2001) Wikipedia knowledge base. http://wikipedia.org
Wu H, Zubair M, Maly K (2006) Harvesting social knowledge from folksonomies. In: Proceedings of the 17th conference on hypertext and hypermedia, Odense. ACM, New York, pp 111–114
Xi W, Zhang B, Chen Z, Lu Y, Yan S, Ma WY, Fox EA (2004) Link fusion: a unified link analysis framework for multi-type interrelated data objects. In: Proceedings of the 13th international conference on World

Wide Web, WWW'04. ACM, New York, pp 319–327. doi:10.1145/988672.988715, http://doi.acm.org/10.1145/988672.988715

Yeh I, Karp PD, Noy NF, Altman RB (2003) Knowledge acquisition, consistency checking and concurrency control for gene ontology (go). Bioinformatics 19(2):241–248

Yelp (2004) Yelp – local directory service with social networking and user reviews. http://yelp.com

Zhou M, Bao S, Wu X, Yu Y (2007) An unsupervised model for exploring hierarchical semantics from social annotations. In: Aberer K, Choi KS, Noy NF, Allemang D, Lee KI, Nixon LJB, Golbeck J, Mika P, Maynard D, Mizoguchi R, Schreiber G, Cudré-Mauroux P (eds) Proceedings of the 6th international the semantic web and 2nd Asian conference on Asian semantic web conference (ISWC'07/ASWC'07), Busan. Lecture notes in computer science, vol 4825. Springer, Berlin/Heidelberg, pp 680–693

Social Capital

Roger Leenders
Tilburg School of Social and Behavioral Science, Department of Organization Studies, Tilburg University, Tilburg, The Netherlands

Synonyms

Capital; Goodwill; Human capital; Social networks; Trust

Glossary

Social Capital Productive resources residing in and resulting from social networks

Social Liability Obstructive resources residing in and resulting from social networks

Human Capital One's stock of competencies, knowledge, skill, education

Political Economics The study of the relationship between politics and economics in society

Definition

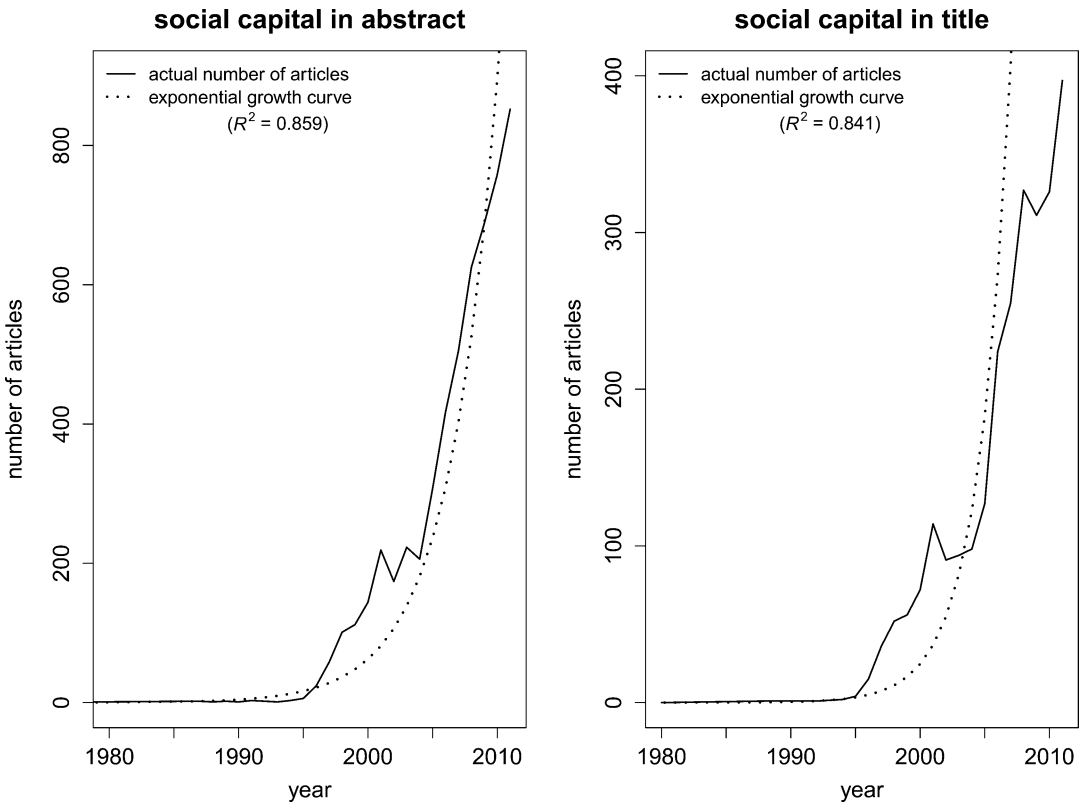
By analogy with notions of physical capital and human capital, "social capital" refers to the features of social organization that facilitate coordination and cooperation for mutual or individual benefit. According to sociologist James Coleman (1990), "Like other forms of capital, social capital is productive, making possible the achievement of certain ends that would not be attainable in its absence. . . . In a farming community . . . where one farmer got his hay baled by another and where farm tools are extensively borrowed and lent, the social capital allows each farmer to get his work done with less physical capital in the form of tools and equipment."

Social capital has been referred to as "the glue that holds society together" and is centrally concerned with the value and implications of relationships as a resource for social action. It is often considered to be the contextual complement to human capital. Social capital theory contends that returns to intelligence, education, and seniority depend considerably on a person's location in the social structure of a market or hierarchy. While human capital refers to individual ability, social capital refers to opportunity (Burt 1997).

During recent years, the concept of social capital has become one of the most popular exports from sociological theory into everyday language (Portes 1998). As a term, "social capital" has become one of sociology's trendiest terms, both in academic literature and popular publications. There seems to be a contagious quality to the concept's predominant focus on positive aspects of human interrelationships. In addition, the concept is attractive to many by providing a broad framework that focuses on non-monetary capital as a source of influence and prosperity.

Historical Background

Although the active use of the concept dates back to the nineteenth century, the social capital concept only became popular in the 1980s and this popularity accelerated in a major way in the 1990s. Figure 1 shows the number of academic articles abstracted in Scopus (1980–2011) with "social capital" in the title or in the abstract, both displaying exponential growth. When books, book chapters, reviews, conference papers,



Social Capital, Fig. 1 The social capital concept in the academic literature

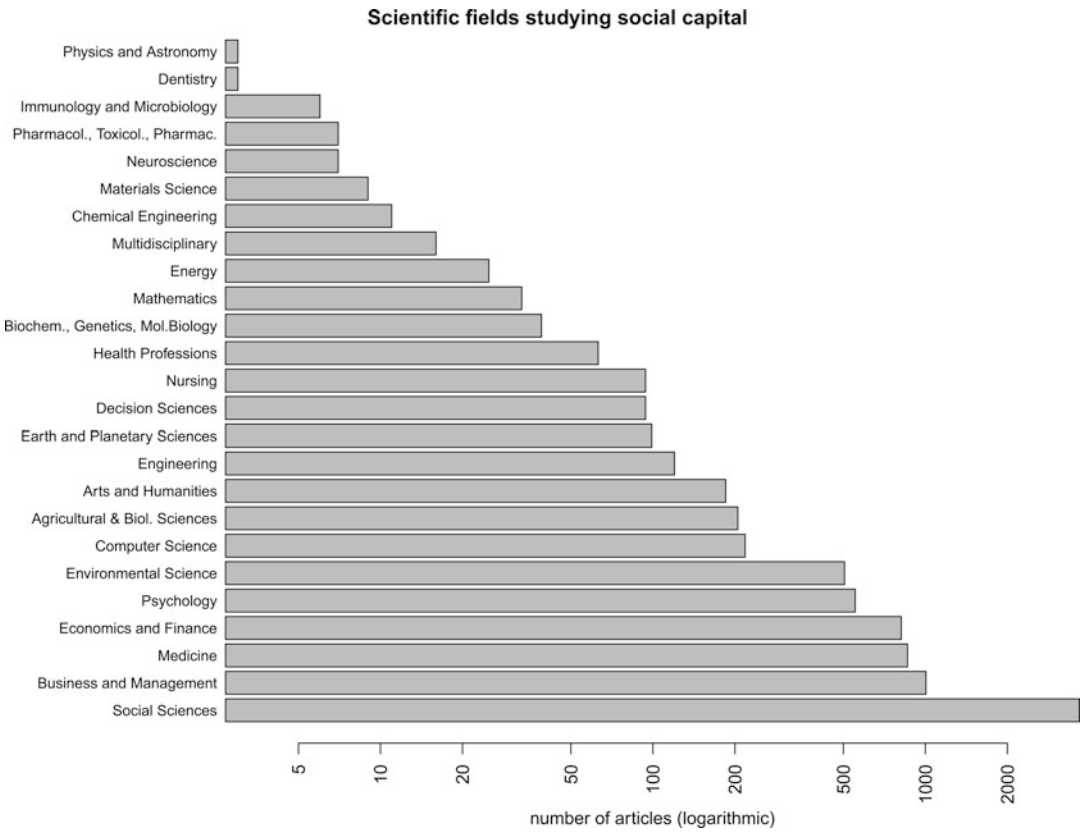
editorials, and popular press are included, the numbers go up further still. Until around 1980 there were virtually no papers that featured social capital. In the late 1990s the World Bank started a *Social Capital Initiative*, a program with the aim of defining, monitoring, and measuring social capital (Grootaert and Van Bastelaer 2001). The initiative is ongoing and active today (http://go.worldbank.org/VEN7OUW280).

Apart from economics and sociology, the concept's original habitats, "social capital" has been adopted by a very wide range of disciplines. Figure 2 gives an overview of the fields (as categorized by Scopus) that published papers with "social capital" in their abstract. The figure shows that an astounding variety of academic disciplines publish papers that employ the concept, including the disciplines "biochemistry," "earth and planetary science," "agricultural science," "computer science," and "medicine"–not exactly social sciences. Notwithstanding this variety in fields, virtually all of these papers consider social capital in terms of the ability of actors to secure benefits by virtue of membership in social networks or other social structures. Obviously, the concept has caught on.

The Evolution of the Concept

In a way, there seem to be two largely separate histories of the "social capital" concept. The first starts in the late nineteenth century and runs into the beginning of the twentieth. The second starts around 1980 and is ongoing. Although today's use of the social capital concept differs from the way it was originally developed over a century ago, it is instructive to describe at least a little bit of the concept's original heritage.

It appears that political economists were the first to use "social capital" in their writings. Alfred Marshall used the term in 1890



Social Capital, Fig. 2 The scientific fields that study social capital

(Marshall 1890). Before that, so had John Bates Clark (1885), Henry Sidgwick (1883), and Karl Marx (1867). In their work, they were opposing what they regarded as the unsocial point of view of classical political economy. Contrasting with the "individual" point of view of capitalists, social capital was "capital from the social point of view." Social capital was an aggregate of tools, inventions, improvements in land, roads, bridges, the organization of the State, and the skill and ability of humans. Also, immaterial elements were added to the concept, such as "goodwill" (Farr 2004, p. 22). As I will show later in this article, goodwill is still at the core of contemporary views of social capital.

The way in which the political economists of the nineteenth century thought about social capital painted a lively picture of corporations, trade unions, friendly societies, brotherhoods, guilds, communes, and cooperatives of endless variation. Through their joint ties, these cooperatives attempted to increase wages, share wealth, and render mutual aid (Farr 2004). Perhaps surprisingly, this picture is in many ways quite close to the social networks approach to social capital that has become the concept's dominant focus in contemporary research.

Farr (2004, p. 25) eloquently describes how the political economists' approach to social capital relates to its contemporary treatment:

> The political economists of the nineteenth century... took capital from the social point of view. Today's social capitalists, apparently, take "the social" from capital's point of view. The one reflected an age coming to terms with capital, the other an age coming to capital for its terms. Then, "social capital" expressed an explicit antithesis to an unsocial perspective upon capital, now, an implicit antithesis to a noncapitalist perspective on society. "Social capital" was once a category of political economy in a period of its transformation,

> now one of economized politics, expressing the general dominance of economic modes of analysis in society and social science. But, in the long view, these perspectives may not be logical antinomies so much as two sides of the same coin. Both, surely, sought or seek to comprehend the social relations constitutive of modern capitalist societies, and to position capital as their governing asset. And both, significantly, did so in the very terminology of "social capital."

The contemporary strain of social capital studies flows primarily from the works of Pierre Bourdieu, James Coleman, and Robert Putnam. In *The Forms of Capital*, Bourdieu (1986) distinguishes between three forms of capital: economic capital, cultural capital, and social capital. For him, social capital is made up of social obligations and connections. It is "the aggregate of the actual or potential resources linked to possession of a durable network of more or less institutionalized relationships of mutual acquaintance and recognition—or in other words, to membership in a group." To Bourdieu, social capital can be broken into two elements: the size of one's connections and the volume of capital (economic, cultural, or symbolic) in these connections' possession. To Bourdieu social capital refers to a sphere of "mutual acquaintance and recognition" (Bourdieu and Wacquant 1992). In his view, social capital cannot be reduced to economic or cultural capital, nor is it independent of them: it acts as a multiplier for the two other forms.

For Coleman, social capital consists of "a variety of different entities with two elements in common: they all consist of some aspect of social structure, and they facilitate certain actions of actors—whether personal or corporate actors—within the structure" (Coleman 1990). One of Coleman's well-known examples is that of the Jewish diamond traders of New York. The merchants were able to have their diamonds appraised through their local networks without the need to resort to costly legal contracts to safeguard against being cheated, because of the strength of the ties between their community members and the ready threat of exclusion if trust was violated. As a result, the traders were able to increase their economic advantage because of their social networks (Coleman 1988).

Coleman's approach derives from his interest in drawing together the insights from two disciplines: economics and sociology. Where in Bourdieu's work social capital serves to multiply economic and cultural capital, in Coleman's work an important function of social capital is in the multiplication of human capital. His main argument was that social capital had a profoundly beneficial effect on the acquisition of educational credentials (Schuller et al. 2001).

The fundamental difference between the Bourdieu and Coleman definitions lies in how and why the social processes develop. For Bourdieu, social processes are constrained by underlying economic organization; in his view the potential of profit is the very reason for the solidarity that makes group existence possible. In fact, Bourdieu argues that these processes may become habitualized and become reinforced by "habitus." For Coleman, on the other hand, they are created by the free will of individuals. In his approach, social capital is created by rational, purposeful individuals who build social capital. As they attempt to maximize their individual opportunities, individuals freely choose to build networks to further their self-interest. Coleman views social capital as a form of contract: individuals must have trust that others will reciprocate their actions and will feel some social obligation to do so.

The disparity in the definition of social capital between Coleman and Bourdieu has consequences in the way social capital needs to be measured. An analysis based on Bourdieu's definition would need to include an understanding of the material conditions driving the formation of social processes. A Coleman-esque analysis needs only to consider motivation at the individual (or aggregated individual) level.

One of Coleman's chief contributions to the social capital literature may be in his relatively straightforward sketch of the concept, which attracted widespread attention among social researchers. Bourdieu's work became popular only after it had been translated from French to English. Coleman's work has probably shaped the contemporary debate more than that of any other author. Since it has been so prominent, it has also been widely criticized. Important criticism comes

from Portes (1998) who charged Coleman with using a "rather vague definition" that "opened the way for re-labelling a number of different and even contradictory processes as social capital" (Portes 1998, p. 5). In particular, Portes argued for the need to draw a clear line between membership of social structures on the one hand and the resources acquired through such membership on the other.

Despite the reservations that have been voiced regarding his work, Coleman's contributions have been both influential and significant. Although now overshadowed by Putnam in the wider public policy debate, Coleman has arguably had much greater influence over scholarship in the debate so far (Schuller et al. 2001, p. 8) and justifiably so.

Probably the currently most well-known author on social capital is Robert Putnam, who has appeared in televised talk shows, was invited to Camp David, and was even featured in *People* magazine. Social capital is now deployed in a great many fields and Robert Putnam undoubtedly is the author whose work is cited across a wider range than any other (Schuller et al. 2001). Much of his work's popularity is due to the use of a clever and compelling metaphor, "Bowling Alone," characterizing the transformation of American social and political life during the postwar era (Putnam 1995, 2000). His work deployed the example of bowling as an activity which used to be highly associational, not only a source of recreational pleasure but also of social interaction, a key component of social capital.

In this work, Putnam argued that a decline in civic culture was occurring in the United States since the 1960s, an idea that resonated with many of his readers. Controlling for political ideology, tax revenues, and several other conditions, Putnam concluded that the best predictor of governmental performance was a strong local tradition of civic engagement, measured by social capital variables such as membership in voluntary organizations and voter participation in elections. Putnam was certainly not the first to call attention to the disintegration of American civic culture, but his work is clearly distinct from earlier authors through its specific focus on the eroding of social capital. According to Putnam, social capital "refers to the collective value of all 'social networks' and the inclinations that arise from these networks to do things for each other." In other words, it refers to features of social organization, such as trust, norms, and reciprocity, that can improve the efficiency of society by facilitating participants to act together more effectively to pursue shared objectives. Like Coleman, Putnam's definition strongly relies on networks and social linkages but Putnam aggregates the social capital of individuals to a "collective social capital" of a population, state, or community. Putnam's main argument is that social capital is a key component to building and maintaining democracy and notices its decline by, among other things, lower levels of trust in government and lower levels of civic participation. He makes the claim that television ("the only leisure activity where doing more of it is associated with lower social capital") and urban sprawl ("every ten minutes of commuting reduces all forms of social capital by 10 %") have had a significant role in making the USA far less "connected." In his analysis, Putnam focuses on the creation of civic norms, which lead to socioeconomic order; this is basically the reverse of Bourdieu's description of the relationship.

Notwithstanding the "celebrity status" his work gained him, Putnam's work has also been extensively criticized. Putnam's arguments have been criticized as being circular and tautological: "social capital is simultaneously a cause and an outcome. It leads to positive outcomes, such as economic development and less crime, and its existence is inferred from the same outcomes" (Portes 1998). Other criticism (e.g., McLean et al. 2002) relates to his lack of sound empirical measures (of both social capital and his dependent variables), inconsistent and incomplete derivation of his causal statements, the presence of implicit ideological underpinnings, and historical inaccuracy. In a 1998 special issue of the *American Behavioral Scientist*, several of his key results were reexamined, many of which found no or only limited support.

Perhaps the strongest points of criticism has been raised by Boggs (2001), who writes that "the author's iconic status does not prevent his book

from being so conceptually flawed and historically misleading that it would seem to require yet another large tome just to give adequate space to the needed systemic critique." Boggs makes some compelling arguments that social capital is not on the decline at all, for example, because "Putnam fails to consider the spread of newer, in many ways more interesting, civic phenomena over exactly that same time span—not only social movements but thousands of self-help and new-age groups, religious movements, and community organizations (resource centers, clinics, bookstores, periodicals, public interest groups, tenants associations, and so forth) often spawned by the larger movements" (Boggs 2001, p. 286). Boggs argues that social capital in fact has resurfaced, but in new (often apolitical) forms. He even takes issue with Putnam's Bowling Alone metaphor, which Putnam derived from the observation that participation in bowling leagues had declined by about 40 %, a sign to Putnam that social networks were eroding. Not only can bowling activity have been rechanneled to even more socially interactive sports like golf or soccer (Lemann 1996), but Boggs argues that people simply switched from bowling in leagues to bowling in more informal groups of friends and relatives: it would be quite rare for people to actually bowl alone…

Basically, every aspect of Putnam's work has been criticized, sometimes dogmatically so and overly harsh. Indeed, Putnam's measures and concept of social capital are relatively weak and lack definition and consistency. Putnam's work popularized "social capital" quickly and across a great many disciplines. It was inevitable that many authors blindly copied Putnam's less-than-fully sound approach, yielding a surge of social capital-based papers that lack rigor and consistency. However, it is hardly fair to blame Putnam for this. Rather, he deserves some praise for bringing an academic concept to the political agenda and the general public in a easily understandable way. It is now up to the (scientific) community to develop sound definitions, measures, and causal models that bring the concept further.

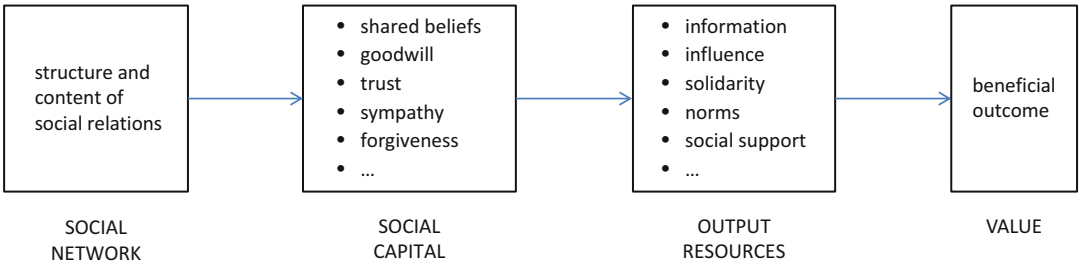
A Model of Social Capital

The explosion of social capital research is having a predictable consequence: the term is proliferating meanings. It has been applied in so many different contexts that it has lost any distinct meaning. Many social science researchers and policy makers may have embraced the term because it provides a hardnosed economic feel while restating the importance of the "social" (Halpern 2005). The concept has therefore been characterized as a "wonderfully elastic term" and a "notion that means many things to many people" and that has taken on "a circus tent quality" (Adler and Kwon 2002, p. 18). As a consequence, social capital may be at a risk of being used as a metaphor only.

The commonalities of most definitions of social capital are that they focus on social relations that have productive benefits. In the remainder of this article, I will adopt a definition and a conceptual broad model of social capital that includes both social capital sources and outcomes, multiple levels of aggregation, allows room for multidimensionality and multidisciplinarity, and can be extended to incorporate "time" as a variable in the social capital process. Of course, as with any definition, its leniency and agility is in how it is interpreted and applied, so I dare make no claim that this definition fits with every research project in any discipline. In fact, it is unlikely that a definition of social capital can be made that fits that bill and would still be useful. My definition is:

> Social capital refers to the social resources that accrue to an actor (or a set of actors). Its source lies in the structure and content of the actor's social relations. These social resources facilitate the attainment of goals of the actor (or set of actors).

This definition is largely based on the definitions provided by Gabbay and Leenders (1999) and Adler and Kwon (2002). The term "social resources" refers to goodwill, norms, sympathy, trust, forgiveness, and shared beliefs that make alters willing or more likely to share resources



Social Capital, Fig. 3 A conceptual model of social capital

that facilitate the attainment of some goal. As Adler and Kwon (2002) put it: "if goodwill is the *substance* of social capital, its *effects* flow from the information, influence, and solidarity such goodwill makes available." In the current definition, the information and influence (or other resources made available through the goodwill within an actor's social ties) needs to support the attainment of some goal for it to constitute to social capital. The reason for this is that one can easily argue that available information that one may not need or understand cannot fruitfully be considered capital. Social capital requires the use or mobilization of the actor's social resources in *purposive actions* (Lin 2001).

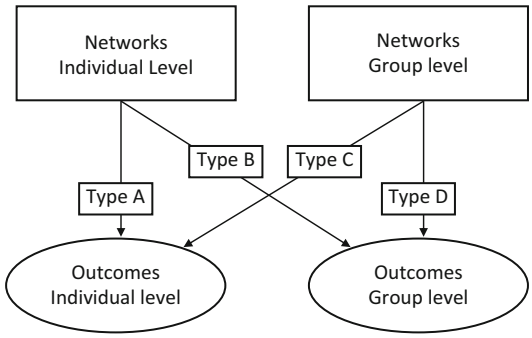
This view on social capital is depicted in Fig. 3. While it is consistent with a large proportion of the social literature, authors vary in their implementation of it. For example, the model categorizes "norms" as flowing from the goodwill that is inherent in the social network. However, there are also authors who view norms as inherent in social networks themselves – the presence of certain norms in a network can even be a reason for an actor to join the network – and these norms then make people likely to share knowledge with their alters; this would put norms in the social capital box. For this article it is not necessary to have a full-blown categorization that fits with the entire literature (which would be impossible, anyway) or with which all researchers would agree (which is equally outside the realm of possibility). What is important is that it fits with *most* of the literature and addresses the basic process that underlies social capital production and effects.

A Multilevel Concept

Much debate in the social capital literature deals with whether social capital resides at the individual level of aggregation or whether it is a group-level phenomenon. James Coleman focused mainly, though not exclusively, on the individual: a person's set of social ties provides that person with benefits. Robert Putnam, on the other hand, mainly focuses on the distribution of social ties within societies and studies how this high-level social structure produces outcomes at the level of the community.

The level-of-analysis discussion relates to two separate questions: what is the level of analysis at which the social ties reside and is their outcome accrued to the group or to the individual? A third, related, question addresses whether the social capital process is driven by *individual* goal seeking behavior or whether it is driven by *communities/collectives* that have preferred outcomes they seek to fulfill.

When studying the collective social capital literature one can only conclude that questions regarding the level at which the relevant social ties or their outcomes reside are unnecessarily restrictive and ignorant of empirical reality. Just as an individual can mobilize his personal contacts' resources for purposive action, so can a formal organization activate various resource networks to achieve its goals (Knoke 1999). There is ample evidence that individuals benefit from their own individual-level social networks as well as from the ties maintained by collectives they are part of. Similarly, collectives such as organizational groups draw the fruits from both their own connectivity with other collectives and the



Social Capital, Fig. 4 Two-level model of social capital

ties maintained by (some of) their individual members. In fact, fine-grained analyses indicate that more levels than just two are often relevant. Figure 4 shows the multilevel nature of social capital, with only two levels for simplicity. It is based on Gabbay and Leenders (2001), who classify the multilevel character of the literature into four main categories.

Type A refers to the lion's share of social capital research performed in organizational settings. Social structure and outcomes are both considered at the level of the individual. A typical example is Burt (1992), showing that managers with disconnected networks achieve faster promotions to managerial positions. Other typical examples include studies on how people mobilize their array of direct and indirect relationships to accomplish personal goals such as finding jobs and achieving upward mobility (Granovetter 1973). Although he doesn't use the term social capital explicitly, Granovetter's (1973) argument is entirely about social capital; the mechanism Granovetter discusses is as follows. The friendship ties people maintain provide them with alters who have the goodwill (and good will) to provide them with valuable information, for example, about possible interesting job opportunities. Friendship ties tend to vary in strength; ego may have drinks with some friends every night of the week, whereas ego interacts with other friends only once or twice a year. The former set of friends (who have strong ties with ego) is likely to all have the same friends as ego, whereas the latter set of friends (who are connected to ego by weak ties) will likely socialize with many others than ego does. The larger the set of weak friendship ties an individual has, the more varied the information that will reach ego. Granovetter's work showed that individuals with many weak ties are more likely to find a suitable job or be upwardly mobile. Thus, the *social network* of individuals provides them with the *social capital* that makes available *output resources* (in this case, information) that bestows them with the *value* of increased opportunities on the job market.

Social capital research of Type B refers to the benefits a collective (e.g., a company) draws from networks of individuals. For example, trustworthy relationships between employees of a firm and the employees of a bank may make it easier for that firm to secure a loan from that bank. Law firms, accounting firms, and consulting agencies considerably draw upon the networks senior consultants have with their clients to bring business to the firm. In firms, successful innovation often requires the firm to bring information about the market into the firm as well as new technology and other resources (such as financial resources). In many firms, much of this is achieved through ties that individual employees maintain with actors outside the firm; they then distribute these resources to the places in the firm that might need them. The effectiveness of this process for the firm highly relies on the number and quality of the ties that these employees maintain and the goodwill and knowledgeability of their network partners.

Type C research refers to situations where networks of corporations or other groups confer advantages to individuals. Examples include joint research and development projects between two firms that create new job opportunities for the individuals working in these positions or that produce the knowledge necessary to do one's job better and become eligible for a bonus or promotion. The networks of consulting firms can assist (junior) consultants in bringing in new projects, and the ties maintained by an academic department can be of great use to a junior academic in need of specialized expertise or research funding.

Finally, in Type D, organizations draw advantages from their own interorganizational networks. Joint venture relationships or joint marketing efforts, allowing for economies of scale or increased expertise, are examples of this type. Through interfirm relations, firm can gain timely and affordable access to new technology. For example, high prestige semiconductor firms tend to establish license alliances in which they gain the rights to produce and sell the proprietary technologies of competing organizations. It is because of their ability to certify the initiatives of other organizations (startups, in particular) that high prestige firms will gain access to the endeavors of others. The correspondence between prestige and access implies that prestigious firms enjoy a powerful positional advantage.

Whether social capital is seen from the group-level or the individual level, Lin (2001) contends that all scholars remain committed to the view that, at the heart of things, it is the interacting members who make the maintenance and reproduction of social capital possible.

The Dark Side

Even though the predominant sentiment is that social networks are beneficial to individuals and groups, there is an increasing realization that there are profound negative sides to them as well. This is often referred to as "negative social capital," "the dark side of social capital," or, more in keeping with the "capital" part of the concept's name, "social liability" (Gabbay and Leenders 2001; Leenders and Gabbay 1999).

An example is violent or criminal gang activity that is encouraged through the strengthening of intragroup relations: this brings social liability to society. Alternatively, membership in certain groups may require individuals to submit to group norms and obligations that reduce individual autonomy (Portes 1998). Social capital in tight-knit communities may create free-riding problems and hinder entrepreneurship. Strong solidarity with ingroup members may overembed the actor in the relationship, which reduces the flow of new ideas into the group, resulting in parochialism and inertia (Adler and Kwon 2002; Gargiulo and Benassi 1999).

Social liability shows why it pays off to explicitly relate social capital to goals or other outcomes: the same mechanism can provide outcomes that are productive for one goal but harmful for the achievement of another. For example, dense ties in a network of an R&D team provide the team members with quick access to knowledge, assisting the team in being efficient (social capital). However, research also shows that this comes at the expense of reduced levels of team creative performance, hampering the R&D team's equally important goal of being truly innovative (social liability).

Similarly, social structures can be beneficial to the fulfillment of a particular outcome at one point in time but become a liability later. An example of Type A research, Gargiulo and Benassi (1999) showed that relational structures that were helpful to managers in the past, later increased the number of coordination failures for which they were responsible. The network had become a constraint, impeding their performance. In his study on network marketing, Gabbay (1997) found that, for some entrepreneurs, strong ties combined with structural holes were beneficial at the initial stages of their business but were harmful for future expansion.

Grapevines – informal, person-to-person communications network of employees which are not officially sanctioned by the organization – are sources of rumors and gossip that spread quickly throughout an organization. Management decisions may travel through grapevines days ahead of their official announcement. Because they feel threatened by it, managers often try to suppress the grapevine but find themselves confronted by a nearly impossible exercise. Grapevines and gossip networks, examples of individual-level social structure, can have detrimental effects on organization-level well-being and productivity (Type B).

Another source of potential organizational social liability is related to the resilience of personal networks. Managers in charge of (re)designing business processes often experience difficulties in breaking through the power structures that exist

among the firm's employees. As a result, many attempts to redesign organizational processes fail or can only be implemented after long and painful struggles between higher management and employees (Type C).

At the fully organizational level (Type D in Fig. 4), long-standing relationships with customers may stifle the firm by monopolizing a disproportionate share of its resources, inhibiting the firm from forming relationships with alternative customers. Similarly, dense long-lived ties with other firms often effectively create blinders, reducing the firm's ability to see new (technological) developments that occur outside the firm's constrained field of vision.

Is It Capital?

By now, it will be clear to the reader why "social" is part of the term "social capital." An article that addresses the history and roots of the social capital concept, starting from its use in nineteenth-century political economics, also has to spend at least a few lines on the question whether social capital is "capital," a question that increasingly appears in academic social capital articles. It seems to me that there are two valid answers to this question, the second perhaps being the most to the point (except, perhaps, to economists).

The first answer addresses the nature of capital itself. Social capital does exhibit a number of characteristics that distinguish it from other forms of capital. Unlike physical capital, social capital can accumulate as a result of its use. Moreover, unlike financial capital, social capital erodes when it is not used. On the other hand, similar to other forms of capital, social capital is not costless to produce, requiring an investment that can be significant (Adler and Kwon 2002; Knoke 1999). The trusting relationships among the members of a sports club or professional organization can require years of meeting and interacting to develop (Grootaert and Van Bastelaer 2001). In addition, like all other forms of capital, social capital is a long-lived asset into which other resources can be invested, with the expectation of future benefits. It is also both appropriable and convertible and can be a substitute for or complement other resources. Based on these arguments, Adler and Kwon (2002) conclude that social capital "falls squarely within the broad and heterogeneous family of resources commonly called capital."

The second potential answer to the question is: "who cares"? The key attribute of capital is that it is an accumulated stock from which a stream of benefits flows. The view that social capital is an asset – that is, that it represents genuine capital – means that it is more than just a set of social organizations or social values. On the output side, it shows how things are getting done in society. On the input side, it shows that it requires a genuine investment to make society prosper. This is important, both from a conceptual and societal point of view. Aside from the intellectual joy the "is it capital" debate can undoubtedly provide to academists at cocktail parties, the social capital literature is probably best served by spending our efforts on developing better ways to measuring social capital and on improving the empirical and analytical rigor in social capital papers than by a debate about the semantic accuracy of the concept's name.

Future Directions: Challenges

Although it was first used in the nineteenth century, social capital is still relatively immature as a concept, especially in its contemporary use. Its rapid proliferation has allowed a diversity of approach, definition, measurement, and causal logic (Schuller et al. 2001). Social capital is used in an extraordinarily wide range of disciplines. One consequence of this is that it is still largely unclear how social capital should be measured. Where such a diversity of definition exists, it is inevitable that an equivalent heterogeneity of measure is used (Schuller et al. 2001). A main challenge for the concept is for its users to develop useful and analytically sound measures of social capital (and of the other parts of the social capital model). It is unlikely that any time soon a measure of social capital will (or can) be developed that is acceptable or useful to the wide range of contemporary

social capital analysts, but one would hope that at least the empirical and statistical rigor of social capital research would be improved in the near future.

A second challenge, one that likely makes the former challenge an even harder one, relates to the changing nature of social relations in modern life. The works of Bourdieu, Coleman, and much of Putnam's work addressed social relations in a "bricks-and-mortar" world, in which social relations were largely created and maintained in a face-to-face manner. Especially over the last decade, social relations increasingly reside in cyberspace as well, and our social environment is transformed into a "clicks-and-mortar" world. Increasingly, social ties are built or maintained on Facebook, LinkedIn, Twitter, and other electronic platforms that are now frequently referred to as "social networks." One can easily see that the claim that social capital is declining can be refuted if one goes beyond the traditional interpersonal offline networks and includes network ties that live in cyberspace. However, can cyber ties be seen as equal to physical ties? At the very least, the answer to this question will be different for different goals. We may have to reconsider findings from earlier research. For example, weak ties may no longer provide such a strong informational advantage when most job openings can easily be found by a single click of the mouse. It is conceivable that investment in online social capital is lower than the investment needed to build offline social capital; at the same time, the social capital (or social liability) that one draws from online ties may also differ from those drawn from offline relationships. At any rate, social capital researchers cannot deny the increasing and pervading importance of cyberrelations if they are to study social capital in today's society and will need to rethink their causal models and social capital measures.

Finally, an important challenge for the social capital literature is how to deal with temporal issues. Social networks are dynamic, those residing in cyberspace perhaps even more so. With social relations being dynamic, it is inevitable that social capital and its outcomes will experience dynamics as well. In general, there is a dearth of time-based theories in the social sciences. Statistical models for network dynamics are now publicly available. However, appropriate theories of network dynamics are still lacking. For the rest of the social capital framework of Fig. 3, both theory and statistical models are missing almost entirely.

Conclusion

One of the key merits of social capital is that it shifts the focus of analysis from behavior by individual agents to the pattern of relations between agents (and their environment). Closely linked to this is that the social capital concept links micro-, meso-, and macrolevels of analysis (Coleman 1990; Schuller et al. 2001).

In addition, social capital research addresses issues that are important to everyone, everyday. It addresses questions related to interpersonal trust, quality of relationships in different contexts, and about equality and inequality in society. Even in academic fields like sociology or economics, there are only few topics that so consistently address issues that are of direct importance to every human being.

A successful future for the social capital literature requires an interdisciplinary approach that bridges some of the current different disciplinary perspectives. Political scientists, sociologists, and anthropologists tend to approach the concept of social capital through analysis of norms, networks, and organizations. Economists, on the other hand, tend to approach the concept through the analysis of contracts and institutions and their impacts on the incentives for rational actors to engage in investments and transactions. Each of these views has merits and the overarching challenge is to take advantage of the complementarities of the different approaches (Grootaert and Van Bastelaer 2001, p. 8). In this manner, we can turn the current proliferation of approaches, often seen as a weakness of the concept and threat to its viability, into a strength, providing the social capital literature with a bright and productive future.

Cross-References

- ▸ Human Behavior and Social Networks
- ▸ Online Communities
- ▸ Personal Networks: The Intertwining of Ties, Internet and Geography
- ▸ Structural Holes
- ▸ Trust in Social Networks

References

Adler PS, Kwon S-W (2002) Social capital: prospects for a new concept. Acad Manag Rev 27:17–40

Boggs C (2001) Social capital and political fantasy: Robert Putnam's "bowling alone". Theory Soc 30:281–297

Bourdieu P (1986) The forms of capital. In: Richardson JG (ed) Handbook of theory and research for the sociology of education. Greenwood Press, New York, pp 241–58

Bourdieu P, Wacquant LJD (1992) An invitation to reflexive sociology. University of Chicago Press, Chicago

Burt RS (1992) Structural holes: the social structure of competition. Harvard University Press, Cambridge

Burt RS (1997) The contingent value of social capital. Adm Sci Q 42:339–365

Clark JB (1885) The philosophy of wealth. Ginn, Boston

Coleman JS (1988) Social capital in the creation of human capital. Am J Sociol 94:95–120

Coleman JS (1990) Foundations of social theory. Belknap Press of Harvard University Press, Cambridge

Farr J (2004) Social capital. Pol Theory 32:6–33

Gabbay SM (1997) Social capital in the creation of financial capital: the case of network marketing. Stipes Publishing, Champaign

Gabbay SM, Leenders RTAJ (1999) CSC: the structure of advantage and disadvantage. In: Leenders RTAJ, Gabbay SM (eds) Corporate social capital and liability. Wolters-Kluwer Academic Publishers, New York, pp 1–14

Gabbay SM, Leenders RTAJ (2001) Social capital of organizations: from social structure to the management of corporate social capital. In: Gabbay SM, Leenders RTAJ (eds) Social capital of organizations, research in the sociology of organizations. JAI Press, New York, pp 1–20

Gargiulo M, Benassi M (1999) The dark side of social capital. In: Leenders RTAJ, Gabbay SM (eds) Corporate social capital and liability. Wolters-Kluwer Academic Publishers, Norwell, pp 298–322

Granovetter MS (1973) The strength of weak ties. Am J Sociol 78:1360–1380

Grootaert C, Van Bastelaer T (2001) Understanding and measuring social capital: a synthesis of findings and recommendations from the social capital initiative. In: World bank social capital initiative working paper 24. http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTSOCIALDEVELOPMENT/EXTSOCIALCAPITAL/0,contentMDK:20194767~menuPK:418848~pagePK:148956~piPK:216618~theSitePK:401015,00.html

Halpern D (2005) Social capital. Polity Press, Cambridge

Knoke D (1999) Organizational networks and corporate social capital. In: Leenders RTAJ, Gabbay SM (eds) Corporate social capital and liability. Wolters-Kluwer Academic, Norwell, pp 17–42

Leenders RTAJ, Gabbay SM (1999) Corporate social capital and liability. Kluwer Academic, Boston

Lemann N (1996) Kicking in groups. Atl Mon (10727825) 277:22–26

Lin N (2001) Social capital: a theory of social structure and action. Cambridge University Press, Cambridge

Marshall A (1890) Principles of economics. Macmillan, London

Marx K (1867) Das Kapital: Kritik der Politischen Ökonomie. Marx-Engels Werke. Dietz Verlag, Berlin

McLean SL, Schultz DA, Steger M (2002) Social capital: critical perspectives on community and "bowling alone". New York University Press, New York

Portes A (1998) Social capital: its origins and applications in modern sociology. Annu Rev Sociol 24:1–24

Putnam RD (1995) Bowling alone: America's declining social capital. J Democr 6:65

Putnam RD (2000) Bowling alone: the collapse and revival of American community. Simon & Schuster, New York

Schuller T, Baron S, Field J (2001) Social capital: a review and critique. In: Baron S, Field J, Schuller T (eds) Social capital: critical perspectives. Oxford University Press, Oxford, pp 1–38

Sidgwick H (1883) The principles of political economy. Macmillan, London

Social Capital of Managers

- ▸ Managerial Networking

Social Classification

- ▸ Folksonomies

Social Communication Network, Case Study

Bo Xu, Deqing Yang, Qi Liu, and Yanghua Xiao
School of Computer Science, Fudan University, Shanghai, China

Synonyms

Call network; Communication network; Interaction network; Mobile network; Social interaction

Glossary

SNA Social network analysis
SP Shortest path
DC Degree centrality
BC Betweenness centrality
CC Closeness centrality
IM Instant messaging

Definition

Social network is formally defined as a set of social actors that are connected by one or more types of relations (Wasserman and Faust 1994). Social actors can be individuals, groups, organizations and even any units that can be connected to other units such as web pages, blogs, emails, instant messages, families, journal articles, neighborhoods, classes, sectors within organizations, positions, or nations (Furht 2010).

Social communication network is one of the most important social networks. In a social communication network, social actors are mostly persons, and the relationship between them is established for the purpose of communication. In a social communication network, social actors use communication tools such as mobile phones, instant messenger softwares (MSN messenger, Google Talk, etc.), and so on to communicate with each other. Social communication networks can be classified into different categories in terms of the client communication tools and the network infrastructure. Typically, those running on telecom network with mobile phone as clients include mobile call network and short message network. Those running on Internet with PC or smart phone as clients include instant messaging network such as MSN, QQ, and Skype. Another typical communication network on the Internet is email network.

Introduction

Social network analysis (SNA) is one of significant steps towards understanding the behavior of actors in the network. The first step of SNA is characterizing the structural properties of the networks. In general, different structural properties imply different principles of users' behaviors. Understanding user behavior is critical for the success of applications built upon these networks. Social communication networks underlie our daily life. All of us are living in social communication networks. Thus, our communication behavior pattern is certainly embedded in these social communication networks. Hence, SNA on social communication networks is of special importance for user behavior understanding. After understanding the network properties of these networks, the next key step is leveraging these properties for a successful application.

The purpose of this article is twofold. First, we showcase the common structural properties of social communication networks. Second, we showcase the applications on these networks.

Key Points

The structural properties of social communication networks in general can be explored from the following aspects: (1) social ties, (2) node strengths, (3) shortest paths and diameter, (4) centrality, and (5) assortativity. We show that social communication networks exhibit similar properties to a general social network but with some exceptions. For example, the degree of a typical

social network follows power-law distribution. But in social communication network, the degree distribution follows Double Pareto-Lognormal (DPLN) distribution.

We have witnessed many successful applications on social communication networks including (1) economic development evaluation, (2) spammer detection, and (3) email worms defense. The diversity of social ties in social communication networks is positively correlated to economic development (Eagle et al. 2010) which allows us to evaluate the regional economical development by the social tie diversity of inhabitants in the region. In general, social communication network reflects people's interaction in social lives, but it may also include some spammers who generate garbage information. When an email user clicks a worm program in the attachments of a worm email, the worm program will find all the email addresses stored on this computer and send its email address to other users. The worm program is called "email worms." By identifying the structure of the network, the state-of-the-art system can successfully resist against the email worms.

Historical Background

Social communication network analysis has been studied for a long time. It dates back to the experiment that was made by social psychologist Stanley Milgram in 1967. He selected two target persons and found some volunteers to let them send the letter from one target person to another by using their own social relationships. Some letters were successfully delivered from one target person to another target person. He found that the average distance of the success delivery is 6, implying that any two persons are linked to each other on average via a chain with "six-degrees-of-separation." However, the experiment data is very small; the number of successful experiments is only 300. Hence, the reliability of the experimental result is an issue, which can be solved by statistical analysis on large-scale social communication network available nowadays.

Structural Measures on Social Communication Networks

In this part, we will review some important aspects to characterize social communication networks, including tie strength, node strength, shortest paths, centrality, and assortativity.

Tie Strength

Edges in a social communication network represent the social ties between two social actors. Typical social ties in social communication networks include sending messages, calling, and sending email.

The strength or weight of a tie between person i and person j, denoted by *tie(i, j),* can be quantified by the aggregate time that i and j spent on the communication with each other or by the total number of communication times between them. These weights are denoted by w_{ij}^D (total duration of communication) and w_{ij}^N (total number of communication times), respectively.

Node Strengths

Based on tie strengths, node strengths can be defined as $s_i^N = \sum_{j \in N(v_i)} w_{ij}^N$ or $s_i^D = \sum_{j \in N(v_i)} w_{ij}^D$, where $N(v_i)$ is the neighbors of i. s_i^N represents the aggregate number of communication times. s_i^D represents the aggregate communication duration.

Shortest Paths

The shortest path between two nodes is one of simple paths with minimal length between them. The *diameter* of a network is the longest shortest path length over all node pairs. In general, it is hard to calculate the exact diameter on a large network due to its quadratic computational complexity. Diameter can be approximated with affordable cost (Magnien et al. 2009). It was found that the average shortest path on mobile social network of a city in China is 5.75 (Dong et al. 2009), which confirms the "six-degrees-of-separation" theory.

Centrality

Centrality measures the importance of users in social communication networks. There are three

typical centrality measures: degree centrality, betweenness centrality, and closeness centrality.

1. Degree Centrality
   Degree of a node is the number of its connections. In social communication networks, it represents the number of contacts the user has. Hence, degree is a natural choice to measure the importance or the activity of the user.
2. Betweenness Centrality
   The betweenness of a vertex i is defined as the fraction of shortest paths that pass through i. More specifically, it is defined as

$$b_i = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

   where σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(i)$ is the number of those paths that pass through i.
3. Closeness Centrality
   The closeness of a node is the inverse of the average distance in the network from the node to all other nodes. Closeness reveals how long it takes for information to spread from one individual to others in the network. High-scoring node tends to have shorter shortest paths to other nodes in the network.

Assortativity

A network is assortatively mixing if the nodes in the network that have many connections tend to be connected to other nodes with many connections (Newman 2002). That is, people with many friends are connected to others who also have many friends. This gives rise to degree-degree correlations in the network, implying that the degrees of two adjacent nodes are not independent (Onnela et al. 2007). The average nearest neighbors degree of a node v_i is $k_{nn,i} = \frac{1}{k_i} \sum_{j \in N(v_i)} k_j$, where k_j is the degree of v_j. By averaging this over all nodes in the network of a given degree k, one can calculate the average degree of the nearest neighbors for degree k, denoted by $< k_{nn}|k >$ (Pastor-Satorras et al. 2001). The network is assortatively mixing if $< k_{nn}|k >$ increases with k and disassortatively mixing if it decreases as a function of k. On edge-weighted networks, weighted average nearest neighbor degrees are also used to characterize strength-strength correlations. There are two typical weighted versions: $k^N_{nn,i} = \frac{1}{s^N_i} \sum_{j \in N(v_i)} w^N_{ij} k_j$ and $k^D_{nn,i} = \frac{1}{s^N_i} \sum_{j \in N(v_i)} w^D_{ij} k_j$. It was found that in a typical social communication network, the degrees of two adjacent nodes are strongly correlated, while the strengths of two adjacent nodes in most cases are not (Onnela et al. 2007).

Mobile Call Network

Mobile phones are widely used in our daily lives. According to the International Telecommunication Union, at the end of 2011, there were 6 billion mobile subscriptions, which accounts for about 87 % of the world population. In a mobile call network, each node is a mobile phone user, and each edge between two users means that they have at least one mobile call. There are several interesting structural properties and application in mobile call network.

Distribution

Power-law distribution is frequently observed in the real world. For example, a common property of many large real networks is their power-law degree distribution. This feature was found to be a consequence of two generic mechanisms: (1) the network grows continuously by the addition of new vertices and (2) new vertices attach preferentially to well-connected vertices. Several recent works (Saramaki and pekka Onnela 2007; Nanavati et al. 2006) studied typical mobile call networks and found that their distributions with respect to degree and many other measures also follow the power-law distribution. However, the study on a larger mobile call network which consists of a million users and a hundred million calls shows that most distributions of this network significantly deviate from power-law and lognormal distribution but fit better- to a less-known distribution: Double Pareto-Lognormal (DPLN) distribution (Seshadri et al. 2008). The distributions following DPLN include the number of phone calls per customer, the total talk time per customer, and the distinct number of calling

partners per customer. Their study further reveals that the DPLN distributions can be consistently observed for networks in different snapshots.

Social Tie Diversity

Social networks form the backbone of social and economic life. Theoretical work suggests that the structure of social relations between individuals may affect personal life or economic development. For example, it was found that weak acquaintance relationships rather than close friendships are more helpful to find a job (Granovetter 1973, 1983). This is well known as weak tie theory. Eagle et al. (2010) found that the economical development is positively correlated to the diversity of social ties in a mobile call network. They use the following steps to study this relationship between network structure and economic development.

Step 1. Mobile network construction. They collected the national mobile call logs on August in 2005 in the UK. The data contains more than 90 % of all mobile phones, which cover more than 99 % of the populations and business landlines in the country. The constructed network consists of 65 million nodes and 368 million edges.

Step 2. Measuring diversity of social ties. They use Shannon entropy (Shannon 2001) to quantify diversity. They propose two diversity metrics: social diversity and spatial diversity. Social diversity of person i is defined as

$$D_{social}(i) = \frac{-\sum_{j=1}^{k} p_{ij} \log(p_{ij})}{\log(k)}$$

where k is the number of i's contacts and p_{ij} is the proportion of i's total call volume that involves j. Spatial diversity of person i can be similarly defined as

$$D_{spatial}(i) = \frac{-\sum_{a=1}^{A} p_{ia} \log(p_{ia})}{\log(k)}$$

where A is the total number of telephone exchange areas and p_{ia} is the proportion of time i spends communicating with a-th of exchange area.

Step 3. Analysis. In this step, they compare the social tie diversity to economic development measured by IMD (Index of Multiple Deprivation) of UK in 2004 UK. IMD is a composite measure of relative prosperity of 32,482 communities encompassing the entire country, based on income, employment, education, health, crime, housing, and the environmental quality of each region. They found that the ranks of both social and spatial network diversity scores are positively correlated to IMD rank. For example, in Stoke-on-Trent, one of the least prosperous regions in the UK has one of the lowest diversity scores in the country.

Short Message Network

Short messages are sent from one mobile phone to another. This inherently is a network with users as vertices and edges as message-sending relationships. Short message has been one of the fastest-growing telecom value-added services worldwide. Due to its own characteristics, there are some special applications on it. In this section, we will showcase an SMS anti-spam system on this network.

As we know, short message service has greatly changed our lives. On some occasions, we prefer to short message rather than phone call to communicate with others. However, an accompanying problem is that message spam has also grown fast. Unsolicited and unwanted commercial advertisements may be sent as messages to mobile phone users. In some cases, fraud messages and rumor messages may be sent over the network.

Many solutions have been proposed to overcome this problem. Wang et al. (2010) uses spammers' behavior features and temporal features to detect spammers.

To distinguish legitimate users from spammers, they summarize many behavior patterns of spammers and normal users. In general, spammers tend to send a large number of messages to legitimate users. The legitimate users in general will not reply to an unknown phone number. Legitimate user's messaging targets are probably their friends, while spammer's messaging

targets are mostly strangers. Furthermore, in a given period, a spammer usually sends only one message to one recipient. These social features can be quantified by out degree, mean weight on out edges, variance of weight on out edges, one weight ratio, reply ratio, partner ratio, and edge ratio (Wang et al. 2010).

Spammers can be divided into fraudulent senders and unauthorized advertisement agencies. They use temporal patterns to distinguish fraudulent senders from advertisement agencies. Fraudulent senders always submit a large number of messages in a short time period. Unauthorized advertisement agencies submit messages at a low frequency; and legitimate users submit messages at a medium frequency.

Instant Messaging Network

Instant messaging programs, such as Microsoft MSN, ICQ, Yahoo Messenger, Tencent QQ, Skype, are very widely used in personal and business communications. A recent report (Leskovec and Horvitz 2008) estimated that approximately 12 billion instant messages are sent each day. These instant messaging tools imply instant messaging networks, where each vertex is a user, and each edge represents the contact relationship between users. Unlike other social communication networks, people tend to use informal language, loose grammar, abbreviations, and minimal punctuation in instant messages.

As a typical instant messaging network, MSN network was investigated in Leskovec and Horvitz (2008). They use anonymized data capturing a month of high-level communication activities in MSN system. They have found the following interesting facts.

First, they found that birds of the same feather flock together. People with similar properties tend to communicate with each other. For example, people with similar ages, the same languages, and geographically close locations tend to communicate with each other more frequently and longer. One of exceptions is gender. People tend to converse more frequently and with longer durations with those with opposite gender.

Second, they found that the instant messaging network is well connected and well clustered with 99.9 % of the users belonging to the largest connected component, and the average clustering coefficient is 0.137. The average shortest path length among Messenger users is 6.6, which is half a link more than "6-degrees-of-separation."

Third, they found that instant messaging network is very robust against intentional attack. They used different attack measures, such as average number of sent messages per user's conversation, average duration of user's conversation and so on, and simulate the intentional attach on the network.

Email Network

Email is a highly effective communication tool. It is inexpensive and only requires Internet connection. Hence, email network is one of most important social communication networks. However, email network is prone to some security issues.

"Email worms" (Zou et al. 2004) are one of the major Internet security threats for our society. There are many different types of worms (Weaver et al. 2003). One typical email worm works as follows: When an email user clicks a worm program in the attachments of a worm email, the worm program will find all the email address stored on this computer and sends the copies of itself to other users. Email worms spread on the email network, which is one of great security challenge to manage email networks.

Newman et al. (2002) found that there is little that computer system administrators can do to control the spread of a virus in the world at large through the study on the email network reconstructed from emails in a university. There are two main methods to defend against the "email worms": random vaccination and targeted vaccination. According to Newman's results, random vaccination has little effect on virus spread, while targeted vaccination seems pretty good. The effectiveness of vaccination strategy obviously depends on the network structure. Zou et al. (2004) investigates the influence of three topologies: power law, small world and random

graph. They found that on power-law topology, email worms spread more quickly, and targeted vaccination is more effective.

Key Applications

In general, analysis on social communication networks allows us to understand users' communication behavior. Specifically, these networks are helpful in the following applications. First, they can be used for the evaluation of regional economical development. The positive correlation between economical development and diversity of social ties in mobile call networks can be used for this application. Second, they can be used for spammer detection. Spammers have different features in the short message network, which allows us to detect spammers. Third, they can be used for friend recommendation. In instant messaging networks, users with similar properties tend to communicate with each, which can be used for friend recommendation. Finally, they can be used for resisting email worm attacks.

Future Directions

These social communication networks allow us to understand human behavior better. However, previous research on social communication network can be extended in many directions. First, social communication networks are inherently evolving. Investigation on the evolution pattern is more important in many real applications. Second, social communication networks contain abundant heterogeneous information. For example, users in instant messaging networks have much profile information. How to employ the heterogeneous information for the analysis of these networks is one of promising direction.

Cross-References

▸ Mobile Communication Networks

References

Dong Z, Song G, Xie K, Wang J (2009) An experimental study of large-scale mobile social network. In: Proceedings of the 18th international conference on World wide web. ACM, pp 1175–1176

Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. Science 328(5981):1029–1031. doi:10.1126/science.1186605

Furht B (2010) Handbook of social network technologies and applications. Springer, New York

Granovetter M (1973) The strength of weak ties. Am J Sociol 78(6):1360–1380

Granovetter M (1983) The strength of weak ties: a network theory revisited. In: Sociological theory, Wiley, New Jersey, pp 201–233

Leskovec J, Horvitz E (2008) Planetary-scale views on a large instant-messaging network. In: Proceedings of the 17th international conference on World Wide Web, WWW '08, Beijing, China. ACM, New York, pp 915–924

Magnien C, Latapy M, Habib M (2009) Fast computation of empirically tight bounds for the diameter of massive graphs. J Exp Algorithmics 13:10:1.10–10:1.9

Nanavati AA, Gurumurthy S, Das G, Chakraborty D, Dasgupta K, Mukherjea S, Joshi A (2006) On the structural properties of massive telecom call graphs: findings and implications. In: Proceedings of the 15th ACM international conference on Information and knowledge management, CIKM '06. ACM, New York, pp 435–444

Newman MEJ (2002) Assortative mixing in networks. Phys Rev Lett 89(20):208701

Newman M, Forrest S, Balthrop J (2002) Email networks and the spread of computer viruses. Phys Rev E 66(3):035101

Onnela JP, Saramaki J, Hyvonen J, Szab G, de Menezes MA, Kaski K, Barabsi AL, Kertsz J (2007) Analysis of a large-scale weighted network of one-to-one human communication. New J Phys 9(6):179

Pastor-Satorras R, Vazquez A, Vespignani A (2001) Dynamical and correlation properties of the internet. Phys Rev Lett 87(25):258701

Saramaki J, pekka Onnela J (2007) Structure and tie strengths in mobile communication networks. Proc Natl Acad Sci 104(18):7332–7336

Seshadri M, Machiraju S, Sridharan A, Bolot J, Faloutsos C, Leskove J (2008) Mobile call graphs: beyond power-law and lognormal distributions. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, Las Vegas, Nevada, USA, KDD '08. ACM, New York, pp 596–604

Shannon C (2001) A mathematical theory of communication. ACM SIGMOBILE Mobile Comput Commun Rev 5(1):3–55

Wang C, Zhang Y, Chen X, Liu Z, Shi L, Chen G, Qiu F, Ying C, Lu W (2010) A behavior-based sms antispam system. IBM J Res Dev 54(6):3
Wasserman S, Faust K (1994) Social network analysis: methods and applications. Structural analysis in the social sciences. Cambridge University Press, Cambridge
Weaver N, Paxson V, Staniford S, Cunningham R (2003) A taxonomy of computer worms. In: Proceedings of the 2003 ACM workshop on Rapid malcode. Washington, DC, USA, ACM, pp 11–18
Zou C, Towsley D, Gong W (2004) Email worm modeling and defense. In: Proceedings, 13th International Conference on Computer communications and networks, ICCCN 2004. Chicago, Illinois, IEEE, pp 409–414

Social Computing

▸ Collective Intelligence for Crowdsourcing and Community Q&A
▸ Web Communities Versus Physical Communities

Social Content Search

▸ Social Web Search

Social Data Analysis

▸ Temporal Analysis on Static and Dynamic Social Networks Topologies

Social Engineering

▸ Reconnaissance and Social Engineering Risks as Effects of Social Networking

Social Engineering/Phishing

Jingrui He[1] and Yada Zhu[2]
[1]Computer Science, School of Engineering & Science, Hoboken, NJ, USA
[2]Computer Science, School of Engineering & Science, IBM T. J. Watson Research Center, Hoboken, NY, USA

Synonyms

E-mail; Fraud; Information; Internet; Social network; Suspicious

Glossary

E-mail Spam Unsolicited e-mails for the purpose of advertisement or committing fraud
Phishing Electronic fraud based on social engineering
Phisher Fraudsters who commit phishing crimes
Phishing Site Websites created by phishers to steal sensitive information from users
Anti-phishing Efforts taken from multiple perspectives to combat phishing crimes
Machine Learning The design and development of algorithms that takes as input empirical data and outputs patterns and predictions for future data

Definition

Nowadays, phishing has gradually become a popular type of electronic fraud that makes use of social engineering to steal sensitive information from users such as user name, password, bank account number, and credit card details (http://www.indiana.edu/~phishing/?about; http://en.wikipedia.org/wiki/Phishing#Early_phishing_on_AOL). Phishing can be carried out via e-mails, instant messages, phone calls, text messages, etc. (http://www.indiana.edu/~phishing/?about; http://en.wikipedia.org/wiki/Phishing#Early_phishing_on_AOL), where

phishers pretend to be a trustworthy party in an attempt to lead the users to disclose the above sensitive information. Based on the collected information, the phishers can withdraw money from the accounts, causing significant financial loss.

To combat phishing crimes, people are making efforts from various aspects. For example, there have been constant efforts towards raising public awareness of this rapidly proliferating cyber crime, so that users are not easily spoofed into giving up sensitive information; researchers from academia and industry have been tracking the recent developments of phishing techniques with the hope of catching them in time; there has also been efforts from the government by filing law suits against phishers and proposing laws to fight this crime.

The purpose of this article is twofold. The first is to introduce the evolution process of the phishing techniques, with an emphasis on its current status, and the second is to look into the techniques for anti-phishing, which shed lights on the future generation of phishing methods.

Introduction

According to http://en.wikipedia.org/wiki/Phishing#Early_phishing_on_AOL, the word "phishing" came into use as a variant of "fishing" in mid-1990s, which is connected to "baits" used therein to induce the users into disclosing sensitive information. Also, the "ph" spelling was used to link phishing scams with some underlying communities, such as the hackers known as "phreaks" (http://www.phishing.org). A typical example of phishing is a fake e-mail masqueraded to come from a bank, which asks the user to follow an embedded link to a phishing site (which often highly mimics the authentic website) and give up his/her bank account information. Another example of phishing takes place in an online chat session, where the phisher pretends to be an agent from the online vendor and requires the sensitive information from the user. In both cases, the baits are the masqueraded identity of the e-mail sender, the embedded link, and the online agent. If the user is tricked into believing this identity and reveals his/her account information, he/she will suffer significant financial loss.

Due to the severe challenge posed by phishing, recent years have seen rapidly growing efforts in anti-phishing. To be specific, in academic, many universities have set up groups devoted to anti-Phishing research, such as the Anti-Fishing Group at Indiana University (http://www.indiana.edu/~phishing/), the Institute for Security Technology Studies at Dartmouth College (http://ists.dartmouth.edu/), the Center for Education and Research in Information Assurance and Security at Purdue University (http://www.cerias.purdue.edu), the Stanford Security Laboratory (http://theory.stanford.edu/seclab/), and the Cylab Usable Privacy and Security Laboratory (http://cups.cs.cmu.edu), to name a few. In industry, a variety of anti-phishing solutions have been proposed, such as the suite of solutions provided by the Anti-Phishing Working Group, the phishing protection services from Dell SecureWorks and RSA, anti-phishing toolbars from eBay, Netcraft, and EarthLink, as well as the anti-phishing filters in Firefox, Internet Explorer, Google Chrome, etc.

Key Points

In the rest of this chapter, we first review the history and the current status of phishing, followed by a discussion of anti-phishing techniques.

Historical Background

According to http://www.phishing.org, the very first phishing attacks happened on American Online (AOL) on January 2, 1996. At that time, phishers sent messages to users through AOL instant messengers and e-mail systems, requesting the users to verify their accounts or to confirm their billing information. Many users gave up the account information upon such requests and

later experienced financial loss. In response to such phishing crimes, AOL and later many banks and online payment systems include warnings in their e-mails and instant messenger chat windows preventing the users to disclose the sensitive information in such scenarios.

The phishing crimes quickly ramped up since late 2003, with the registration of domains suggesting legitimate sites such as eBay and PayPal, which were used as phishing sites. Phishers then sent out e-mails to the users, leading them to these phishing sites and asking them to update their credit card information. Later the phishing techniques evolved into using pop-up windows of online banks to gather account information from the users, which was proven to be very effective in 2004.

According to the surveys by Gartner between 2005 and 2007, there was continuous increase in the percentage of phished web users in the USA (Herley and Florêncio 2008): 0.5 % in 2005, 1.05 % in 2006, and 2.18 % in 2007, resulting in huge financial loss of these victims. Similar as in the USA, the losses in the UK from web banking fraud, most of which are phishing fraud, almost doubled from 2004 to 2005 http://en.wikipedia.org/wiki/Phishing#Early_phishing_on_AOL.

The most up-to-date phishing techniques have been summarized in a variety of websites, such as http://en.wikipedia.org/wiki/Phishing#Early_phishing_on_AOL; http://www.phishing.org, and http://ists.dartmouth.edu/. These include e-mail spamming, web-based delivery (aka man-in-the-middle), instant messaging, Trojan hosts, link manipulation, key loggers, session hacking, system reconfiguration, content injection, phishing through search engines, phone phishing, and malware phishing (http://www.phishing.org). In addition, according to http://en.wikipedia.org/wiki/Phishing#Early_phishing_on_AOL and http://www.csionsite.com/2012/phishing/, phishing with specific targets are sometimes referred to as spear phishing and whaling. Furthermore, some smart phishers make use of advanced techniques to get around anti-phishing software (e.g., by making use of images instead of text http://en.wikipedia.org/wiki/Phishing#Early_phishing_on_AOL) or to gain trust of the potential victims (e.g., by including personal information obtained from social networks (Jagatic et al. 2007)). All the above evidence highlights the urgency of effective anti-phishing techniques.

Key Techniques

There are several different ways to combat phishing, including end-user education, legislation, and technology developed specially to fight against phishing. This section discusses the use of technology to protect against phishing website and e-mail.

Many technical solutions have been used to identify a web page as a phishing site, including blacklists (fraudulent sites), heuristics, page analysis, ratings, and their combinations. Blacklisting is a widely used approach in phishing detection mechanisms which maintains a list of known phishing websites and check websites against the list. This method has been implemented in numerous browser-integrated anti-phishing tools, such as Internet Explorer (IE), Google Safe Browsing (Schneider et al. 2007), NetCraft toolbar (NetCraft 2007), Firefox, and eBay toolbar (eBay 2013). The IE browser queries lists of blacklisted and white-listed domains from Microsoft servers and makes sure that the user is not accessing any phishing sites. The Google Safe Browsing uses blacklists of phishing URLs to identify phishing sites. The users are warned before they attempt to navigate to a known phishing site. Blacklists can be created using a set of classification rules based on previous phishing patterns, manually classified by the user or crowd sourced by users of a given service (Wilson and Argles 2011).

The effectiveness of a blacklist is determined by the coverage and quality of the list and the time it takes to include a phishing site. The quality shows the number of safe sites is falsely included into the list. Study shows that the URLs that have been verified by users tend to be classified with lower false-positive rate (Sun et al. 2010). Timeliness may be a challenge for

blacklisting because the average lifetime of phishing sites is only a few days or maybe a few hours for the low cost of creating a phishing site. Ludl et al. use 10,000 phishing URLs to test the effectiveness of the blacklists maintained by Google and Microsoft (Ludl et al. 2007). They demonstrate that blacklists provided by Google can recognize almost 90 % of live phishing sites, while IE contained only 67 % of them. They also find that on average it takes Microsoft 6.4 h to add an initially not blacklisted entry with a standard deviation of 6.2 h. For Google, it takes somewhat longer, 9.3 h on average with a standard deviation of 7.2 h. Sheng et al. (2009) use 191 fresh phish that are less than 30 min old to conduct two tests on eight blacklist-based anti-phishing toolbars. By hour 2, 63 % of phishing campaigns in their dataset are finished, but only 7.9 % of those phish are taken down. On average, 33 % of the websites are taken down within 12 h, around half are taken down after 24 h, and 27.7 % are still alive after 48 h. They conclude that blacklists are not effective when protecting users initially, as most of the tools catch less than 20 % of phish at hour 0. In addition, they show that blacklists are updated at different speeds and vary in coverage, as 47–83 % of phish appear on blacklists 12 h from the initial test. They also demonstrate that two tools use heuristics to complement their blacklists trigger catch significantly more phish initially than those using only blacklists. However, it takes a long time for phish detected by heuristics to appear on blacklists. Ramachandran et al. measure the effectiveness of 8 spam blacklists in real time by analyzing a 17-month trace of over 10 million spam messages collected at an Internet "spam sinkhole" and by correlating this data with the results of IP-based blacklist lookups (Ramachandran and Feamste 2006). In their study, whenever a host spammed their domain, they examine whether that host IP is listed in a set of Domain Name Service-based Blackhole Lists (DNSBLs) in real time. Their study indicates that about 80 % of the received spams are listed in at least one of eight blacklists, but even the most comprehensive blacklist has a false-negative rate of about 50 %.

Heuristic techniques analyze whether a page possesses suspicious behavior, e.g., examining the characters of the URLs and site's hostname. Since a phishing site is usually a mimicry of a legitimate site, page analysis or content-based method detects phishing by examining their similarity in terms of page properties, such as the number of password fields, the number of links, or the organization's logo. Using a search with the extracted keywords, it retrieves candidates for the legitimate site. If the page on the user's browser and the one of the candidate sites have the same domain name, the target site is judged legitimate, otherwise, a phishing site. Rating methods determine phish sites based on user ratings. Each site's rating is computed by aggregating all rates given for that site, with each user's rating of a site weighted according to that user's record of correctly identifying phishing sites. Heuristic, content analysis and rating are employed by numerous anti-phishing products, for example, Spoof Guard is based on heuristic and ratings; Calling ID toolbar is based on heuristic; Cloudmark Anti-Fraud toolbar is based on ratings; and EarthLink toolbar is based on the combination of heuristic and user rating.

Heuristics can detect attacks as soon as they are launched, without the need to wait for blacklists to be updated. However, attackers can design their attacks to avoid heuristic detection. In addition, heuristic approach may produce false positives, incorrectly labeling a legitimate site as phishing. On the other hand, page analysis techniques also have high false-positive rates due to the similarity between the phishing pages and the legitimate ones (Wilson and Argles 2011). User ratings might become meaningless if URLs of legitimate sites are too complex to be known or recognized by users. In response to this challenge, (Ludl et al. 2007) analyze a large number of phishing pages and explore the page properties that can be used to identify phishing pages. These features from the HTML source of a page include the following: the number of forms, input fields (e.g., the number of input fields, text fields, password fields, and hidden fields), links (e.g., the number of internals links to internal links to resources located in the page's domain

as well as external links to resources stored on other sites), white-listed references, and script tags. Zhang et al. (2011) introduce a content analysis-based large-scale anti-phishing gateway. When the http(s) traffic is intercepted by the gateway, the system fetches and filters the target URL. If the URL is not prefiltered by the black and white hash repository, the system fetches the web page content and extracts features. They build a phishing page template database as a repository. After feature extraction, the system calculates the similarity scores between the evaluated web page and each template in the database. They evaluate the performance of the detection system based on 118,165 positive URLs and 92,970 negative URLs. The maximum false-positive rate is below 0.1 %, and the average false-positive rate and false-negative rate are 0.05 and 1.78 %, respectively. The system demonstrates better performance than several other approaches. Whittaker et al. (2010) present a logistic regression classifier based on features that describe the composition of the web page's URL, the hosting of the page, and the page's HMTL content as collected by a crawler. The evaluation of the classifier is based on two datasets. The first one contains 446,152,060 URLs and the second contains 74,816,740 URLs. The phishing pages make up 1.1 % of each dataset. The study shows that the classifier can maintain a false-positive rate well below 0.1 %.

Due to inevitable false positives, directly blocking users' connections to suspected phishing sites is unacceptable. Therefore, phishing site warning mechanisms become mandatory in popular browsers including Firefox and IE. If a web page is correctly identified as a phishing site, a user is directed to a warning page and not allowed to proceed without interacting with the warning page. If the user chooses to ignore the link, the warning page disappears and the user is exposed to the risk of phishing. Otherwise, the user is directed to a default page. A hybrid solution, AntiPhish (Kirda and Kruegel 2005), integrates phishing warning and page analysis for phishing identification. It keeps track of where sensitive information is being submitted. If it detects that confidential information such as a password is being entered into a form on a suspicious website, a warning is generated and the pending operation is cancelled.

However, users tend to ignore the warnings or have learned to bypass the warnings. Wu et al. (2006) conduct a study of three simulation anti-phishing toolbars to determine how effective they are at preventing users from visiting websites that the tools have determined to be fraudulent. They find that many participants do not notice warning signals or assume the warnings are invalid. In a follow-up study the authors test anti-phishing toolbars that produce pop-up warnings and block access to fraudulent websites until overridden by the user. These pop-up warnings reduce the rate at which users fall for fraudulent sites, but do not completely prevent all users from failing for these sites. Egelman et al. (2008) compare the effectiveness of active and passive phishing warning. They designed two phishing websites to mimic the login pages of Amazon and eBay, the most commonly phished nonbank websites. They divide the 60 participants into four groups: Firefox warning, active IE warning, passive IE warning, and no warning at all. The results show that over 45 and 90 % of participants ignore the strong warning or the passive warning, respectively. Similarly, Schneider et al. (2007) demonstrate that over 50 % participants of a warning usability test ignore the warning and enter their credentials, despite the strong wording of the warning page.

Traditional phishing begin with e-mail spam. SMTP (Simple Mail Transfer Protocol) (Jonathan 1982) is the protocol to deliver e-mails in the Internet. It is a simple protocol which lacks necessary authentication mechanisms. Information related to sender, such as the name and e-mail address of the sender, can be counterfeited in SMTP. Therefore, attackers can send out spoofed e-mails that are seems from a friend, relative, or a reputable business where victims might have an account. A number of solutions have been proposed to solve the anti-phishing problem at the e-mail level. Since the phishing e-mail usually contains some socially engineered message

asking users to submit information or to visit the phishing website, filters and content analysis are used to prevent phishing e-mails from reaching their addresses' inbox. For example, MailScanner (Julian Field 2007) is an anti-spam package for e-mail gateway systems in attempts to combat e-mail fraud by examining e-mail contents. ClamAV (2012) is another toolkit for e-mail scanning making use of blacklisting and phishing signature, such as the use of a specific phrase or looking for the PayPal (https://www.paypal.com/home) logo that many phishing e-mails contain.

The effectiveness of such techniques relies on critical factors, such as natural language processing, filter training using machine learning approaches, and the availability of anti-phishing tools in the e-mail system. Chandrasekaran et al. (2006) use the distinct structural features present in e-mail to classify phishing e-mails. A total of 25 features consisting of a mixture of style marker (e.g., account, risk, bank, risk, and vocabulary richness) and structural attributes (e.g., the structure of the greeting in the body and the structure of the subject line of the e-mail) are considered. Features are ranked based on their relevance to e-mail classification. A total of 400 e-mails, out of which 200 are phishing e-mails, are used in training and evaluating the model. The results demonstrated a detection rate of 95 %. Similarly, based on structural features of the phishing e-mails, Abu-nimeh et al. (2008) investigate phishing detection in a mobile environment utilizing modified Bayesian Additive Regression Trees (BART). The algorithm modification intends to reduce the computation time and memory overhead of MCMC simulations. 6,561 raw e-mails are used in building the dataset, from which 1,409 e-mails are phishing. The legitimate e-mails are collected from financial institutions such as Bank of America, eBay, and Chase and regular communication e-mails. The dataset constitutes of 60 style marker features and 10 structural attribute features, respectively. The results show a detection rate of 97 % and a false-positive rate of 3 %. However, no matter how effective, some phishing e-mails can still successfully get through the filters and reach potential victims.

Future Directions

The battle between phishing and anti-phishing is far from over. With the advancement of anti-phishing techniques, phishers constantly come up with new ways of stealing sensitive information from users, by pretending to come from their close friends, by including fake US Airways itineraries, by quietly changing the content in one of the browser tabs, etc. Therefore, it is necessary to raise the awareness of phishing crimes among the general public, to keep the anti-phishing tools up-to-date regarding the newly developed crime patterns, and to even predict the emergence of novel phishing patterns.

Conclusion

In this chapter, we focus on social phishing, which is a common social engineering technique for conducting fraud. Ever since its first appearance in the mid-1990s, it has evolved into a variety of sophisticated forms. In the future, to effectively combat phishing, coordinated efforts have to be made from multiple aspects, e.g., education, legislation, and improved anti-phishing techniques.

Acknowledgments

Research was sponsored by the US Defense Advanced Research Projects Agency (DARPA) under the Social Media in Strategic Communication (SMISC) program, Agreement Number W911NF-12-C0028. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies, either expressed or implied, of the US Defense Advanced Research Projects Agency or the US Government. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

Cross-References

- ▸ Counter-Terrorism, Social Network Analysis in
- ▸ Fraud Detection Using Social Network Analysis, a Case Study
- ▸ Spam Detection, E-mail/Social Network

References

Abu-nimeh S, Nappa D, Wang X, Nair S (2008) A distributed architecture for phishing detection using Bayesian Additive Regression Trees. eCrime Researchers Summit, Atlanta, GA

Chandrasekaran M, Narayanan K, Upadhyaya S (2006) Phishing email detection based on structural properties. In: Proceedings of the NYS cyber security conference. Albany, NY

ClamAV. ClamAV. http://www.clamav.net. Accessed 23 July 2012

eBay (2007) eBay tool bar. http://anywhere.ebay.com/browser/firefox/. Accessed 11 Feb 2013

Egelman S, Cranor LF, Hong J (2008) You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In: CHI, Florence

Herley C, Florêncio D (2008) A profitless Endeavor: phishing as Tragedy of the commons. In: NSPW. Victoria, BC

http://cups.cs.cmu.edu. Accessed 18 July 2012

http://en.wikipedia.org/wiki/Phishing#Early_phishing_on_AOL. Accessed 18 July 2012

http://ists.dartmouth.edu/. Accessed 11 Feb 2013

http://theory.stanford.edu/seclab/. Accessed 18 July 2012

http://www.cerias.purdue.edu. Accessed 18 July 2012

http://www.csionsite.com/2012/phishing/. Accessed 23 July 2012

http://www.indiana.edu/~phishing/. Accessed 18 July 2012

http://www.indiana.edu/~phishing/?about. Accessed 18 July 2012

http://www.phishing.org. Accessed 23 July 2012

https://www.paypal.com/home. Accessed 23 July 2012

Jagatic T, Johnson N, Jakobsson M, Menczer F (2007) Social phishing. Commun ACM 50(10):94–97

Jonathan BP (1982) Simple Mail Transfer Protocol. RFC821: http://freesoft.org/CIE/RFC/821/index.htm. Accessed 11 Feb 2013

Julian Field (2007) MailScanner. http://www.mailscanner.info. Accessed 23 July 2012

Kirda E, Kruegel C (2005) Protecting users against phishing attacks. Comput J, 49:2006

Ludl C, McAllister S, Kirda E, Kruegel C (2007) On the effectiveness of techniques to detect phishing sites. In: DIMVA '07: proceedings of the 4th international conference on detection of intrusions and malware, and vulnerability assessment. Springer, Berlin/Heidelberg, Lucerne, Switzerland, p 2039

NetCraft (2007) Netcraft anti-phishing tool bar. http://toolbar.netcraft.com/. Accessed 23 July 2012

Ramachandran A, Feamster N (2006) Understanding the network-level behavior of spammers. In: SIGCOMM '06: proceedings of the 2006 conference on applications, technologies, architectures, and protocols for computer communications. Pisa, Italy, pp 291–302

Schneider F, Provos N, Moll R, Chew M, Rakowski B (2007) Phishing protection design documentation. https://wiki.mozilla.org/Phishing_Protection:_Design_Documentation. Accessed 23 July 2012

Sheng S, Wardman B, Warner G, Cranor LF, Hong J, Zhang C (2009) An empirical analysis of phishing blacklists. In: CEAS 2009: sixth conference on email and anti-spam. Como, Italy

Sun B, Wen Q, Liang X (2010) A DNS based anti-phishing approach. In: In second international conference on networks security, wireless communications and trusted computing, Beijing

Whittaker C, Ryner B, Nazif M (2010) Large-scale automatic classification of phishing pages. In: NDSS'10. San Diego, California

Wilson C, Argles D (2011) The Fight against phishing: technology, the end user and legislation. In: The international conference on information society (i-Society), London

Wu M, Miller RC, Garfinkel SL (2006) Do security toolbars actually prevent phishing attacks? In: Proceedings of the SIGCHI conference on human factors in computing systems, Montreal

Zhang J, Wu C, Guan H, Wang Q, Zhang L, Ou Y, Xin Y, Chen L (2011) An content-analysis based large scale anti-phishing gateway. In: 12th IEEE international conference on communication technology, Nanjing

Social Factor

- ▸ User Behavior in Online Social Networks, Influencing Factors

Social Graph Dataset

- ▸ Social Network Datasets

Social Grids

- ▸ Social Network Analysis and Company Linguistic Identity

Social Group Evolution

▸ Community Evolution

Social Groups in Crowd

Jarosław Wąs[1] and Krzysztof Kułakowski[2]
[1]Faculty of Electrical Engineering, Automatics, Computer Science and Biomedical Engineering, Department of Applied Computer Science, AGH University of Science and Technology, Krakow, Poland
[2]Faculty of Physics and Applied Computer Science, Department of Applied Informatics and Computational Physics, AGH University of Science and Technology, Krakow, Poland

Synonyms

Familiar groups in crowd; Mesoscale structures in crowd

Glossary

Crowd A temporary gathering of persons
Dyad A group consisting of two persons
Triad A group consisting of three persons
Small Group A group enough for all members to interact simultaneously. It is possible for all members to communicate or be acquainted with each other

Definition

According to different authors, a social group is a set of people with a common fate, with a direct interaction between them, with a social relationship between them, or who consider themselves as members of the same social category.

A crowd is a large group of people, gathered at one time and place, connected by a common aim.

A social group in crowd is defined as two or more human beings, who are allocated in the crowd and who are connected by and within social relationships.

Most frequently crowd consists of a set of social groups like couples, groups of friends, or families.

Introduction

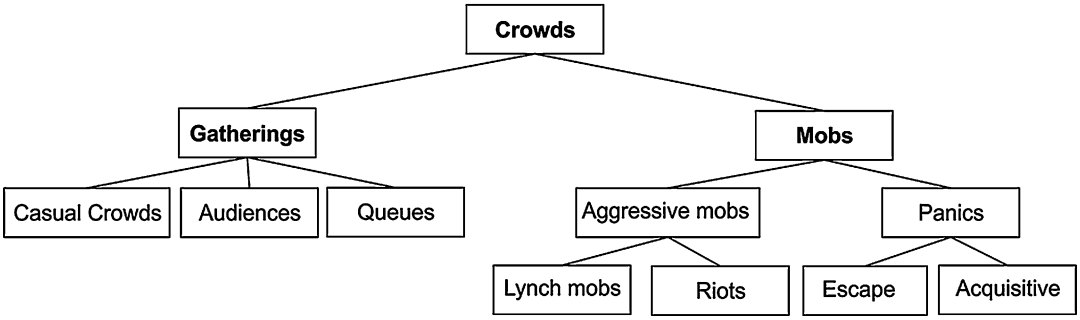
The occurrence of social groups in human crowds is a very common phenomenon. It is estimated that, depending on the situation, about 50–75 % of people walk in groups and hold together in a crowd (Aveni 1977; Moussaïd et al. 2010).

A group in crowd is interpreted as two or more persons who are connected by interpersonal relationships. We can distinguish several methods of analyzing crowd dynamics and crowd behavior: from the macroscopic level when the crowd is treated as a whole, through the microscopic level when we consider the behavior and dynamics of individuals, and finally the analysis of the behaviors of particular groups of people in the crowd - the mesolevel.

It seems that the mesolevel analysis of crowd is crucial in terms of crowd behavior seen as a whole Moussaïd et al. (2010).

Crowd Classification

Social groups are part of vast majority of crowds. What is a crowd? Forsyth (2005) defines crowd as "a temporary gathering of individuals, who share a common focus on interest." The occurrence and character of these groups depends on the type of crowd. According to Forsyth (2005) one can distinguish two different types of crowd: *gatherings* and *mobs*. Both of these types of crowds are different from the perspective of a situational context: gatherings mean more ordered aggregation of persons like audiences, queues, or street crowds, while a mob is described as an acting, disordered crowd, often aggressive in character. In some cases, in a social group or crowd, the deindividuation phenomenon may occur, as described by



Social Groups in Crowd, Fig. 1 Crowd classification according to Forsyth (2005)

Zimbardo (1969). In this situation one can observe loss of self-awareness, and reduced responsibility, loss of self-regulation, emotional and impulsive behavior of individuals (Fig. 1).

Modeling and Simulation of Crowd Dynamics

Models of crowd dynamics and crowd behavior are used for simulations of evacuation, simulations of mass events, design of pedestrian traffic in public utility facilities, and, finally, in the entertainment industry (in the creation of movies, games, and special effects).

One can distinguish two main kinds of crowd dynamics models: macroscopic, where pedestrians are considered as fluid particles in hydrodynamics equations (Henderson 1974), and microscopic approach, where pedestrians are considered as individuals or groups (Köster et al. 2011). Actually, most of the crowd dynamics models are based on the microscopic approach, as it entails the mapping of behavior of particular individuals or groups.

The most common method of microscopic modeling of crowd dynamics is Social Force Model (Helbing and Molnar 1995). In this model, time and space are continuous. The model is based on differential equations equivalent to the Newton's second law of dynamics. There, each pedestrian i with mass of m_i moves according to the following equation:

$$m_i \frac{dv_i}{dt} = F_a$$

where

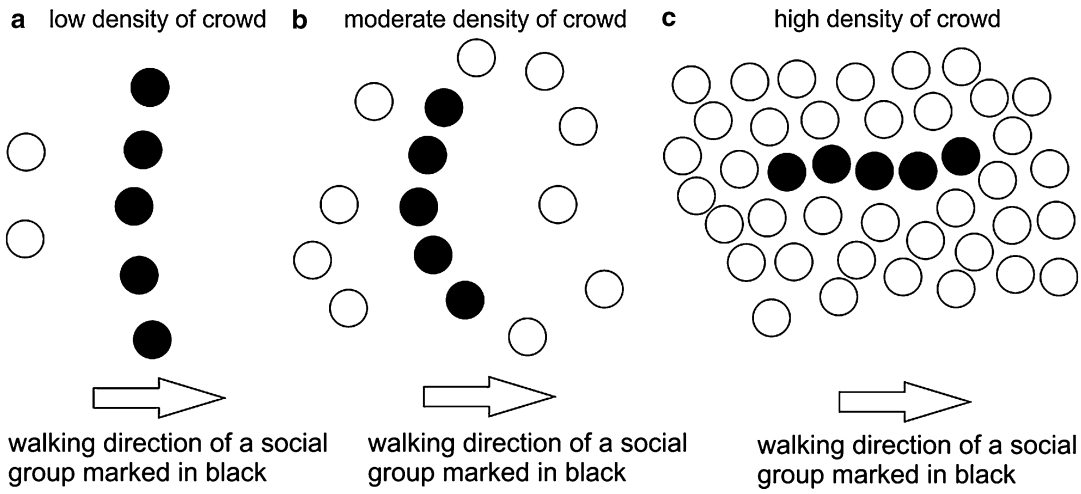
v_i – actual velocity of pedestrian i
m_i – mass of pedestrian i
F_a – the vector of forces, which takes into account personal desire force and interaction forces

The method has a lot of variants and extensions. The most important fact is that using some variants of the method, we can take into account group attraction forces and we can map group dynamics and group behavior (Moussaïd et al. 2010).

Another popular method of crowd dynamics modeling is Cellular Automaton (CA). It is a rule-based dynamical model, where time and space are discrete. Majority of implementations of CA models are interpreted as agent-based models (Burstedde et al. 2001; Wąs et al. 2012). Pedestrians represented as autonomous agents are allocated in a lattice. Agents move on the lattice according to a transition rule f that modifies the configuration C_t of the agents allocated in the lattice (their environment) at certain interval Δt:

$$C_{t+1} = f(C_t)$$

The transition function f can be implemented using floor field (FF), which is defined on a supplementary lattice. Floor field is a set of rules assigned to the lattice, which take into account different parameters, determinating a type of the floor field: distance from a pedestrian to an aim (static floor field) (Burstedde et al. 2001), following predecessors of a pedestrian (dynamic floor field) (Burstedde et al. 2001), omitting obstacles



Social Groups in Crowd, Fig. 2 Exemplary patterns of walking small, social groups for low, moderate, and high density of flow. Members of walking social group are marked in *black*

(Georgoudas et al. 2010), or anticipating potential collisions (anticipation floor field) (Suma et al. 2012).

Most of the models based on cellular automata assume that the crowd is made up of individuals (Burstedde et al. 2001). It was demonstrated recently that it is possible to define rules of behavior for groups as well (Köster et al. 2011; Bandini et al. 2011).

Social Groups Behavior

In a crowd we can often observe a situation, when a group of people intentionally walk together (for instance, family members, couples, or friends). It is in opposition to a different situation, when several proximate pedestrians fortuitously walk close to each other (Moussaïd et al. 2010). How to recognize these cases?

Social groups in crowd behave differently depending on crowd density, size, and purposes of the particular group or type of crowd.

At low densities of crowd, group members usually tend to walk side by side creating a line perpendicular to the walking direction (Fig. 2a). When the density increases, the linear walking formation is bent forward, turning it into a V-like pattern (Fig. 2b). These spatial patterns can be well described by a model based on *social communication* among group members (Moussaïd et al. 2010). In very high densities, V-like patterns are transformed into a lane aimed towards the direction of motion (Fig. 2c).

The speed of the group is related to the density of the crowd and the size of the group. Speed of movement in high densities can be estimated based on the fundamental diagram (the relation between density and flow) (Seyfried et al. 2005; Chattaraj et al. 2009). It should be stressed that in smaller densities, there is a rule that the larger the group, the smaller the velocity of motion, and this relationship between size of the group and its velocity is linear (Klüpfel 2007; Moussaïd et al. 2010).

If we consider crowd as a social network with individuals represented as nodes of a graph (Fig. 2), then the social groups are interpreted as subgraphs, namely, network motifs (Milo et al. 2002; Juszczyszyn et al. 2009). Network motifs in this case may include two nodes (a dyad), three nodes (a triads), or up to a dozen nodes.

Sociological Aspects

Specific properties and identity of a group as opposed to an individual has always been a matter

of interest. *Senatores boni viri, senatus autem bestia* (Senators are good men, but the Senate is an evil beast) was a common opinion in ancient Rome; this sentence indicates a contemporary knowledge belief held at the time that the Roman senate was different (worse) than a set of its members. Ability to control crowds is a necessary skill for politicians, artists, military, and religious leaders. Besides their personal charisma, their methods include rhetorics and dedicated institutions, like army, church, and theatre. In more modern times, this role is played also by media. In recent 10 years, interpersonal communication became possible by means of Internet and smartphones. The latter technology allows members of a large crowd interact in real time; this kind of communication transforms a crowd to an autonomous system. Its power has been demonstrated during a series of events known as the Arab Spring.

The idea of collective thinking has been of interest for centuries. The very concept of democracy relies on the belief that decisions taken by many people can be ahead of those by a single ruler. It makes sense to ask what is the quality of decisions made by a crowd? According to a traditional notion derived from Gustave Le Bon, crowd is irrational. Being part of the crowd, an individual gains feeling of power and loses responsibility. On the contrary, individualistic theories deny the existence of anything like collective thinking. Often, a famous statement by Floyd H. Allport is quoted: "The individual in the crowd behaves just as he would behave alone only more so." Today, both these positions are rejected as unsupported by experience.

If crowd is different from a sum of individuals, interaction between them is an issue of primary importance. During this interaction, some individuals appear to be more influential than others. Once such a leader is able to create an impression of unanimity within some group, the group accepts his leadership. Simultaneously, the group itself is established, with identity defined by the content of accepted messages. It is clear that this acceptance depends, among others, on relations between personal features of individuals and leaders. During the process of group formation, the group identity continues to evolve. According to John C. Turner, the direction of this evolution is such as to minimize intragroup differences with respect to intergroup differences. Also, group decisions are more extreme and more risky than initial attitudes of group members; the effect is known as group polarization (Turner 1975; Cooper et al. 2001).

In a longer time scale, the very existence of groups has far-reaching consequences. However, immediate group formation in crowd can also be observed in situations of emergency, when communication is limited to a given area, as within the hearing range. According to the emergent norm theory by Ralph Turner and Lewis Killian, group members perceive their group as unanimous (Turner and Killian 1987). As they follow the group action, the illusion of unanimity can become a self-fulfilling prophecy. If this is the case, we expect homogeneity of group behavior. On the contrary, differences between groups tend to grow and can lead to an intergroup hostility, even if the group formation is a purely random process.

Social Groups and Evacuation

Individual persons, as well as social groups, are the constitutive units of emergency evacuations.

When considered from the perspective of an evacuation study, a social group is characterized by several types of characteristics (Santos and Aguirre 2005):

- Operational context of a group (characteristic of their environment)
- Individual characteristics of members (age, sex, physical fitness, health, competences, etc.)
- Density understood as a function of physical space occupied by the group and the size of the group
- Relationships among the members including leadership, communication channels, cohesiveness, etc.

When a social group is faced with an emergency that makes it necessary to evacuate, the key parameter is the decision making (Aveni 1977). One of the most important determinants of evacuation timing is the size of a group. The larger the group, the more difficult it is to take the decision to begin evacuation as a response to the emergency. It should be stressed that "in the large group there will be more variation and differences of opinion and relevant experiences about what to do that must be reconciled before the emergent norm is created" (Santos and Aguirre 2005). Response time for an emergency is significant component of evacuation time, and it is determined by occurrence and characteristics of social groups.

Social groups also strongly influence the evacuation effectiveness during movement phase of evacuation, because members of the groups often create blocks (cluster patterns). In practice, single individuals (not members of the group) who want to overtake the group have great difficulties to do this especially in constrained spaces like narrow corridors and stairways. because they are exposed to "the set of norms and new statuses guiding the behavior of these collectivities which they cannot evade" (Santos and Aguirre 2005).

In extreme cases (caused by real or imagined reasons) evacuation situation may cause sudden, overpowering terror called panic, when an individual or the whole group is affected at once. In this case, relationships within the specified group have a large impact on behavior. During panic it is often possible to observe anti-social behaviors, but strong relationship within a group often leads to altruism and strong cooperation in the group according to the sentence "families survive together or die together" cited by Köster et al. (2011).

Group Structure of Crowd

Usually a mere observation of a crowd does not allow inferring about existence and content of groups there, and dedicated tools are necessary. More than often, these tools are borrowed from statistical mechanics. In particular, the concept of modularity was proposed by Mark Newman. This quantity allows to evaluate if a given group structure is statistically meaningful.

Suppose we have a weighted network. Its nodes are pedestrians, and the links describe the similarities between the nodes. In particular, we can measure the trajectories of pedestrians; the value $w(l, j)$ assigned to the link between nodes l, j is an absolute value of the Pearson's correlation coefficient between pedestrians l and j (Rodgers and Nicewander 1988). The correlation can be calculated for positions or velocities or both. Suppose that we have a proposition of the network structure. This means that all nodes are divided into groups. The modularity Q is defined (Blondel et al. 2008) as

$$Q = \frac{1}{m} \sum_{lj} \left\{ w(l, j) - \frac{k(l)k(j)}{m} \right\} \delta(l, j)$$

where

$k(j) = \sum_{l} w(l, j)$, $m = \sum_{lj} w(l, j)$ and $\delta(l, j)$ is equal to one if nodes l, j belong to the same group according to the proposed division; otherwise it is zero.

The challenge is to find the proposed division which gives the maximal value of Q. For large networks this task is NP-complete, then it cannot be treated with exhaustive methods. Instead, approximate algorithms have been proposed. One of them – the so-called agglomerative method – is to connect two nodes which give the largest Q; subsequent nodes are added according to the same rule. Starting from N separated nodes, we end up with a single connected cluster. Somewhere at this path, Q has a maximum; this is the approximated partition. This, however, does not prove that it is statistically meaningful. If the maximal value of Q is at least 0.3, our confidence increases. We recommend (Fortunato 2010) for a review.

The method of detecting correlations of trajectories and velocities can be used to monitor crowd dynamics in real time (Helbing and Mukerji 2012).

Key Applications

Real-time monitoring of large gatherings supported by a software able to identify collective motion and interpersonal correlations should be helpful for predictions and prevention of stampede disasters, like the one in Duisburg, Germany, in 2010 (Helbing and Mukerji 2012).

Future Directions

An interdisciplinary research conducted by sociologists, psychologists, physicists, computer scientists, and fire and transportation engineers can advance our understanding of mutual influence of majority and minority in crowd. In particular, the social mechanisms which rule this influence are not known yet (Brown 2000). Analysis of data on crowd dynamics collected during large gatherings is an example of a research strategy which can build bridges between social theory, field experiments, and applications.

Acknowledgments

The authors acknowledge financial support of FP7 project SOCIONICAL, No 231288.

Cross-References

- ▸ Collective Intelligence, Overview
- ▸ Community Evolution
- ▸ Extracting Individual and Group Behavior from Mobility Data
- ▸ Human Behavior and Social Networks
- ▸ Motif Analysis
- ▸ Simulation
- ▸ Spatio – Temporal Proximity and Social Distance

References

Aveni AF (1977) The not-so-lonely crowd: friendship groups in collective behavior. Sociometry 40(1):96–99

Bandini S, Rubagotti F, Vizzari G, Shimura K (2011) A cellular automata based model for pedestrian and group dynamics: motivations and first experiments. Parallel Comput Technol 6873: 125–139

Blondel VD, Guillame JL, Lambiotte R, Lefebvre E (2008,10) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008: 10008

Brown R (2000) Group processes: dynamics within and between groups. Blackwell, Oxford

Burstedde C, Klauck K, Schadschneider A, Zittarz J (2001) Simulation of pedestrian dynamics using a two-dimensional cellular automaton. Phys A Stat Mech Appl 295(3–4):507–525

Chattaraj U, Seyfried A, Chakroborty P (2009) Comparison of pedestrian fundamental diagram across cultures. Adv Complex Syst 12(3):393–405

Cooper J, Kelly KA, Weaver K (2001) Attitudes, norms, and social groups. In: Hogg MA, Tindale SR (eds) Blackwell handbook of social psychology: group processes. Blackwell, Malden

Forsyth DR (2005) Group dynamics. Wadsworth, Belmont

Fortunato S (2010) Community detection in graphs. Phys Rec 486(3–5):75–174

Georgoudas IG, Koltsidas G, Sirakoulis GC, Andreadis I (2010) A cellular automaton model for crowd evacuation and its auto-defined obstacle avoidance attribute. In: International conference on cellular automata for research and industry. Springer, Ascoli-Piceno, pp 21–24

Helbing D, Molnar P (1995) Social force model of pedestrian dynamics. Phys Rev E 51(5):4282–4286

Helbing D, Mukerji P (2012) Crowd disasters as systemic failures: analysis of the love parade disaster. EPJ Data Sci 1:7

Henderson LF (1974) On the fluid mechanics of human crowd motion. Transp Res 8(6):509–515

Juszczyszyn K, Musial K, Kazienko P, Gabrys B (2009) Temporal changes in local topology of an e-mail based social network. Comput Inform 28(6):763–779

Klüpfel H (2007) The simulation of crowd dynamics at very large events. Calibration, empirical data, and validation. In: Pedestrian and evacuation dynamics. Springer, New York, pp 285–296

Köster G, Seitz M, Treml F, Hartmann D, Klein W (2011) On modelling the influence of group formations in a crowd. Contemp Soc Sci J Acad Sci 6(3):397–414

Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. Science 298(5594):824–827

Moussaïd M, Perozo NG, Helbing D, Theraulaz G (2010) The walking behaviour of pedestrian social groups and its impact on crowd dynamics. PLoS One 5(4):e10047

Rodgers JL, Nicewander WA (1988) Thirteen ways to look at the correlation coefficient. Am Stat 42:59–66

Santos G, Aguirre B (2005) A critical review of emergency evacuation simulation. In: Workshop on building occupant movement during fire emergencies, Gaithersburg, pp 27–52

Seyfried A, Steffen B, Klingsch W, Boltes M (2005) The fundamental diagram of pedestrian movement revisited. J Stat Mech Theory Exp 2005(10):10002. doi:10.1088/1742-5468/2005/10/P10002

Suma Y, Yanagisawa D, Nishinari K (2012) Anticipation effect in pedestrian dynamics: modeling and experiments. Phys A Stat Mech Appl 391(1–2):248–263

Turner JC (1975) Social comparison and social identity: some prospects for intergroup behaviour. Eur J Soc Psychol 5:5–34

Turner R, Killian L (1987) Collective behavior. Prentice Hall, Englewood Cliffs

Wąs J, Lubaś R, Myśliwiec W (2012) Proxemics in discrete simulation of evacuation. In: 10th international conference on cellular automata for research and industry, Santorini. Springer, pp 768–775

Zimbardo P (1969) The human choice: individuation, reason, and order versus deindividuation, impulse, and chaos. In: Nebraska symposium on motivation. Nebraska, pp 237–307

Recommended Reading

Helbing D, Johansson A (2010) Pedestrian, crowd and evacuation dynamics. Encyclopedia of complexity and systems science, vol 16. Springer, New York, pp 6476–6495

Social History of Computing and Online Social Communities

Joseph M. Kizza and Li Yang
Department of Computer Science and Engineering, The University of Tennessee-Chattanooga, Chattanooga, TN, USA

Synonyms

Crime in online communities; Ethics; Privacy; Social media; Social networks

Glossary

OSNs Online social networks (OSNs) are social networks with underlining electronic communication infrastructure links enabling the connection of the interdependencies between the network nodes

mOSNs Mobile OSNs (mOSNs) are newer OSNs that can be accessed via mobile devices and can deal with the new mobile context

IMN Instant Messaging Network (IMN) supports real-time communication between two or more individuals

SNS Social networking services (SNS)

Definition

It is almost unimaginable that a modern person can live a meaningful life today without a mobile device as a conduit to an online social mesh of friends. These online social "gatherings" have slowly replaced the traditional face-to-face social gatherings that make us humans. While these online ecosystems are now packed with all sorts of interesting items that keep members coming back and new ones enrolling, the basic element of "presence" which transforms into "telepresence" in the virtual gatherings of any social gathering remains the same. The history of this amazing transformation of social gatherings mimics the history of social computing, the focus of this entry. The development of the different media of social gatherings and communication is linked with computer technology developed. In fact the nature of these social media developed in line with the computing technology. The history of social computing cannot be discussed comprehensively without talking about these online media. And these online social media cannot be justifiably discussed without investigating individual rights and how these media affect participants' individual attributes. Therefore, ethical, privacy, and security issues in these ecosystems are all involved in protecting personal privacy. On the central point of ethical implications of life in the social network, unlike in the traditional network, governance is not centralized, but community based with equally shared authority and responsibility by all users. But the mechanisms are not yet defined, and where they are being defined, it is still too early to say whether they are

effective. The complexity, unpredictability, and lack of central authority are further enhanced by a virtual personality, anonymity, and multiple personality. These three characteristics are at the core of the social and ethical problems in online social networks in particular cyberspace in general; the larger and more numerous these communities become, the more urgent the ethical concerns become.

Introduction

Social networks are at the core of social computing! In this discussion, therefore, the history of social computing is going to be discussed through the prism of social networks and their evolution into online social ecosystems, as we have them today. So a social network is a theoretical network where each node is an individual, a group, or an organization that independently generates, captures, and disseminates information and also serves as a relay for other members of the network. This means that individual nodes must collaborate to propagate the information in the network. The links between nodes represent relationships and social interactions between individuals, groups, organizations, or even entire society.

The concept of social networking is not new. Sociologists and psychologists have been dealing with and analyzing social networks for generations. In fact social networks have been in existence since the beginning of human. Prehistoric man formed social networks for different reasons including security, access to food, and the social well-being.

As Joseph Kizza (2013) observes, social networks begin with an individual reaching out to another individual or group for a social relationship of sorts, and it snowballs into a mesh of social relationships connecting many individuals and/or groups. In general, social networks come in all sizes and are self-organizing, complex, and agile depending on the nature of relationships in its links. As they grow in size, social networks tend to acquire specific elements and traits that make them different. These traits become more apparent as the network size increases. The type of social interactions, beliefs, and other traits usually limit the size of the social network. It is important to note that as the social network grows big, it tends to lose the nuances of a local system; hence if certain qualities of the network properties are needed, it is better to keep the size under control.

Online Social Networks (OSNs)

As computing technology developed, social networks started evolving into online social networks. **Online social networks** (OSNs) are social networks with underlining electronic communication infrastructure links enabling the connection of the interdependencies between the network nodes. The discussion in this entry will focus on these OSNs. In particular we will focus on two types of online social networks (Kizza 2013):

- The traditional OSNs such as Facebook and Myspace. Many of these can be accessed via mobile devices without the capability of dealing with mobile content.
- The mobile OSNs (mOSNs) which are newer OSNs that can be accessed via mobile devices and can deal with the new mobile context.

The interdependency between nodes in the OSNs supports social network services among people as nodes. These interdependencies as relations among people participating in the network services define the type of OSNs.

Types of Online Social Networks

The growth of the OSNs over the years since the beginning of digital communication saw them evolving through several types. Let us look at the most popular types using a historical chronology (Kizza 2013):

Chat Network The chat network was born out of the digital chatting anchored on a *chat room*. The chat room was and still is a virtual room online where people "gather" just to chat. Most chat rooms have open access policies meaning that anyone interested in chatting or just reading

others' chats may enter the chat room. People can "enter" and "exit" any time during the chats. At any one time, several threads of the public chats may be going on. Each individual in the chat room is given a small window on his or her communication device to enter a few lines of chat contributing to one or more of the discussion threads. This communication occurs in real time and whatever one submits to the chat room can be seen by anyone in the chat room. Chat rooms also have a feature where a participating individual can invite another individual currently in the public chat room into a private chat room where the two can continue with limited "privacy." To be a member of the chat room, you must create a user name and members of the chat room will know you by that. Frequent chatters will normally become acquaintances based on user names. Some chat room software allows users to create and upload their profiles so that users can know you more via your profile.

Although chat rooms by their own nature are public and free for all, some are monitored for specific compliance based usually on attributes like topics under discussion.

With the coming of more graphical-based online services, the use of chat room is becoming less popular especially to youth.

Blog Network Another online social network is the bloggers network. "Blogs" are nothing more than people's online journals. Avid bloggers keep diaries of daily activities. These diaries sometimes are specific on one thread of interest to the blogger or a series of random logs of events during a specific activity. Some blogs are comment on specific topics. Some bloggers have a devoted following depending on the issues.

Instant Messaging Network (IMN) The IMN supports real-time communication between two or more individuals. Like chat rooms, each participant in the IMN must have a user name. To IM an individual, one must know that individual's user name or screen name. The initiator of the IM is provided with a small window to type the message, and the recipient is also provided with a similar window to reply to the message. The transcript of the interchange is kept scrolling up both users' screens. Unlike the chat room however, these exchanges of short messages are private. Like in chat networks, some IMN allows users to keep profiles of themselves.

Online Social Networks (OSNs) These are a combination of all the network types we have discussed above and other highly advanced online features with advanced graphics. There are several of these social networks including Facebook, Twitter, Myspace, Friendster, YouTube, Flickr, and LinkedIn. Since these networks grew out of those we have seen before, many of the features of these networks are those we have discussed in the above networks. For example, users in these networks can create profiles that include their graphics and other enclosures and upload them to their network accounts. They must have a user name or screen name. Also communication, if desired, can occur in real time as if one is using chat or IM capabilities. In additional to real time, these networks also give the user the delayed and archiving features so that the users can store and search for information. Because of these additional archival and search capabilities, network administrators have fought with the issues of privacy and security of users as we will see later in this entry. As a way to keep users' data safe, profiles can be set to a private setting, thus limiting access to private information by authorized users.

Online Social Networking Services

An online social networking service is an online service accessible via any internet-enabled device with the goal of facilitating computer-mediated interaction among people who share interests, activities, backgrounds, or real-life connections. Social networking services (SNS) offer users functionalities for identity management (i.e., the representation of the owner, e.g., in form of a profile) and enable furthermore to keep in touch with other users (and thus the administration of own contacts) (Koch et al. 2007).

Most online social network services consist of:

- User profile management: People construct user profile in social networks for a particular group of audience or a particular task. The profile is used and managed as a social identity that they used to present to each other and analyze each other.
- Social or business links of interests: Users of social networks can search experts or peers based on different criteria such as interest, company, or name. They can also proactively receive recommendations for contacts of interests from social networks.
- Context awareness: This helps to identify common backgrounds of users in social networks. For example, users could have common contacts, common interests, the same university, or the same company. Context awareness helps to build trust among users, which are essential for a successfully collaboration (Kramer 1999).
- Contact management: This combines all functionalities that manage and maintain users' personal network. Examples include tagging people and access restrictions to profile in social networks.
- Network awareness: This includes any change or update of users in one's personal network. This includes awareness of indirect communication, News Feeds, and user notification.
- Exchange: This enables information sharing directly (e.g., messages) or indirectly (e.g., photos or messages via bulletin boards). Examples of exchange in social networks include messages and photo albums.

Currently, the most popular online social network services fall in categories that range from friends based, music and movie, religion, business, and many other interests. In each of these categories, let us give a sample of the current services:

- General and friends-based social networks
  - Facebook
  - Myspace
  - Hi4
- Movie and music social networks
  - Last.fm
  - Flixster
  - iLike
- Mobile social networks
  - Dodgeball
  - Loopt
  - Mozes
- Hobby and special interest social networks
  - ActionProfiles
  - FanIQ
- Business social networks
  - LinkedIn
  - XING
  - Konnects
- Reading and books social networks
  - Goodreads
  - Shelfari
  - LibraryThing

The Growth of Online Social Networks

OSNs have blossomed as the Internet exploded. The history and the growth of OSNs have mirrored and kept in tandem with the growth of the Internet. At the infant age of the Internet, computer-mediated communication services like Usenet, ARPANET, LISTSERV, and bulletin board services (BBS) helped to start the growth of the current OSNs as we know them today. Let us now see how these contributed to the growth of OSNs.

BITNET was an early world leader in network communications for the research and education communities and helped lay the groundwork for the subsequent introduction of the Internet, especially outside the US (Fox 2000). Both BITNET and Usenet, which were invented around the same time in 1981 by Ira Fuchs and Greydon Freeman at the City University of New York (CUNY), were both "store-and-forward" networks. BITNET was originally named for the phrase "Because It's There Net," later updated to "Because It's Time Net" (Fox 2000). It was originally based on IBM's VNET e-mail system on the IBM Virtual Machine (VM) mainframe operating system. But it was later emulated on other popular operating systems like DEC, VMS, and Unix. What made BITNET so popular was its support of a variety of mailing lists supported by the LISTSERV software (ICANN 2005).

BITNET was updated in 1987 to BITNET II to provide a higher bandwidth network similar to the NSFNET. However, by 1996, it was clear that the Internet was providing a range of communication capabilities that fulfilled BITNET's roles, so CREN ended their support and the network slowly faded away (ICANN 2005).

Bulletin Board System (BBS) A Bulletin Board System (BBS) is software running on a computer allowing users on computer terminals far away to login and access the system services like uploading and downloading files and reading news and contribution of other members through emails or public bulletin boards. In "Electronic Bulletin Boards, A Case Study: The Columbia University Center for Computing Activities," Janet F. Asteroff (Evolving the High Performance Computing and Communications Initiative to Support the Nation's Information Infrastructure—Executive Summary 2013) reports that the components of computer conferencing that include private conferencing facilities, electronic mail, and electronic bulletin boards started earlier than the electronic bulletin board (BBS). Asteroff writes that the concept of an electronic bulletin board began from 1976 through ARPANET at schools such as the University of California at Berkeley, Carnegie Mellon, and Stanford University. These electronic bulletin boards were first used in the same manner as physical bulletin boards, i.e., help wanted, items for sale, public announcements, and more. But electronic bulletin boards soon became, because of the ability of the computer to store and disseminate information to many people in text form, a forum for user to debate on many subjects. In its early years, BBS connections were via telephone lines and modems. The cost of using them was high; hence, they tended to be local. As the earlier form of the World Wide Web, BBS use receded as the World Wide Web grows.

LISTSERV It started in 1986 as automatic mailing list server software which broadcast emails directed to it to all on the list. The first LISTSERV was conceived of by Ira Fuchs from *BITNET* and Dan Oberst from EDUCOM (later EDUCAUSE) and implemented by Ricky Hernandez also of EDUCOM, in order to support research mailing lists on the *BITNET* academic research network (Kizza 1999).

By the year 2000, LISTSERV ran on computers around the world managing more than 50,000 lists, with more than 30 million subscribers, delivering more than 20 million messages a day over the Internet (Kizza 1999).

Other Online Services As time went on and technology improved, other online services come along to supplement and always improve on the services of whatever was in use. Most of the new services were commercially driven. Most of them were moving toward and are currently on the web. These services including news, shopping, travel reservations, and others were the beginning of the web-based services we are enjoying today. Since they were commercially driven, they were mostly offered by ISPs such as AOL, Netscape, Microsoft, and the like. As the Internet grew millions of people flocked onto it and the web and services started moving away from ISP to fully fledged online social network companies like Facebook, Flicker, Napster, LinkedIn, Twitter, and others.

Gaining Knowledge from Social Networks

When more and more people are making their opinions available in social networks, it is possible to find out about the opinions and experiences of those in the vast pool of people that are neither our personal acquaintances nor well-known professional critics. Figuring out "What other people think" has always been an important piece of information for most of us during the decision-making process. Organizations are attempting to extract insights from opinions of their consumers for revenue increase and competitiveness improvement. The Twitter as an example of a social network consists of 40 million Twitter users, including billions of tweets, more than 1 billion relationships between users, and millions of posts, hashtags, URLs, and emoticons. Through analyzing and exploiting the Twitter data, it is possible to formulate and answer a variety of interesting problems/questions, such as the trending topics, brands, and pop culture,

to assess the sentiment or popularity around any area of interest, followers count, tweet counts by catalog, and more. For instance, the problems or questions related to Twitter may be "What's the twitter traffic distribution by hours, days, weeks, months, and years?" "Sort all the URLs twitted in descend order" "What background color Twitter users like most?" "Who is the person who twitted the most in the three year period?" "Who is the Twitter user who has the most followers by month and year?" "Which geographic location has the most Twitter users?" and so forth.

Ethical and Privacy Issues in Online Social Networks

Privacy is a human value consisting of a set of rights including solitude, the right to be alone without disturbances; anonymity, the right to have no public personal identity; intimacy, the right not to be monitored; and reserve, the right to control one's personal information, including the dissemination methods of that information. As humans, we assign a lot of value to these four rights. In fact, these rights are part of our moral and ethical systems. With the advent of the Internet, privacy has gained even more value as information has gained value. The value of privacy comes from its guardianship of the individual's personal identity and autonomy.

Autonomy is important because humans need to feel that they are in control of their destiny. The less personal information people have about an individual, the more autonomous that individual can be, especially in decision making. However, other people will challenge one's autonomy depending on the quantity, quality, and value of information they have about that individual. People usually tend to establish relationships and associations with individuals and groups that will respect their personal autonomy, especially in decision making.

As information becomes more imperative and precious, it becomes more important for individuals to guard their personal identity. Personal identity is a valuable source of information. Unfortunately, with rapid advances in technology, especially computer and telecommunication technologies, it has become increasingly difficult to protect personal identity.

Privacy Issues in OSNs

Privacy can be violated, anywhere including online social network communities, through intrusion, misuse of information, interception of information, and information matching (Web Surpasses One Billion Documents 2000). In online communities, intrusion, as an invasion of privacy, is a wrongful entry, a seizing, or acquiring of information or data belonging to other members of the online social network community. Misuse of information is all too easy. While online, we inevitably give off our information to whoever asks for it in order to get services. There is nothing wrong with collecting personal information when it is authorized and is going to be used for a legitimate reason. Routinely information collected from online community members, however, is not always used as intended. It is quite often used for unauthorized purposes, hence an invasion of privacy. As commercial activities increase online, there is likely to be stiff competition for personal information collected online for commercial purposes. Companies offering services on the Internet may seek new customers by either legally buying customer information or illegally obtaining it through eavesdropping, intrusion, and surveillance. To counter this, companies running these online communities must find ways to enhance the security of personal data online.

As the number and membership in online social networks skyrocketed, the issues of privacy and security of users while online and the security of users' data while off-line have taken center stage. The problems of online social networking have been exhibited by the already high and still growing numbers especially of young people who pay little to no attention to privacy issues for themselves or others. Every passing day, there is news about and growing concerns over breaches in privacy caused by social networking services. Many users are now worry that their personal data is being misused by the online service providers.

As the growth in online social networks continues unabated, the coming in the mix of the smart mobile devices is making the already existing problems more complex. These new devices are increasing the number of accesses to OSNs and increasing the complexity of the privacy issues, including (Wresch 1996):

- The presence of a user. Unlike in the most traditional OSNs where users were not automatically made aware of the presence of their friends, most mobile OSNs (mOSNs) now allow users to indicate their presence via a "check-in" mechanism, where a user establishes their location at a particular time. According to Wresch (1996), the indication of presence allows their friends to expect quick response, and this may lead to meeting new people who are members of the same mOSN. Although the feature of automatic locate by oneself is becoming popular, it allows leakage of personal private information along two tracks: the personal information that may be sent and the destination to which it could be sent.
- Location-based tracking system (LTS) that are part of our mobile devices. This is a feature that is widespread in the mobile environment. However, users may not be aware that their location can be made known to friends, and friends of friends who are currently online on this mOSN, their friends in other mOSNs, and others who may lead to leakage of personal information to third-parties.
- Interaction potential between mOSNs and traditional OSNs. According to Wresch (1996), such connections are useful to users who, while interacting with a mOSN, can expect some of their actions to show up on traditional OSNs and be visible to their friends there. However, a lot of their personal information can leak to unintended users of both the traditional OSNs and the mOSNs.

In addition to almost free access to a turn of personal data on OSNs, there is also a growing threat to personal data ownership. For example, who owns the data that was altered or removed by the user which may in fact be retained and/or passed to third parties? Fortunately users are beginning to fight for their privacy to prevent their personal details from being circulated further than they intended it to be. For example, Facebook's 2006 News Feed and Mini Feed features are designed to change what Founder and CEO Mark Zuckerberg called Facebook's old "Encyclopedic interface," where pages mostly just list off information about people, to the current stream of fresh news and attention content about not only the user but also the user's friends and their activities (Walsh 2013). The first, News Feed, brought to the user's home page all new activities on all friends and associate links including new photos posted by friends, relationship status changes, people joining groups, and many others, thus enabling the user to get an abundance of information from every friend's site every day. Although these features adhered to Facebook's privacy settings, meaning that only people a user allowed to view the data were able to see it, it still generated a firestone from users across the world. Over 700,000 users signed an online petition demanding the company to discontinue the feature, stating that this compromised their privacy (Walsh 2013). Much of The criticism of The News Feed was that it gave out too much individual information.

Since online social networks are bringing people together with no physical presence to engage in all human acts that traditionally have taken place in a physical environment. As these cybercommunities are brought and bound together by a sense of belonging, worthiness, and the feeling that they are valued by members of the network, they create a mental family based on trust, the kind of trust you would find in a loving family. However, because these networks are borderless, international in nature, they are forming not along well-known and traditional identifiers such as nationalities, beliefs, authority, and the like, but by common purpose and need with no legal jurisdiction and no central power to enforce community standards and norms.

Strengthening Privacy in OSNs

As more and more people join OSNs and now the rapidly growing mOSNs, there is a growing need for more protection to users. Chew et al. suggest

the following steps needed to be taken (Chew et al. 2013):

- Both OSN and mOSN applications should be explicit about which user activities automatically generate events for their activity streams.
- Users should have control over which events make it into their activity streams and be able to remove events from the streams after they have been added by an application.
- Users should know who the audience of their activity streams is and should also have control over selecting the audience of their activity streams.
- Both OSN and mOSN application should create activity stream events which are in sync with user expectation.

Other suggestions that may help in this effort are:

- Use secure passwords.
- User awareness of the privacy policies and terms of use for their OSNs and mOSNs.
- Both OSNs and mOSNs providers should devise policies and enforce existing laws to allow some privacy protection for users while on their networks.

Ethical Issues in Online Social Communities

Online social communities including online social network are far from the traditional physical social communities with an epicenter of authority with every member paying allegiance to the center with a shared sense of responsibility. This type of community governance with no central command, but an equally shared authority and responsibility, is new, and a mechanism needs to be in place and must be followed to safeguard every member of the community. But these mechanisms are not yet defined, and where they are being defined, it is still too early to say whether they are effective. The complexity, unpredictability, and lack of central authority are further enhanced by Kizza (2013):

- *Virtual personality*: You know their names, their likes, and dislikes. You know them so well that you can even bet on what they are thinking, yet you do not know them at all. You cannot meet them and recognize them in a crowd.
- *Anonymity*: You work with them almost every day. They are even your friends; you are on a first-name basis, yet you will never know them. They will forever remain anonymous to you and you to them.
- *Multiple personality*: You think you know them, but you do not because they are capable of changing and mutating into other personalities. They can change into as many personalities as there are issues being discussed. You will never know which personality you are going to deal with next.

These three characteristics are at the core of the social and ethical problems in online social networks in particular and cyberspace in general; the larger and more numerous these communities become, the more urgent the ethical concerns become. With all these happening in online social networks, the crucial utilitarian question to ask is what is best way and how we can balance the potential harms and benefits that can befall members of these online social networks and how if possible to balance these possibilities. Of late, the news media has been awash with many of these online ills and abuses, and the list is growing including potential for misuse, cyberbullying, cyber-stalking and cyber-harassment, risk for child safety, psychological effects of online social networking, and free speech.

Security and Crimes in Online Social Communities

Online crimes, in tandem with the growth of computing and telecommunication technologies, are one of the fastest growing types of crimes, and they pose the greatest danger to online communities, e-commerce, and the general public in general. An *online crime* is a crime like any other crime, except that in this case, the illegal act must involve either an Internet-enabled electronic device or computing system either as an object of a crime, an instrument used to commit a crime, or a repository of evidence related to a crime. Also online crimes are acts of unauthorized intervention into the working of the telecommunication networks and/or the sanctioning of authorized access

to the resources of the computing elements in a network that lead to a threat to the system's infrastructure or cause a significant property loss. The International Convention of Cyber Crimes and the European Convention on Cyber Crimes both list the following crimes as online crime (K-izza 2005):

- Unlawful access to information
- Illegal interception of information
- Unlawful use of telecommunication equipment
- Forgery with use of computer measures
- Intrusions of the Public Switched and Packet Network
- Network integrity violations
- Privacy violations
- Industrial espionage
- Pirated computer software
- Fraud using a computing system
- Internet/e-mail abuse
- Using computers or computer technology to commit murder, terrorism, pornography, and hacking

As we discussed before, the online contents are accessible from different locations without noticeable delay. Because of the decentralized architecture of the Internet, personal publication through the web becomes more feasible and affordable, while still maintaining a high exposure to the target audience. At the same time, the lack of regulations makes the online social community a pretty free realm where the geographical border dims in the online communities. Information can be spread anonymously with little interference from governments via the online community. Costs of the community are relatively low compared with other media. Various communities benefit from the online features of the community. We will analyze a dark web as a case study here to illustrate how terrorist/extremist organizations and their sympathizers exchange ideology, spread propaganda, recruit members, and plan attacks. The terrorists, extremists, and their sympathizers can benefit from web techniques and online communities. They exchange ideology, spread propaganda, recruit members, and even plan attacks through the online community. Especially, because of the ubiquity of the online community, the previously isolated terrorists/extremist cells are able to collaborate more efficient than any time before and to form a more compact community virtually. Dark webs contain rich information about the dark groups, such as ideologies, recent topics, and news.

Several research works have been conducted to analyze web of terrorist cells or criminal activities. M. Sparrow (1991) applied social network analysis to criminal activities and observed three problems associated with criminal network analysis. They are incompleteness of analyzing data as a result of missing nodes and links that the investigators will not uncover, fuzzy boundaries resulting from the difficulty in deciding who to include and who not to include, and the dynamic property of analyzed networks. V. E. Krebs (2001) uses public information reported in major newspapers such as the NewYork Times and the Wall Street Journal to map networks of terrorist cells. Their research unrevealed a picture of a covert network after the tragic events of September 11, 2001. P. Klerks (2001) describes the development of criminal network analysis. The approaches start from manual analysis. An analyst constructs an association matrix by identifying criminal associations from raw data. Then a graphic-based approach is proposed to automatically generate graphical representation of criminal networks. Recently social network analysis has been used to provide more advanced analytical functionality to assist crime investigation. J. Xu and H. Chen (2005) and Koch et al. (2007) use data mining techniques to reveal various structures and interactions within a network. Discovering topics from dark websites helps in developing effective combating strategies against terrorism or extremists. The latent or topics are buried in large-scale web pages and hosted by dark websites. This work employs information retrieval (IR) techniques to discover hidden topics in a known dark web, such that the discovered latent topics can provide insights into social communities.

Modeling text corpora extracted from websites help find short description of a topic such that essential statistical relationships are

preserved from for the basic tasks such as classification, summarization, and similarity judgment (Sparrow 1991). In the field of information retrieval, a basic vocabulary of words or terms is chosen, and each document in the corpus is reduced to a vector of real numbers, each entry representing ratios of word counts. In the popular *tf-idf* scheme (Kerbs 2001), term frequency (tf) count is compared to an inverse document frequency (idf) count, which measures the number of occurrences of a word in the entire dataset. The tf-idf scheme generates a term-by-document matrix X whose columns contain the tf-idf values for each of the documents in the corpus. Latent semantic indexing (LSI) (Deerwester et al. 1990) is proposed to further reduce description length and reveal more inter- or intra document statistical structure. LSI uses a singular value decomposition of the X matrix to identify a linear subspace in the space of tf-idf features that capture most of the variance in the collection.

Hofmann (1999) presented probabilistic LSI (pLSI) model to model each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of topics. Thus, each word is generated from a single topic, and different words in a document may be generated from different topics. While Hofmann's word is a useful step toward probabilistic modeling of text, it provides no probabilistic model at the level of documents. Yang et al. (2009) discovered latent topics from the dark web by Latent Dirichlet Allocation (LDA) (Blei and Jordan 2003) which improves upon pLSI by placing a Dirichlet Prior on topic distribution to reduce overfitting and bias the topic weights from each document toward skewed distributions with few dominant topics.

Conclusion

The growth of online social communities, emanating from the old social gatherings of days before computing, has given us all a bonanza to and means to access information in amazing ways. Online communities have created opportunities for us unprecedented in the history of human where one individual can reach millions of others anywhere on the globe in seconds. The history and development of computing has made all this possible. However, with the easiness and abundance of resources at our disposal availed to us by online communities, there has also been evils that have been enabled by these large ecosystems. To be able to safeguard personal privacy, security, and dignity, we must pay special attention and develop protocols and best practices that must make everyone in these communities safely enjoy the experiences presented in these ecosystems. The battle is not yet worn and the way forward is not clear yet just because the next move in new technologies is not predictable.

Cross-References

▸ Consequences of Publishing Real Personal Information in Online Social Networks
▸ Corporate Online Social Networks and Company Identity
▸ Ethical Issues Surrounding Data Collection in Online Social Networks
▸ Online Social Network Privacy Management
▸ Privacy Preservation and Location-based Online Social Networking
▸ Topology of Online Social Networks
▸ User Behavior in Online Social Networks, Influencing Factors

References

Blei Ng, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022

Bylaws for Internet Corporation for Assigned Names and Numbers (ICANN) (2005). www.icann.org/general/bylaws.htm. Retrieved April 2013

Chew M, Balfanz D, Laurie B (2013) (Under)mining privacy in social networks, Google Inc. http://w2spconf.com/2008/papers/s3p2.pdf. Retrieved April 2013

Deerwester S, Dumais S, Landauer T, Furnas G, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf Sci 41(6):39–407
Evolving the High Performance Computing and Communications Initiative to Support the Nation's Information Infrastructure—Executive Summary (2013). http://www.nap.edu/openbook.php?record_id=4948. Retrieved April 2013
Fox R (2000) News track: age and sex. Commun ACM 43(9):9
Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of the twenty-second annual international SIGIR conference, Berkeley
Kizza JM (1999) Ethical and social issues in the information age. Springer, New York
Kizza JM (2005) Computer network security. Springer, New York
Kizza JM (2013) Ethical and social issues in the information age, 5th edn. Springer, London
Klerks P (2001) The network paradigm applied to criminal organizations: theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands. Connections 24(3):53–65
Koch M, Richter A, Schlosser A (2007) Services and applications for IT-supported social networking in companies. Wirtschaftsinformatik 6/49:448–455
Kramer RM (1999) Trust and distrust in organizations: emerging perspectives, enduring questions. Ann Rev Psychol 50:569–598
Krebs VE (2001) Mapping networks of terrorist of cells. Connections 24:43–52
Sparrow MK (1991) The application of network analysis to criminal intelligence – an assessment of the prospects. Soc Netw 13:251–274
Walsh M (2013) Court backs student on Facebook page criticizing teacher, NewsWeek. http://blogs.edweek.org/edweek/school_law/2010/02/court_backs_student_on_faceboo.html. Retrieved April 2013
Web Surpasses One Billion Documents: Inktomi and NEC Research Institute Complete First Web Study (2000) Inktomi News & Events, Jan 2000
Wresch W (1996) Disconnected: haves and have-nots in the information age, "Information Age Haves and Have-Nots." Rutgers University Press. ISBN-10:0813523702
Xu J, Chen H (2005) Analyzing and visualization. Commun ACM 48:100–107
Yang L, Liu F, Kizza JM, Ege RK (2009) Discovering latent topics from dark websites. In: IEEE symposium on computational intelligence in cyber security. IEEE Xplore, Nashville, TN

Social Indexing

▸ Folksonomies

Social Influence

▸ Web Communities Versus Physical Communities

Social Influence Analysis

Tiziana Guzzo, Fernando Ferri, and Patrizia Grifoni
Institute of Research on Population and Social Policies – IRPPS, National Research Council – CNR, Rome, Italy

Synonyms

Social networks members; Social networks users

Glossary

Confounding Variables Unknown variables exist (e.g., common location, gender, school, and several other external factors), which may cause friends to behave similarly with one another

Correlation Factor Correlation between variables is a measure of how well the variables are related. The most common measure of correlation in statistics is the Pearson correlation

Edge-Reversal Test Reserves the direction of all edges. Social influence spreads in the direction specified by the edges of the graph, and hence reversing the edges should intuitively change the estimate of the correlation

Homophily A user in the social network tends to be similar to his/her connected neighbors

Induction An action of a user is triggered by an action of another user

Selection People tend to create relationships with other people who are already similar to them

Shuffle Test Shuffles the activation time of users. It is based on the idea that influence does not play a role, and then the timing of activation

should be independent of the timing of activation of others

The Influence Maximization Problem Aims to identify an initial set of users in a social network that could maximize the spread of influence such that other users will adopt the new product in the shortest time

Definition

Social influence refers to change of a person's behavior after an interaction with other people, organizations, and in general society. It consists of the process by which the individual opinions can be changed by the influence of other individual(s) (Friedkin 1998). It is characterized by three main features:

- Conformity, that occurs when an individual expresses a particular opinion in order to meet the expectations of a given other, though he/she does not necessarily hold that belief that the opinion is appropriate
- Power, that is the ability to force someone to behave in a particular way by controlling his/her outcomes
- Authority, that is the power that is believed to be legitimated by those who are subjected to it

Webster's dictionary defines influence as "the power or capacity of a person or things in causing an effect in indirect or intangible ways." It could be defined as the combination of all things that may change or have some effects on a person's behavior, thoughts, actions, or feelings. It can be represented by peer pressure, persuasion, marketing, sales, and conformity.

This phenomenon in social networks refers to the behavioral change of individuals affected by others in a network. Social influence analysis in online social networks, studies people's influence by analyzing the social interactions between its members.

Introduction

Three broad categories of social influence were identified by Kelman (1958): (i) *compliance*, when people appear to agree with others while keeping their dissenting opinions private; (ii) *identification*, when people are influenced by someone who is liked and respected, such as a famous celebrity; and (iii) *internalization*, when people accept a belief or behavior and agree both publicly and privately.

The social environment and personal interactions have powerful effects on human behavior that in fact is always influenced by each other.

In literature three types of reference group influences are identified: *informational influence, utilitarian influence,* and *value-expressive influence* (Park and Lessig 1977; Bearden and Etzel 1982):

- *The informational influence* acts when individual would like to improve its knowledge and have best and useful information in order to optimize its choices (Kelman 1961).
- *The utilitarian influence* is based on the compliance process and acts when individual would like to satisfy a group's expectation in order to achieve a favorable reaction from it (Kelman 1961).
- *The value-expressive influence* is based on the identification process and acts when individual would like to be similar to the group in order to belong to it (Kelman 1961).

The exponential growth of online social networks such as Facebook, Twitter, MySpace, Flickr, and Pinterest, Instagram is playing an important role in shaping the users' behavior on the Web. Fowler and Christakis (2008) introduced the theory of three degrees of influence to explain the great influence that social networks have on people' behavior. According to them, people have an influence on friends which in their turn influence their friends, meaning that actions can influence people they have never met. They claim that "everything we do or say tends to ripple through our network, having an impact on our friends (one degree), our friends' friends (two degrees), and even our friends' friends' friends (three degrees). Our influence gradually dissipates and ceases to have a noticeable effect on people beyond the social frontier that lies at three degrees of separation."

The probability to be influenced by an influencer depends on four factors:

– *Relevance* (the right information): the user's information needs have to coincide with the influencer's expertise.
– *Timing* (the right time): information has to be delivered when the user needed it.
– *Alignment* (the right place): few channel of overlap between the user and the influencer there must be.
– *Confidence* (the right person): users have to trust the influencer with respect to his/her information needs.

Historical Background

Social relationships are key components of human life, and they have been historically connected to time and space limitations; these restrictions have been partially removed with the Internet diffusion. In particular, the emergence of social networks has created a new social dimension where individuals can increase their social awareness interacting with old and new friends; share information about data, products, and services; and be more informed about different aspects of everyday lives anywhere and anytime. The interest in social network studies has been growing massively in recent years. Psychologists, anthropologists, sociologists, economists, and statisticians have given important contributions, making it actually an interdisciplinary research area. In the last years several methods to collect and visualize network data have been developed in order to analyze relationships between people, groups, and organizations.

In a social network, members (nodes associated with others nodes) are influenced by others for various reasons. Social influence is a directional effect from node A to node B. Some nodes can have intrinsically higher influence than others due to network structure. Social Network Analysis is the study of social relations among a set of actors (nodes). The nodes in the network are the people and groups, while the links show relationships or flows between the nodes. The analysis allows "to determine if a Virtual Social Network is tightly bounded diversified or constricted, to find its density and clustering, and to study how the behaviour of network members is affected by their positions and connections" (Scott 2000). The importance of a node in the network is measured by its centrality. The three most important individual centrality measures are (http://www.orgnet.com/sna.html):

- The degree centrality refers to the number of direct connections a node has.
- The betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes.
- The closeness centrality that is the inverse of farness, which in turn is the sum of distances to all other nodes.

A node with high centrality is usually more highly influential than other nodes. According Katona et al. (2011), demographics and user's position can predict their influential power on their neighbors. Social Network Analysis analyzes which members are individuals or peripheral in a network; it identifies bonding and bridging and who has influence in the network. Many mathematical techniques are available to measure networks (Wasserman and Faust 1994). Hoppe and Reinelt (2010) demonstrate how to use these metrics to understand and evaluate specific leadership networks.

Farrow and Yuan (2011) explored the strength of network ties to show how Facebook influences the attitudes of the alumni to volunteer for and make charitable gifts to their alma mater fortifying consistency between attitude and behavior.

Social influence analysis aims at qualitatively and quantitatively measuring the influence of one person on others. There are different methods and algorithms for measuring social influence, and they will be analyzed in the following sections.

Qualitative Measures for Analyzing Social Influence

According to Anagnostopoulos et al. (2008), influence of a person on another can act for three reasons: (i) induction, (ii) homophily, and (iii) confounding variables (factors). They applied

statistical analysis on data from a large social system in order to identify social influence as a source of correlation between the actions of individuals with social ties. They proposed two tests, the Shuffle test and the Edge-Reversal test, to identify induction as cause of social correlation when the time series of user actions is available. This approach is based on the assumption that timing of actions should matter if induction is a likely cause of correlation.

Goyal et al. (2010) followed a similar approach, proposing to establish relationship between users by scanning log of user action. According to them the influence probability between two users is determined by common actions and time issues. The approach based on homophily was followed by Crandall et al. (2008) that used cosine similarity to compute the similarity between two people. They proposed a probabilistic model which samples activities of people based on their history and those of their neighbors and a background distribution. This concept was stressed also by Matsuo and Yamamoto (2009). They studied user's behavior on an e-commerce site and found that users generally trust other users who have similar behavior with them. Other studies analyzed the correlation between social similarity and influence. Singla and Richardson (2008) studied the probability of relationships between two users by measuring their similarity. According to them users with common features (age, gender, zip code, word, and queries issued) chat more likely to each other; then influence probabilities could be estimated by user's similarity.

These studies used different approaches to analyze influence probabilities, but they did not address the issue of identifying influential users of the network. This issue will be analyzed in the next section introducing studies that used quantitative measures.

Quantitative Measures for Analyzing Social Influence

The problem to quantify the strength of social influences and differentiate social influences from different angles (topics) was addressed by Tang (2009). They studied the topic-based social influence analysis on large networks. The goal was to simultaneously analyze nodes' topic distributions (or user interests), similarity between nodes (users), and network structure. They proposed a Topical Factor Graph (TFG) model to incorporate all information into a unified probabilistic model and present Topical Affinity Propagation (TAP) for model learning.

Most studies about social influence analysis considered positive interactions (agreement, trust) between individuals; Li et al. (2011) also considered negative relationships (distrust, disagreement) between individuals and conformity of people (the inclination of a person to be influenced). They proposed an algorithm called CASINO (Conformity-Aware Social INfluence cOmputation) which quantifies the influence and conformity of each individual in a network by utilizing the positive or negative relationships between individuals. This algorithm consists of three phases. In the first phase, a set of topic-based subgraphs that represent the social interactions associated with a specific topic are extracted from a social network. In the second phase, the edges (relationships) between individuals are labeled with positive or negative signs. Finally, in the third phase, the influence and conformity indices of each individual in each signed topic-based subgraph are computed.

The problem of dynamic social influence analysis was addressed by Wang et al. (2011). They proposed a pairwise factor graph (PFG) model to quantify the influence between two users in a large social network. Different types of factor functions capture information such as users' attribute information, social similarities/weights, and network structures, which form the basic components of the factor graph model. An algorithm was designed to learn the model and make inference to obtain all the marginal probabilities. They further proposed a dynamic factor graph (DFG) model to incorporate the time information.

Domingos and Richardson (2001) first studied the problem of which individuals is necessary to

target to have a large cascade of further adoptions. The problem was considered in a probabilistic model of interaction; heuristics were given for choosing customers with a large overall effect on the network. Kempe et al. (2003) faced the same issue of choosing influential sets of individuals by formulating it as a discrete optimization problem and proposing an approximation algorithm that was applicable to general cases. This optimization problem has a complexity NP and "the greedy algorithm can guarantee the influence spread within (1-1/e) of the optimal influence spread."

Kimura and Saito (2006) propose shortest-path-based influence cascade models and provide efficient algorithms to compute influence spread. However, these algorithms are not scalable for large graphs; to solve the problem Chen et al. (2009, 2010) designed a new heuristic algorithm. This algorithm, scalable to millions of nodes and edges, allows controlling the balance between the running time and the influence spread of the algorithm. With respect to the work of Kimura that used simple shortest paths on the graph, which are not related to propagation probabilities, Chen used maximum influence paths and local structures such as arborescences.

Key Applications

In social networks very important is the effect of "word of mouth," since idea, opinions, and recommendations propagate very quickly and with an exponential grow. This concept is very frequently applied in different fields like marketing, recommendations, healthcare, and politics.

Many companies have recently started to capture data on the social interaction between consumers in social networks, with the objective of understanding and leveraging how this interaction can generate social influence. Consumers can really modify their opinions about products and/or services according to the social influence process; this process also impacts on knowledge diffusion about products and services. Social network emerges as one of the most authoritative and influential sources of knowledge about products and services related to the area of interest of a community. They have the aptitude to generate knowledge sharing among consumers and facilitate the collaboration and exchange of ideas among consumers. In this context, viral marketing involves customers in commercial strategies for recommending commercial products to their friends through the customer social networks. According to De Bruyn and Gary (2004), viral marketing is a "consumer-to-consumer (or peer-to-peer) communication, as opposed to company-to-consumer communications, to disseminate information about a product or a service, hence leading to its rapid and cost-effective market adoption." In this context the problem of the influence maximization that aims to identify individuals to target to have a large cascade of further adoptions assumes a great relevance. Several studies introduced in the previous section (such as Domingos and Richardson 2001; Kempe et al. 2003; Kimura and Saito 2006; Chen et al. 2009, 2010) addressed this issue.

The emergence of e-commerce has led to the development of recommender system, a personalized information filtering technology used to identify a set of items that will be of interest to a certain user. Mao et al. (2012) explored social influence for item recommendation. Previous approaches mostly incorporated social friendship into recommender systems by heuristics. They captured quantitatively social influence and proposed a probabilistic generative model, called social influenced selection (SIS), extracting social influence and preferences through statistical inference. Moreover, they developed a new parameter learning algorithm based on expectation maximization (EM) to face the problem of multiple layers of hidden factors in SIS.

Social networks are rapidly transforming also the healthcare field. People are more and more connected to the Web in order to search, share, and exchange information and find support from

other people. According to a 2008 survey carried out by Icrossing, the Internet has been the most used source to find information about health and wellness in the previous 12 months. Patients, thanks to the Web, can share the same illness with people all over the world and feel themselves less alone. In addition according to Edelman's study (2008), people have more trust in a person with which they can identify themselves than business, government, and media subjects. Which is the impact of these activities on heath conditions? A study of Christakis and Fowler found that health status can be influenced by the health status of the neighbors. How do you manage the inaccurate information disseminating on health social networks? Some studies were carried out to identify influential users in order to optimize the spread of health information (Krulwich and Burkey 1995; Zhang et al. 2007).

Another emerging key application of social influence on social networks is the political field.

A strength of the first Obama election campaign was his strategic use of social media. Analysts are now studying the impact of tools such as Facebook, Twitter, and YouTube had on election results. A recent Pew Research study (Rainie and Smith 2012) analyzed politics on social networks and found that users after discussing a political issue or reading posts about it on these sites change their points of view and political involvement. Bond et al. (2012) hypothesized that voting behavior is significantly influenced by messages on Facebook. They found that political self-expression, information seeking, and real-world voting behavior of millions of users were directly influenced by messages. This had an indirect effect through social contagion also in the users' friends and friends of friends. Close friends had four times more influence than the message itself. Furthermore, they stated that "online mobilization works because it primarily spreads through strong-tie networks that probably exist offline but have an online representation."

Fowler (2005) based on observational data found that behavior of each act of voting spreads through the network generating on average an additional three votes.

Future Directions

Social influence analysis studies are in its beginnings, and so in the future more methods and techniques will be developed. A challenge for future works will be to develop efficient, effective, and quantifiable methods for analyzing the persuasion and influence phenomenon within social networks. Until now, studies have mainly focused on conceptual models and small-scale simulations. In the future as online social networks enable for the first time to measure social influence over a large population, they should include more large-scale data mining algorithms to analyze social network data. It will allow having more realistic results for large-scale applications in different fields and in different social and informational settings.

Cross-References

▸ Centrality Measures
▸ Friends Recommendations in Dynamic Social Networks
▸ Mobile Communication Networks
▸ Modeling Social Preferences Based on Social Interactions
▸ Origins of Social Network Analysis
▸ Political Networks
▸ Questionnaires for Measuring Social Network Contacts
▸ Recommender Systems: Models and Techniques
▸ Recommender Systems Using Social Network Analysis: Challenges and Future Trends
▸ Social Recommendation in Dynamic Networks
▸ Social Recommender System
▸ Trust in Social Networks
▸ User Behavior in Online Social Networks, Influencing Factors

References

Anagnostopoulos A, Kumar R, Mahdian M (2008) Influence and correlation in social networks. In: KDD'08, Las Vegas, pp 7–15

Bearden WO, Etzel MJ (1982) Reference group influence on product and brand purchase decisions. J Consum Res 9:183–194

Bond RM, Fariss CJ, Jones JJ, Kramer ADI, Marlow C, Settle JE, Fowler JH (2012) A 61-million-person experiment in social influence and political mobilization. Nature 489:295–298

Chen W, Wang Y, Yang S, (2009) Efficient influence maximization in social networks. In: KDD'09, Paris

Chen W, Wang C, Wang Y (2010) Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: KDD'10, Washington, DC

Crandall D, Cosley D, Huttenlocher D, Kleinberg J, Suri S (2008) Feedback effects between similarity and social influence in online communities. In: KDD'08, Las Vegas, pp 160–168

De Bruyn A, Gary LL (2004) A Multi-Stage model of word of mouth through electronic referrals. eBusiness Research Center Working Paper

Domingos P, Richardson M (2001) Mining the network value of customers. KDD'01, San Francisco, pp 57–66

Edelman (2008) Edelman trust barometer. www.edelman.com/trust/2008. Accessed 18 Sept 2012

Farrow H, Yuan YC (2011) Building stronger ties with alumni through Facebook to increase volunteerism and charitable giving. J Comput Mediat Commun 16(3):445–464

Fowler JH (2005) Trunout in a small world In: Zuckerman AS (ed) The Social logic of politics: personal networks as contexts for political behavior. Temple University Press, Philadelphia, pp 269–287

Fowler JH, Christakis NA (2008) The dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham heart study. Br Med J 337:a2338

Friedkin N (1998) A structural theory of social influence. Cambridge University Press, Cambridge. doi:10.1017/CBO9780511527524

Goyal A, Bonchi F, Lakshmanan LVS (2010) Learning influence probabilities in social networks. In: Proceedings of the proceedings of the third ACM international conference on Web search and data mining, New York. ACM, New York, pp 241–250

Hoppe B, Reinelt C (2010) Leadersh Q 21:600–619

Icrossing (2008) iCrossing's how America searches: health and wellness. Retrieved 10 May 2009

Katona Z, Zubcsek PP, Sarvary M (2011) Network effects and personal influences: the diffusion of an online social network. J Mark Res 48(3):425–443

Kelman H (1958) Compliance, identification, and internalization: three processes of attitude change. J Confl Resolut 2(1):51–60

Kelman HC (1961) Processes of opinion change. Public Opin Q 25:57–78

Kempe D, Kleinberg J, Tardos E (2003) Maximizing the spread of influence through a social network. In: KDD'03, Washington, DC, pp 137–146

Kimura M, Saito K (2006) Tractable models for information diffusion in social networks. In: Proceedings of the 10th European conference on principles and practice of knowledge discovery in databases, Berlin, pp 259–271

Krulwich B, Burkey C (1995) Contact finder: extracting indications of expertise and answering questions with referrals. In: Symposium on intelligent knowledge navigation and retrieval, Cambridge, pp 85–91

Li H, Sourav SB, Sun A (2011) CASINO: towards conformity-aware social influence analysis in online social networks. In: Proceedings of the 20th ACM conference on information and knowledge management, CIKM 2011, Glasgow, October 24–28

Mao Y, Xingjie L, Wang-Chien Lee (2012) Exploring social influence for recommendation – a generative model approach. In: Proceeding SIGIR'12 proceedings of the 35th international ACM SIGIR conference on research and development in information retrieval, Portland, pp 671–680

Matsuo Y, Yamamoto H (2009) Community gravity: measuring bidirectional effects by trust and rating on online social networks. In: Proceedings of the 18th international conference on world wide web, Madrid. ACM, New York, pp 751–760

Park W, Lessig VP (1977) Students and housewives: differences in susceptibility to reference group influence. J Consum Res 4:102–110

Rainie L, Smith A (2012) Politics on social networking sites. Pew Research Center's internet & American Life Project. http://pewinternet.org/~/media/Files/Reports/2012/PIP_PoliticalLifeonSocialNetworkingSites.pdf. Accessed 4 Sept 2012

Scott J (2000) Social network analysis: a handbook. Sage, London

Singla P, Richardson M (2008) Yes, there is a correlation: from social networks to personal behavior on the web. In: WWW'08, Beijing, pp 655–664

Tang J (2009) Social influence analysis in large-scale networks. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, Paris. ACM, New York, pp 807–816

Wang C, Tang J, Sun J, Han J (2011) Dynamic social influence analysis through time-dependent factor graphs. In: International conference on advances in social networks analysis and mining Kaohsiung, Taiwan, 25–27 July 2011, IEEE Computer Society, Los Alamitos

Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge

Zhang J, Ackerman M, Adamic L (2007) Expertise networks in online communities: structure and algorithms. In: Proceedings of the 16th international conference on World Wide Web, Banff. ACM, New York, pp 221–230

Social Interaction

▸ Incentives in Collaborative Applications
▸ Networks at Harvard University Sociology
▸ Origins of Social Network Analysis
▸ Privacy and Disclosure in a Social Networking Community

Social Interaction Analysis for Team Collaboration

Ognjen Scekic, Mirela Riveni, Hong-Linh Truong, and Schahram Dustdar
Distributed Systems Group, Vienna University of Technology, Vienna, Austria

Synonyms

BPEL4People; Collaboration analysis; Collaboration metrics; Collaboration platforms; Crowdsourcing; Human-Based services (HBS); Interaction patterns; Process mining; Rewarding; Social trust; Task assignment; Team collaboration; Team formation

Glossary

Actor Entity (human or computer) possessing a capability to act intelligently and process specific assignments (activities/tasks)

Task Piece of work to be solved, typically complex enough to require knowledge or processing power of a large number of individual actors

Atomic Task Task that can be handled by an individual actor

Composite Task Task that must be handled by multiple actors due to size or complexity. A composite task can be broken down into atomic tasks

Collaborative Process (Collaboration) Joint effort of a (limited) number of actors with the goal of performing a task. A collaborative process has a limited duration and requires coordination among actors (due to task dependencies)

Team Set of actors taking part in a collaborative process. Team lifetime is considered equal to the lifetime of the collaborative process

Task Assignment The art to divide a (composite) task into (sub) tasks and assign them to appropriate actors

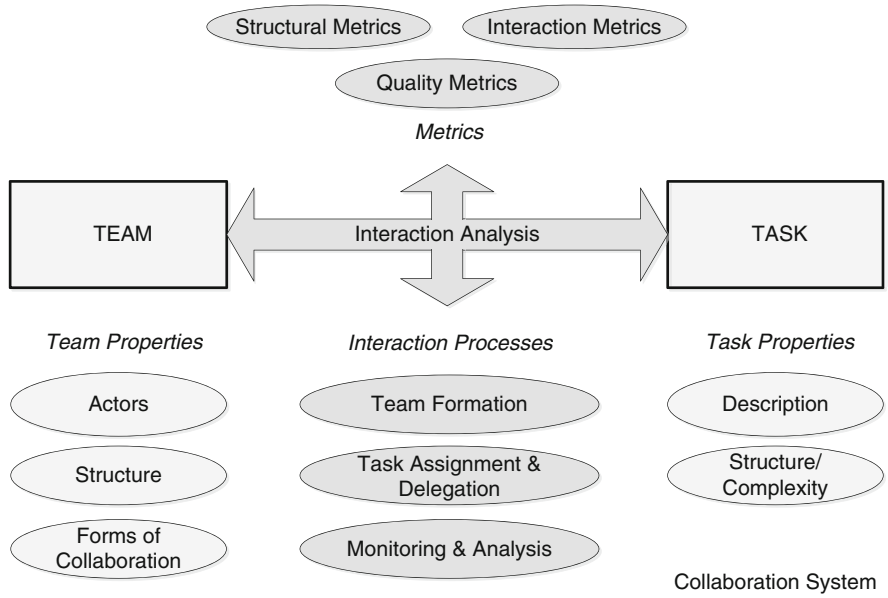
Team Formation Process consisting of identifying appropriate actors for performing all atomic tasks and establishing of internal coordination and functioning rules in the team

Metric Precisely defined, context-specific measure of some properties

Collaboration System (Platform) Information system supporting execution of collaborative processes

Definition

With the advent of Web 2.0 and social networks, millions of users around the world were given the opportunity to collaborate, share ideas, and coordinate their efforts easier than ever before. These developments lead to an increased interest to exploit these opportunities, both in the research community and in the industry. Such collaborative efforts are supported by different types of *collaboration systems*, providing automated or semiautomated actor management (e.g., modeling, reputation, and rewarding), task management (e.g., modeling, creation, division, scheduling, aggregation, and monitoring), and process execution environment (e.g., actor communication and coordination).



Social Interaction Analysis for Team Collaboration, Fig. 1 Elements of a collaboration system

In order to better understand how these systems work, in this entry we look into different types of collaboration systems. We describe team structures and discuss different forms of collaborations they support. In particular, we focus on interaction processes that are supported by the system and discuss different metrics used to describe and analyze such systems. Figure 1 depicts the fundamental elements of a collaboration system that we discuss in this entry.

Introduction

The idea of combining research on how humans work, communicate, and cooperate and the research on how computer systems can efficiently support such collaborations led to the creation of an interdisciplinary research area known as *computer-supported cooperative work* (CSCW) in the 1980s (Grudin 1994). Initially, the research was focused on small-scale collaborations, e.g., within companies or interest groups. With the wide adoption of Internet technologies, service-oriented architectures (SOA), mobile and cloud computing, and especially social networks, nowadays it is possible to carry out large-scale collaborations, possibly involving thousands of collaborators across boundaries of multiple organizations and countries. Some examples of today's well-established types of computer-supported human collaboration systems include:

- *Human Computation Systems* – Systems in which human actors perform assigned tasks in a precisely defined sequence (e.g., by following an algorithm). The execution is explicitly controlled and coordinated by the system and expected to yield precise results (Law 2011).
- *Workflow Management Systems* – Systems that allow modeling of tasks and their execution scenarios. Notable representatives of such systems are the various business process management (BPM) systems. Although tasks can be performed by human actors, the traditional understanding of the notion of a workflow system does not include an integrated management of human-performed tasks.
- *Mixed Systems* – Systems where both human and computer actors process the tasks. Humans are deeply integrated into the system, making both types of actors first-class citizens of it. The decision on who processes a particular task can be made by the system.

While computer-performed tasks are accurate, employing humans for certain tasks requires dealing with uncertainties both in terms of human behavior and the quality of results.

- *Crowdsourcing Systems* – Systems in which the task is offered, rather than assigned explicitly, to an unknown and usually large group of people who can freely accept and perform the tasks (also see the ▶ Social Network Analysis of Crowdsourcing chapter).

These types of systems clearly enable different collaboration types. Depending on the type of system and type of problem to be solved, different team structures are possible. The team structure guides the interactions and collaboration among team members and consequently plays an important role in a team's performance. Therefore, this entry explores team formation processes and team collaboration types. We discuss three main team collaboration types: static, ad hoc, and open collaboration. We then focus on interaction analysis and discuss appropriate interaction metrics.

Team Collaboration Analysis

Team Properties

We consider three important team properties: (a) *actors* making up the team, with their different skills, qualities, and personalities; (b) *structure*, which represents a set of interaction paths among the actors; and (c) different *forms of collaboration* among the actors.

Actors and Team Structure

Actor teams are usually modeled as undirected or directed graphs with nodes representing people or teams of people and edges representing social relationships between them (Newman 2010). Often, the edge is associated with a weight describing the amount of interaction between the two nodes it connects and annotated with a context, representing the type of the relationship (e.g., friendship, prior professional collaboration, and trust). Therefore, a team network can be modeled as a graph consisting of nodes representing actors, sets of skills forming their profiles, edges representing relationships, and associated contexts of relationships (Caverlee et al. 2008).

Forms of Computer-Supported Team Collaboration

Static Collaboration

Static collaboration is characterized by well-defined, long-lasting/repetitive processes (tasks), executed by human actors with specific assigned roles. Such kind of collaborations is usually found in companies that encode and execute their daily business use cases as business processes by using workflow technologies. This collaboration type makes no use of the underlying social networks connecting the actors to alter or enhance the collaboration in any way. As such, this approach works well only in cases where the predictability of the process execution is high and where no adaptability is required.

Ad Hoc Collaboration

Unlike static collaboration, the ad hoc collaboration is suitable when performing highly dynamic tasks that change in time or complex tasks that occur only once and are not repeated. In this type of collaboration, tasks are initially defined, but the actors performing them are provisioned only at runtime. Ad hoc collaborations often cross organizational boundaries and are distributed in nature, in terms of software services used and actors executing the tasks as well as in terms of control. Actor provisioning can be fully automated or partially performed by the actors themselves, often relying on social and other underlying networks connecting the actors.

Ad hoc collaborations are primarily supported by SOA-based collaboration systems. One approach to abstract human actors as services in mixed systems is through Human-Based Services (HBS). However, HBS are still not considered as a mature technology. Another approach to building ad hoc collaborations is to build upon existing crowdsourcing platforms and extend them with necessary features.

Open Collaboration

In open collaborations a task can be actively shaped by the actors. The actors (often belonging to a professional community or an interest-based community) contribute freely to the task resolution during runtime. A task is not strictly assigned to a particular actor, but instead it is editable by (m)any community members upon their wish. In this case the coordination between the actors can affect the quality of the task (Kittur and Kraut 2008). Data quality is controlled by the system itself and/or by a designated entity, but the quality is mainly evaluated by feedback information from actors. Open collaboration is particularly suitable for longer running, best-effort tasks, with no strict quality and time constraints, but requiring distributed know-how.

Open-source development, Wikipedia, and community-based Q&A Web sites are among the best examples of open collaboration. Examples of open collaboration enabling technologies and platforms include cloud services (e.g., Amazon EC2), sharing and collaboration platforms (e.g., DropBox, Google Docs, and Mendeley) and open-source repositories (e.g., GitHub and SourceForge).

Task Properties

Task Description

Considering the general nature of the tasks that can be handled by a team composed of human actors, describing tasks precisely and unambiguously is extremely difficult. The difficulty lies in expressing the information that needs to be interpreted by each actor in the same way. At the same time, the effort required to interpret a task's objectives must be considerably smaller compared to the effort required to perform the task itself.

Two different approaches for task description are *informal* and *formal*:

- Informally describing tasks means expressing the required outcomes in natural language, accompanied with simple examples. This approach is usually taken by today's crowdsourcing platforms that handle simple tasks. Also, informal description may be preferred in cases where tasks require aesthetic judgment or when the required outcome of the task is too vague to be expressed more precisely (e.g., on Web sites running creativity contests).
- Formally describing tasks means employing a specific notation that precisely defines how the task should be processed and what should the outcome be. Formal task description is usually used in specific environments, most notably in business process modeling (BPM). Initial versions of the most prominent business control-flow languages, such as BPEL, did not support specification and invocation of human interactions. An extension to BPEL, known as BPEL4People (Kloppmann et al. 2005), was proposed in 2005 to allow modeling of human interactions within business processes by introducing the concept of *people activities*. A people activity can be described according to the WS-HumanTask (Amend et al. 2007) specification. In this way, humans can be internally represented as Web Services and integrated into the system.

Task Structure and Complexity

Task structure directly influences the team structure. Different task structures and complexities demand specific types of collaboration in terms of communication form, coordination protocols, adaptation schemes, and outcome type.

Subtask interdependencies are one of the fundamental factors determining the task structure and task complexity. Tasks can be *parallel* and *sequential*. Parallel tasks contain subtasks that can be executed independently in parallel, while a sequential task is composed of subtasks whose execution must follow a strict order. A subtype of sequential tasks is *iterative* tasks, where the output of one actor is given as input to another actor for subsequent task execution. An experiment and analysis of parallel and iterative approaches in open systems can be found in Little et al. (2010).

Apart from subtask interdependencies, other, nonstructural factors can influence a task's complexity, such as (a) number of atomic tasks; (b) growth (Dustdar and Bhattacharya 2011) – the number of atomic tasks can grow in runtime, necessitating team-size adaptability; and (c) task

cardinality – tasks can be designed to be executed by one or many actors in one-to-one, many-to-one, many-to-many, and few-to-one fashion. See Quinn and Bederson (2011) for details and examples.

Interaction Processes

Team Formation

The problem of team formation consists of selecting suitable actors to perform a given task (out of a larger group of available actors) and organizing them in a collaborative structure. The first problem with identifying "suitable" actors is that suitability is highly contextdependent and difficult to define precisely. Furthermore, suitability can have many different aspects. For example, the minimal suitability requirement for an actor is to possess the skills to perform the task. But, at the same time, for a successful teamwork, factors like trust, motivation, experience, and personal relations with other team members can be equally important.

Initially, the research focused on locating individual best-matching actors for a required set of skills and other individual properties. However, a group of top individuals does not guarantee the quality of their collaboration. Subsequent research efforts began taking into account the underlying social relations among the actors (e.g., friendships, managerial relations, previous business interactions, interests, connectedness, and social trust). After selecting suitable actors, the next step in ensuring a successful collaboration is setting up a collaborative organization and environment. Although collaboration patterns in a team often resemble those in the underlying social networks, other factors like coordination cost, userpreferences, and context are also important.

Whichever the properties considered, they are always measurable and quantifiable, meaning that the problem of team formation can be ultimately expressed as an optimization problem where we want to optimize certain performance aspects of the team as a whole (speed, quality, cost, and response time). In general, team formation can be as follows:

- *Self-organizing* – The actors themselves lead the team formation in a collective-intelligence fashion and set up the collaboration environment.
- *Centralized* – Team formation and setting up of collaborative environment is managed by the system.

Wikipedia and open-source community are striking examples of how self-organizing teams can perform well. The assumption is that the actors taking part in collaboration will perform best if they are given the possibility to modify and adapt the collaborative environment. This includes also the initial team formation. For example, in Gaston and DesJardins (2005), the authors investigate a system that enables actors to locally modify their collaborative environment according to their social network preferences (i.e., to *rewire* the local network topology) with the goal of achieving globallynoticeable, collective performance improvement.

The most problematic aspect of self-organizing teams is the discrepancy between local and global effects. Although we rely on the collective intelligence of the actors, in practice, actors may not know how or when to modify the local network to achieve global improvements, since their actions are based upon their partial views only.

Centralized team formation is entirely handled by the system. Internally, the system can employ an algorithm or human actors to assemble the team:

- *Human-managed team formation* relies on human actors offering their referrals and recommendations via Web Services, thus leveraging crowdsourcing techniques to identify the best candidates from their social networks. An example of such a system is PeopleCloud (Lopez et al. 2010).
- *Algorithmic team formation* relies on an algorithm to select actors and assemble the team. A lot of research efforts have been directed in this sense, producing a number of different algorithms. In Schall and Dustdar (2010), the authors modify the well-known page ranking algorithms PageRank and HITS to identify the best team members, based on

their previous interactions. In Lappas et al. (2009) and Anagnostopoulos et al. (2012), the goal is to minimize the total coordination cost of the newly established team, while in Dorn and Dustdar (2010), the optimal team is chosen as a tradeoff between skill coverage and actor connectivity. In Caverlee et al. (2010), the social trust between the team members is regarded as the most important factor in forming efficient collaborations.

Task Assignment and Delegation

Routing and Delegations

Task delegation mechanisms are being explored as forms of coordination and load balancing in human computation. The concept of *social routing* is introduced in Dustdar and Gaedke (2011) as a form of delegation of tasks by task owners to actors from their social, professional, other context-based community networks or the crowd. The so-called social routine can be a software service that actually does the task forwarding across different types of networks depending on the requirement of the actor wishing to delegate the task.

Historical data on delegations (e.g., the executed/delegated tasks ratio) can serve as a good indicator of actor's role and performance qualities. For example, a high number of task delegations testify of a coordinating/managing role. On the other hand, if an actor has a very large number of delegated tasks but a low number of executed tasks, it can be inferred that the actor is lazy. Thus, delegation data can be used as metrics in actor selection and team formation algorithms. Moreover, delegation measures can be used in trust inference mechanisms. If the receivers of delegated tasks are considered trustworthy, new trust-based links will be created between the delegator and the delegates (Skopik et al. 2010).

Delegation Patterns in Business Process Activities

The four main delegation patterns, detailed in Kloppmann et al. (2005), are as follows:

- *Nomination* pattern allows predefined actor(s) to decide to whom to assign a task.
- *Escalation* pattern allows transfer of responsibility for task execution to other human actors when the originally assigned actor cannot meet task's time constraints.
- *Chained execution* pattern forces the actors to perform a specific sequence of actions, where the concrete actions may be determined only in runtime.
- *Four-eye Principle* pattern allows two actors to take a public or a private decision on the same issue independently (*separation of duties*).

Algorithmic Task Life-Cycle Management

In cases when subtasks are clearly delimited and subtask dependencies are static and do not change in time, parallelizing a task execution is fairly easy. Some application domains, such as crowdsourcing systems, are characterized by exactly such properties. This has led researchers to dedicate a lot of effort to automate task life-cycle management transparently for the programmer, by developing a number of programming language extensions/libraries that work on top of existing commercial crowdsourcing systems, such as Amazon Mechanical Turk. The extensions are typically able to automatically split a task; to assign/offer the subtasks to the actors in the crowd respecting the dependency, cost, and time constraints; and to merge the processed subtasks into the final resulting task. Additionally, automated quality control processes may be also offered. Most commonly, these are based on peer reviews or on a combination of redundant processing and majority rule. For example, an image that needs to be tagged may be submitted to multiple actors, but the aggregated result will contain only tags suggested by multiple independent actors. The data quality requirements can have a direct influence on task assignment, as they may introduce assignments not explicitly required by the user, but performed transparently by the system. In fact, the main purpose of algorithmic handling of task assignment is exactly to move the burden of task life-cycle management from the user to the system.

Collaboration systems can manage task assignments automatically throughout the entire

execution time, repeating them when needed. For example, Little et al. (2009) shows a system offering the possibility of iterative task execution, by reassigning previously processed tasks a number of times in order to improve the final quality of work by incrementally building upon previous work. In Marcus et al. (2011) a system can autonomously decide when to assign pleasing tasks to specific actors in order to motivate/reward them.

Another major advantage of algorithmic task assignment is the cost optimization. For large-scale collaborations, the system is able to assign the tasks in such a way to reduce the coordination costs better than human managers could do. For example, the task can be assigned to actors possessing similar professional skills and backgrounds, or the system can adjust task prices and time allotments based on the feedback obtained from monitoring data (Barowy and Berger 2012).

Collaboration Monitoring and Analysis

Monitoring and analyzing collaborative processes is necessary to gather important metrics regarding the performance of teams and actors and the quality of processed tasks. Such metrics are then used to detect bottlenecks, improve performance, and decide on appropriate compensation of the actors. As these metrics play a fundamental role in determining overall collaboration efficiency and costs, every collaboration system must support some kind of monitoring and analysis functionalities.

Monitoring can be performed during the runtime of a collaborative process (*active monitoring*) or it can be performed post-runtime, e.g., by *log mining*. Log mining is usually considered a part of more complex analysis processes, known as *workflow/process mining* (van der Aalst 2011; Zhang and Serban 2007).

Active monitoring is suitable for detecting anomalies that require quick responsive actions and team adaptations. An example of monitoring and analyzing SOA-based collaborative processes can be found in Truong and Dustdar (2009).

Log mining, on the other hand, is used to gather less obvious information about the internal functioning of the team, since it considers the backlog of all recorded actions performed during previous collaborations. This allows discovery and prediction of critical execution paths, expected workload distribution, actor performance, and identification of previously unknown collaborative social networks, e.g., the network of most trusted colleagues or the groups of workers that together collaborate most efficiently as a team.

Collaboration Metrics and Patterns

Metrics characterizing collaborations can be divided into three major categories (see Table 1):

- *Structural metrics* – Defining the mathematical properties of the social/collaborative network connecting the actors
- *Interaction metrics* – Defining various properties of individual actors or actor groups, emerging as the result of past interactions
- *Quality metrics* – Defining quality criteria for actor performance and for task outcome data

Structural Metrics and Network Patterns

Structural metrics and network patterns are based on mathematical properties of the social graph connecting the actors in a collaboration team. They provide useful insights into the functioning and self-organization of actors in a team. Structural metrics are well researched. Here is a brief overview of some of the main structural metrics:

- *Centrality measures* – They include various metrics that identify the importance of an actor within a network in different contexts of importance. Some of the most important centrality metrics are *degree centrality*, *closeness centrality*, *betweenness centrality*, and *eigenvector centrality*. See also Chaps. Centrality Measures of Social Networks and ▶ Similarity Metrics on Social Networks.
- *Structural groups* – They refer to various group patterns that can be identified within networks, such as *core* (denoting a subset of actors within a network where each actor is connected to at least k other actors within the same subset), *k-component* (denoting a subset of actors in which each two actors are connected by at least k independent paths),

Social Interaction Analysis for Team Collaboration, Table 1 Overview of metrics and patterns used in collaboration systems

Structural metrics		Centrality measures (degree, closeness, betweenness, eigenvector, etc.)
		Structural groups (cores, components, cliques)
		Transitivity, reciprocity
		Similarity, equivalence
Interaction metrics	Actor level	Trust, reputation Functional/skill coverage
		Task familiarity, team familiarity
	Group level	Structural groups
		Team size
		Link quality, interaction intensity
		Collaboration patterns (delegations, escalations, redundant processing, iterative processing, etc.)
Quality metrics	Quality-of-Data (QoD)	Uncertainty, Completeness, accuracy, freshness, relevancy etc.
	Performance	Availability, response time, success rate, etc.
	Rewarding & incentives	Effort, productivity, quality of work

and *clique* (denoting a subset of actors all directly connected to each other).

- *Transitivity and reciprocity* – Transitivity reflects the "friend-of-a-friend" concept, i.e., if an actor *a* is connected by an edge to another actor *b*, and *b* is connected to *c*, then *a* is also connected to *c*. Reciprocity, on the other hand, denotes the probability that actor *b* points to actor *a* if actor *a* points to *b*.
- *Similarity* – It is defined by *structural equivalence* and *regular equivalence* metrics. See Newman (2010).

Details about all these and other metrics, as well as about ranking algorithms, can be found in Newman (2010).

Interaction Metrics

Interaction metrics can be defined at two levels: *individual level* (targeting individual actors) and *group level* (targeting multiple actors or the entire team). Individual interaction metrics describe a property of an individual actor that is shaped by the interaction in which the actor has participated. Group interaction metrics describe properties of particular interactions between actors, possibly including the collaboration as a whole.

Certainly, the most important actor-level metrics are *skill coverage* and *trust*. Skill coverage represents a degree to which an actor or a team possess necessary skills to perform a task. This metric is important because it describes how much a team's set of skills deviates from the optimal one for a given task. The problem of matching skills is equivalent to the problem of functional matching in Web Service compositions.

Trust, as a computational concept, was formalized in Marsh (1994), and since then it has been seen as a metric of great importance for selection of appropriate actors during the team formation phase. Trust is defined as an indicator of an actor's expectation about another actor's future behavior based on knowledge from previous interactions, which inherently involves a degree of uncertainty about this behavior and its outcomes. Trust is highly context dependent and one actor may have information about several scope-specific trust values for another actor. A scope can be the membership in a professional network,

social network, or a collaboration team. For more information, see entry. ▸ Computational Trust Models.

Inferring trust is important in several cases:

- For actor discovery and team formation algorithms, when determining actor suitability for specific tasks
- For team optimization, adaptation, and risk management purposes
- For delegation mechanisms, e.g., when selecting a collaborator that may be a part of the extended team structure for the purpose of load balancing in cases of unexpected load

We can distinguish three types of trust based on the type of actors and interactions that are taken into account for its inference:

- *Local trust* or *direct trust* (sometimes also called *private reputation*) – First-hand trust, inferred from the outcome of an actor's previous interactions with the trustee
- *Recommendations* – Second-hand trust inferred from the outcome of past interactions between a well-trusted entity and the trustee
- *Global trust* or *reputation* – Aggregated community trust, inferred from outcomes of past interactions between third-party actors and the trustor (Skopik et al. 2009)

Other actor-level metrics include *task familiarity* and *team familiarity* (Espinosa et al. 2007). These are especially important for open collaboration where the system cannot assign a task to appropriate and trusted actors. If some of the actors within an open collaboration are already familiar with other actors, the coordination will be positively affected.

Team familiarity is important in large teams where effective team coordination is more difficult. Team familiarity is a function of multiple other metrics such as quality of prior interactions with a coworker previously not belonging to the same team or prior experience with the same team structure and organization. Hence, this measure is closely related to *trust*.

Task familiarity is best explained with an example of open-source software development team. The bigger the number of interdependent modules, the more complex is the task. This increases the amount of information to be processed by human actors, thus it is important that actors have a reasonable amount of task familiarity. Details of a model for performance analysis of teams based on task familiarity and team familiarity can be found in Espinosa et al. (2007).

Group-level metrics describe performance properties of a collaboration. One of the fundamental metrics describing collaborations is the team size. The bigger the number of collaborating actors, the more communication and coordination among them is needed. For example, in Kittur and Kraut (2008) the authors use Wikipedia to analyze how the number of editors and the coordination methods affect the article quality in terms of accuracy, completeness, and clarity.

A metric indicating *interaction intensity* between an actor and other important actors is measured in specific interaction contexts. It is used in the aforementioned *DSARank* ranking algorithm (Schall and Dustdar 2010).

The relevance of the connections to important actors is the most important factor in determining the reputation of an actor. The reliability of the feedback information in reputation systems depends on the reputation of actors providing the feedback. Reputation information is valuable when an actor lacks information based on direct experiences with another actor. However, when this information is available and appropriate, the private or direct trust weights more than trust values based on reputation data. In this case the weight of data from direct interactions should be determined by calculating the minimum number of direct/local trust or rating values that should be maintained by an actor for the actor providing the service/executing a task (Noorian et al. 2012).

Collaboration cost is an important metric because of its direct business influence. This metric takes into account not only the price of task processing paid to the actors, but rather the total costs, including the communication and coordination costs. It is used as the basis for the cost optimization algorithms, as shown in the aforementioned systems – Quirk (Marcus et al. 2011) and AUTOMAN (Barowy and Berger 2012).

Automatically discovering *collaboration patterns* naturally occurring among actors opens up a possibility to identify particularly (un)successful collaboration groups or execution sequences. This information can in turn be used to optimize collaborative process. Identifying collaboration patters is one of the central topics of process mining.

Quality Metrics

Quality-of-Data (QoD) Metrics. As collaboration systems deal with various human-performed tasks, and the data quality primarily depends on the type of tasks, trying to develop a general set of quality metrics makes little sense. For example, metrics listed in Table 1, such as data completeness, freshness, and accuracy, are well-known metrics but their definition is highly dependent on the goal of their use. Instead, different metrics are developed for particular application domains. However, it is exactly the fact that humans participate in the collaborative processes that introduces a concept common to all the application areas – that of *uncertainty* or *inaccuracy* (Parameswaran and Polyzotis 2011). The main sources of uncertainty are caused by the dynamic and unexpected behavior of humans: humans make mistakes, are subjective, and can employ malicious behavior. Thus, approaches for dealing with uncertainty should be included in supporting systems.

Different research communities deal with uncertainty differently. However, all approaches rely on some probability metrics that quantify our belief that a single task is performed correctly. In principle, all approaches can be divided into two categories:

- *Optimistic approaches* – Processed tasks are returned along with a confidence (accuracy) estimate. The data user accepts the results, but must be aware that a certain percentage of the results will be wrong.
- *Pessimistic approaches* – The system applies various mechanisms for error detection and correction and usually resubmits the task to multiple actors until the merged result satisfies the required quality threshold.

Actor performance quality metrics are similar to the "traditional" Web Service metrics, like average execution time, number of invocations, and availability. On the group and collaboration level, these metrics measure and predict the existence of various invocation patterns, i.e., the probabilities that certain services will be called in a particular order with respect to other services. A detailed discussion on interaction metrics can be found in Truong and Dustdar (2009).

Incentives and rewarding are important and effective mechanisms for indirectly influencing quality and motivation of human actors in collaborations. The principal metrics in use in today's computer-supported collaboration systems are:

- *Effort* – It measures an actor's determination to perform a task. The main purpose of this metric is to provide a way to compare the performance of both experienced and inexperienced actors. For example, an inexperienced actor may put in a lot of his time and resources only to perform a task worse or slower than an experienced actor. However, for the purpose of incentivizing, a higher effort level should be compensated with a higher reward, because it will ultimately lead to better experienced actors.
- *Productivity* – It expresses the number of units processed in a time period. This metric is suitable for piecework and easily quantifiable tasks (e.g., bug reporting, image tagging, text translation).
- *Quality of work* – This metrics expresses the quality of the working process of an actor. It should not be confused with the Quality-of-Data (QoD) of processed tasks. This metric is used to assess actors when the task's QoD cannot be easily determined or when it cannot say much about the actor. For example, actors that help other actors, waste less resources, provide creative ideas, or take responsibility should be also rewarded. In such cases, the subjective opinions of other relevant actors (i.e., *peers*) can be used to quantify these elusive actor qualities.

In order to acquire the rewarding metrics, collaborative systems use different *evaluation methods*, relying both on human and machine actors:

- *Individual evaluation methods*
  - *Quantitative methods* – They represent a quantitative measurement of an individual actor's contribution as measured by the system itself. Such metrics can represent the number of processed tasks, average speed, responsiveness, acceptance rate, etc. These methods are considered fair and cheap to implement, but unfortunately they are applicable only in cases where actors work on easily quantifiable tasks.
  - *Subjective methods* – In cases where the quality of work is a property understandable to humans only, a quantitatively expressed subjective assessment by a human actor replaces a quantitative metric measured by the system itself. This is the case with artistic or designer tasks. The advantages are the simplicity and cost, but a serious drawback is the inevitable lack of objectivity.
- *Group Evaluation Methods*
  - *Peer evaluation methods* – They are used to express an aggregated opinion of an interest group. The members of evaluation group usually express their votes by scoring tasks or actors on a fixed scale or by investing amounts of virtual credits expressing their confidence (placing bets). The quality and effectiveness of these methods are influenced by the size of the composition of the evaluation group.
  - *Indirect evaluation methods* – In certain situations human actors can be evaluated by comparing the status of the artifacts they previously produced with the status of the artifacts produced by other members of the same community. The artifacts can be Web pages, projects, articles, photos, and programming code. These comparisons are usually performed with the help of sophisticated algorithms. Examples are the Google's PageRank algorithm, impact factor for scientific publications, or Klout's algorithm for measuring social network influence. Advantages and disadvantages of these methods are dependent on the properties of the algorithm.

Future Directions

Although a considerable amount of work is done in the area of interaction analysis in social networks, there is much less work conducted on team-based metrics and analysis. Many open questions still remain to be tackled. Some of them are (i) understanding the interdependencies between metrics for better analysis of different collaboration systems, testing and evaluating these team-based metrics, and (ii) utilizing these metrics in the most appropriate way for task adaptation. Another future research direction in team collaboration in mixed systems is to develop metrics that can be used to compare human-and software-based actors.

Cross-References

▸ Computational Trust Models
▸ Creating a Space for Collective Problem-Solving
▸ Crowdsourcing and Human Computation, Introduction
▸ Crowdsourcing and Social Networks
▸ Distance and Similarity Measures
▸ Incentives in Collaborative Applications
▸ Similarity Metrics on Social Networks
▸ Social Computing
▸ Virtual Team
▸ Wikipedia Collaborative Networks

References

Amend M, Ford M, Endpoints A, Keller C, Rowley M (2007) WS-HumanTask specification, v1.0. http://public.dhe.ibm.com/software/dw/specs/ws-bpel4people/WS-HumanTask_v1.pdf

Anagnostopoulos A, Becchetti L, Castillo C, Gionis A, Leonardi S (2012) Online team formation in social networks. In: Proceedings of the 21st international conference on world wide web – WWW '12. ACM, New York, p 839. doi:10.1145/2187836.2187950, http://dl.acm.org/citation.cfm?id=2187836.2187950

Barowy D, Berger E (2012) AUTOMAN: a platform for integrating human-based and digital computation.

http://www.cs.umass.edu/~emery/pubs/AutoMan-UMass-CS-TR2011-44.pdf

Caverlee J, Liu L, Webb S (2008) Socialtrust: tamper-resilient trust establishment in online communities. In: Proceedings of the 8th ACMIEEECS joint conference on digital libraries. ACM, pp 104–113. http://portal.acm.org/citation.cfm?id=1378889.1378908

Caverlee J, Cheng Z, Eoff B, Hsu CF, Kamath K, Kashoob S, Kelley J, Khabiri E, Lee K (2010) SocialTrust++: building community-based trust in social information systems

Dorn C, Dustdar S (2010) Composing near-optimal expert teams: a trade-off between skills and connectivity. On the Move to Meaningful Internet Systems: OTM 2010, pp 472–489. http://www.springerlink.com/index/H434570G12787H57.pdf

Dustdar S, Bhattacharya K (2011) The social compute unit. IEEE Internet Comput 15(3):64–69. doi:http://dx.doi.org/10.1109/MIC.2011.68, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5755601

Dustdar S, Gaedke M (2011) The social routing principle. IEEE Internet Comput 26–29

Espinosa JA, Slaughter SA, Kraut RE, Herbsleb JD (2007) Familiarity, complexity, and team performance in geographically distributed software development. Organ Sci 18(4):613–630. doi:10.1287/orsc.1070.0297, http://dx.doi.org/10.1287/orsc.1070.0297

Gaston ME, DesJardins M (2005) Agent-organized networks for dynamic team formation. In: Proceedings of the fourth international joint conference on autonomous agents and multiagent systems – AAMAS '05, p 230. doi:10.1145/1082473.1082508, http://portal.acm.org/citation.cfm?doid=1082473.1082508

Grudin J (1994) Computer-supported cooperative work: history and focus. Computer 27(5):19–26. doi:10.1109/2.291294, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=291294

Kittur A, Kraut RE (2008) Harnessing the wisdom of crowds in wikipedia. In: Proceedings of the ACM 2008 conference on computer supported cooperative work – CSCW '08, San Diego. ACM, New York, pp 37–46. doi:10.1145/1460563.1460572, http://portal.acm.org/citation.cfm?id=1460563.1460572 http://portal.acm.org/citation.cfm?doid=1460563.1460572

Kloppmann M, Koenig D, Leymann F, Pfau G, Rickayzen A, Schmidt P, Trickovic I (2005) WS-BPEL extension for people (July):1–18. public.dhe.ibm.com/software/dw/specs/ws-bpel4people/BPEL4People_white_paper.pdf

Lappas T, Liu K, Terzi E (2009) Finding a team of experts in social networks. In: Proceedings of the 15th ACM international conference on knowledge discovery and data mining 7120(4), p 467. http://portal.acm.org/citation.cfm?doid=1557019.1557074

Law E (2011) Defining (human) computation. In: Workshop on crowdsourcing and human computation

Little G, Chilton LB, Miller R, Goldman M (2009) TurKit: tools for iterative tasks on mechanical turk. IEEE. doi:10.1109/VLHCC.2009.5295247, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5295247

Little G, Chilton LB, Goldman M, Miller R (2010) Exploring iterative and parallel human computation processes. In: Proceedings of the 28th of the international conference extended abstracts on human factors in computing systems – CHI EA '10, p 4309. doi:10.1145/1753846.1754145, http://portal.acm.org/citation.cfm?doid=1753846.1754145

Lopez M, Vukovic M, Laredo J (2010) PeopleCloud service for enterprise crowdsourcing. In: 2010 IEEE international conference on services computing, pp 538–545. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5557275

Marcus A, Wu E, Karger DR, Madden S, Miller R-C (2011) Platform considerations in human computation. In: Workshop on crowdsourcing and human computation

Marsh SP (1994) Formalizing trust as a computational concept. University of Stirling, Stirling

Newman MEJ (2010) Networks: an introduction

Noorian Z, Fleming M, Marsh S (2012) Preference-oriented QoS-based service discovery with dynamic trust and reputation management. In: Proceedings of the 27th annual ACM symposium on applied computing – SAC '12, p 2014. doi:10.1145/2245276.2232111, http://dl.acm.org/citation.cfm?doid=2245276.2232111

Parameswaran A, Polyzotis N (2011) Answering queries using humans, algorithms and databases. Technical Report, Stanford University. http://ilpubs.stanford.edu:8090/986/

Quinn AJ, Bederson BB (2011) Human computation: a survey and taxonomy of a growing field. In: Proceedings of the 2011 annual conference on human factors in computing systems – CHI '11. ACM, New York, p 1403. http://dl.acm.org/citation.cfm?id=1978942.1979148

Schall D, Dustdar S (2010) Dynamic context-sensitive PageRank for expertise mining. In: Proceedings of the second international conference on social informatics. Springer, Berlin/Heidelberg, pp 160–175. http://dl.acm.org/citation.cfm?id=1929326.1929338

Skopik F, Schall D, Dustdar S (2009) The cycle of trust in mixed service-oriented systems. In: 2009 35th Euromicro conference on software engineering and advanced applications, pp 72–79. doi:10.1109/SEAA.2009.20, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5349860

Skopik F, Schall D, Dustdar S (2010) Modeling and mining of dynamic trust in complex service-oriented systems. Inf Syst 35(7):735–757. doi:10.1016/j.is.2010.03.001, http://linkinghub.elsevier.com/retrieve/pii/S0306437910000153

Truong Hl, Dustdar S (2009) Online interaction analysis framework for ad-hoc collaborative processes in SOA-based environments. Framework pp 260–277. doi:10.1007/978-3-642-00899-3_15, http://www.springerlink.com/index/u1075446t8qr727q.pdf

van der Aalst WMP (2011) Process mining. Springer, Berlin/Heidelberg. doi:10.1007/978-3-642-19345-3, http://www.mendeley.com/research/no-title-avail/http://www.sciencedirect.com/science/article/pii/S0166361503001945http: //www.springerlink.com/index/10.1007/978-3-642-19345-3

Zhang P, Serban N (2007) Discovery, visualization and performance analysis of enterprise workflow. Comput Stat Data Anal 51(5):2670–2687. doi:10.1016/j.csda.2006.01.008, http://linkinghub.elsevier.com/retrieve/pii/S0167947306000132

Social Knowledge Network

▸ Automatic Document Topic Identification Using Social Knowledge Network

Social Media

▸ Mapping Online Social Media Networks
▸ Mining Trends in the Blogosphere
▸ NodeXL: Simple Network Analysis for Social Media
▸ Privacy Issues for SNS and Mobile SNS
▸ Reconnaissance and Social Engineering Risks as Effects of Social Networking
▸ Social History of Computing and Online Social Communities
▸ Social Media Policy in the Workplace: User Awareness
▸ Social Networking in Political Campaigns
▸ Topology of Online Social Networks
▸ Web Communities Versus Physical Communities

Social Media Analysis

▸ Analysis and Mining of Tags, (Micro)Blogs, and Virtual Communities
▸ Collective Intelligence, Overview
▸ Sentiment Analysis in Social Media
▸ Twitris: A System for Collective Social Intelligence

Social Media Policy in the Workplace: User Awareness

Mohd Heikal Husin[1] and Jo Hanisch[2]
[1]Service Computing, School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia
[2]Strategic Information Management Group, School of Computer and Information Science, University of South Australia, Adelaide, SA, Australia

Synonyms

Impacts of policy; Legal issues; Risks; Social media; Training

Glossary

Governance The act of governing and relates to decisions that define expectations or verify performance

Bureaucratic Having the characteristics of a bureaucracy or a bureaucrat.

CRM Customer relationship management

Awareness Knowledge or perception of a situation or fact

Policy There are two main types of policies – public policies and private policies. In this paper, the research focuses on the private policies or organizational policies which are limited in available resources as well as legal coercion

Longitudinal Research A research study that involves repeated observations/interviews over a period of time

Definition

Organizational policy and social media are two of the most highly discussed topics within organizations today, especially within governments. Policy is typically described as a principle or process to guide decisions in order to achieve rational outcomes or to address evident problems (von

Solms and von Solms 2004). The difference between a policy and a procedure is that a policy will contain the "what" and "why," while a procedure contains the "what," "how," "where," and "when" (Colebatch 2006, pp. 313; 317). Policies are generally adopted by a board or senior management body within an organization (Wergin 1976). They guide senior management in making both subjective-based decisions (on the relative merits of factors which are difficult to objectively test, such as work-life balance) and objective-based decisions (operational in nature and easier to objectively test, such as security policy) (Wergin 1976).

Social media are web-based applications that have emerged from outside the organization. They provide an interactive and open approach to collaboration, communication mainly through the use of Web 2.0. Mergel and Schweik (2012) highlight that Web 2.0 derives its power from users for all activities, and this indirectly differentiates Web 2.0 from other standard technologies implemented within organizations, such as CRM or any other management systems. This poses several problems, including the following: (1) Controlling the level of openness that is needed within government organizations. This is due to the lack of an effective regulatory framework or policy within organizations associated with social media. (2) The existence of information leakage as employees might accidentally share confidential information about their work through social media. This information could either damage or threaten the reputation of the department. (3) Lengthy bureaucratic approval processes, especially when it comes to allowing employees to either access new information or when providing information to other users, such as citizens and nongovernmental organizations. While this process provides a security barrier for the organization, it also deters the interaction among employees who are interested in the new information or providing information and feedback to other users. Placing too many restrictions contradicts the rationale for using social media.

There have been some developments in lowering risks of employing social media tools within organizations. Husin and Hanisch (2011a) propose a policy development framework, highlighting the important components within an effective social media policy. Osimo (2008) presents lessons learned, such as enabling authentication policies and partnering with certain Web 2.0 applications instead of centrally implementing all applications; and Tapscott et al. (2007, p. 18) developed steps to manage change for new governance designs that lead to innovative and agile processes within governments through social media.

While the available research provides good arguments in terms of the importance of planning implementation for social media as well as the development of an effective social media policy, the relationship between users' awareness of a social media policy and the adoption rates of social media remains largely unexplored. So the research question that has been developed for this paper is: "How does the awareness of a social media policy influence the use of social media among users in an organization?"

This entry considers why user adoption is important and why a policy is essential for organizations to maintain control over a new technology. The overview of steps that were undertaken for the research is provided, leading to the results of the longitudinal research. The conclusion includes a summary of the research outcomes and the relevance of the research to an organization which intends to develop an effective policy.

Introduction

The successes of a technology implementation within organization are dependent on the users. Rogers (2003, pp. 171, 177) highlights that users could either adopt or reject a technology and these decisions are based on either a need or an awareness of a technology. Adoption rates are expected to be lower within a working environment where key decisions are made through different parties within a department (Ba-

jwa et al. 2005; Onyechi and Abeysinghe 2009). The complexity of the decision is related to different processes and policies depending upon the organization (Shumarova and Swatman 2008).

There are also a number of reasons why adoption among users could be affected such as low trust levels (Johnston 2007), corporate culture, and the requirement of more training for new technology which is something that users would mainly try to avoid (Husin and Swatman 2010). So in order to limit the barriers to adoption, an effective policy is crucial.

When an organization implements a new technology, the need for an effective organizational policy is essential in order to provide a sense of security for the organization (Husin and Hanisch 2011a). But more often than not, such policies tend to be looked over by users (Althaus et al. 2008). This may be attributed to the generally extensive comprehensiveness of policies (especially in the public sector) and the associated perception of lack of relevance for the user (Althaus et al. 2008). So an appropriate policy development process is essential in ensuring that users have an understanding of the organization's rules and in ensuring that the policy contains important components (Hrdinova et al. 2010; Husin and Hanisch 2011b; Woodford 2005). This also allows the policy to be developed more effectively, while maintaining relevance to the users from the perspective of their work.

For an authoritative-based organization such as in public sector, policies are usually viewed in three ways (Althaus et al. 2008, p. 6):

(1) As an authoritative choice
    Clearly viewed as the method for government to exercise their power and guarantee results through a series of hierarchical decisions.
(2) As a hypothesis
    All policies go through an iterative process or "error making" which enhances and changes the policy to be more effective.
(3) As an objective of government action
    Policies act as a guide for a department to achieve the intended results (Moule et al. 1995).

So, it is natural for employees within the public sector to view the authoritative choice as essential while using a social media tool, but this should not be the case as mentioned by Husin and Hanisch (2011a) and Hrdinova et al. (2010) due to the flexibility and openness that the tool promotes for an organization.

This research considers the social media policy from an organizational perspective (refer to Glossary) and aims to identify the influences that a user's awareness of the policy has on their social media usage.

How the Study Was Conducted

Due to the nature of ethical requirements, the government agency that participated in the research remains anonymous. For the purpose of this paper, it will be referred to as Agency A. Agency A was in the early stages of implementing a standardized social media platform which would be accessible to all their employees for their daily activities. The researcher conducted semiformal interviews as well as quick questionnaires during the platform training session for employees.

As the research was a longitudinal approach over the period of 2010 and 2011, the interview sessions were conducted with ten users from within Agency A with two interviewees continually participating due to their role with the implementation process for the social media platform. The interview included questions concerning their opinions about social media, examples of usage of social media, and their awareness of the social media policy. It should be noted that Agency A consists of a number of internal departments with many of the interviewees spread across different locations.

During the initial implementation process, Agency A conducted training for their employees, involving an hour and a half of "hands-on" time with the social media platform. The participation from the employees was encouraging enough for the training sessions to be held every month since October 2010. The quick questionnaires were circulated at the end of 5 ran-

domly selected training sessions which brought the total of 81 respondents. The questionnaires asked participants about their level of social media usage within their own departments, tools which they deem useful for the respondents, and their expectations from using social media in their work.

Results of the Research

The results are based on the analysis conducted through the interviews as well as the quick questionnaires. The levels of awareness for the existence of a social media policy among the respondents from the training sessions are high as evident from Table 1.

Table 1 shows the overall level of awareness among the questionnaire respondents at 39.51 %, with female respondents at 33.33 % and the male respondents at 6.17 %. A main reason for the stark contrast of gender number is due to a high number of female employees (62 % out of the total employees) within Agency A. Nevertheless, the results still indicate that there is awareness for the social media policy among employees along with a number of employees unaware of the policy (28.39 %). The results from the respondents about their positions and whether they found the social media platform useful in their work are shown in Table 2.

The majority of the respondents were on the employees level (60.49 %), the middle management (23.45 %), and senior management (7.41 %), while the remainder was reluctant to disclose their position. This shows that the interests in using the social media platform in daily work activities are still evident (39.51 %) even with the existence of the social media policy. This is a good sign as it shows that higher management supports the use of the social media platform. As some of the feedback from the interview sessions states,

> More higher management should use the platform so it gives a sense to employees, that yes, the platform is an official tool for them to use. – Interviewees 6 and 7

As Agency A consisted of different departments, the level of social media usage is quite varied. The research found that most of the departments that are using social media are aware of the social media policy. Table 3 shows the different levels of usage and awareness.

The department which was "going ahead" with utilizing social media in their daily activities had a higher awareness of the policy (13.58 %) which is followed by the department which is "trying out" the tools (12.35 %). But coincidentally, the latter department also had the highest level of unawareness for the social media policy (8.64 %). The interesting result was that departments who were "fully using" had the lowest number of social media policy awareness (1.23 %) compared to the other levels of usage.

In the interview sessions, majority of the participants were aware of the social media policy and have either read or had a quick review of the policy. More than 70 % of the participants were still using the social media platform frequently without any indifference to the policy. The participants were using the platform to communicate ideas, comment on non-work-related information, and even share common interests with their colleagues.

An example of the social media usage was by Interviewee 6 where an employee was looking for an available meeting room through the social media platform. As Agency A is spread across different locations with different meeting room sizes; traditionally, the employee would need to either email or contact via telephone the appropriate parties to find an available meeting room. But instead, the employee accessed the social media platform and sent a mass broadcast for assistance via the available micro blogging tool. Within 30 min, the employee had a reply from another employee located in a different location who had booked a meeting room for the initial employee.

Even with the clear benefits of the social media platform in Agency A, there are a few employees who are using the platform mainly because the tool's usage is mandatory by their department. Even though the employees were quite happy with using the social media platform

Social Media Policy in the Workplace: User Awareness, Table 1 Level of awareness for social media policy

		Awareness of social media policy				
		N/A	No	Yes	Total	Percentage
Gender	F	22	10	27	59	72.84
	M	4	13	5	22	27.16
Total		26	23	32	81	100

Social Media Policy in the Workplace: User Awareness, Table 2 Position vs. social media usefulness

		Awareness of social media policy				
		N/A	No	Yes	Total	Total percentage
Position of user	Employee	12	18	19	49	60.49
	Middle M	7	14	8	19	23.45
	N/A	4	1	0	5	6.18
	Other	0	0	2	2	2.47
	Senior M	3	0	3	6	7.41
Total		26	23	32	81	100

Social Media Policy in the Workplace: User Awareness, Table 3 Level of usage within departments vs. awareness

		Awareness of social media policy				
		N/A	No	Yes	Total	Total percentage
Level of social media usage	Fully using	8	2	1	11	13.58
	Going ahead	3	3	11	17	20.99
	I am not sure	4	6	7	17	20.99
	N/A	1	3	0	4	4.94
	Planning	1	1	3	6	7.4
	Trying out	9	7	10	26	32.1
Total		26	23	32	81	100

for any work-related activity, there is not much interest among the employees in any social activity that comes with the social media platform within working hours. From an analysis of the interviews, some participants recalled that

> They were paid to work and not to socialize during working hours. (Interviewees 9 and 10).
> It doesn't matter to me if someone wants to share their interest but I don't see the point of doing so in working hours. (Interviewee 9)

The interesting point is that the participants who mentioned the quotes above are highly interested in the social media policy available in Agency A. In a way, they view the policy as a useful guide for how they should interact on the social media platform.

Conclusions

The results show that balance is needed in order to cater for different users of the social media platform. On one side, there is the socially-based user (highly interactive and willing to share information), while the other is the work restrictive-based user (critical only on work-related issues, with no interest in the social side). Both user categories have awareness of the policy but vary in their usage of social media.

The results indicate that influencing factors on the uptake of social media include the level of training, the ability to use social media for work-related activities, and the level of use by senior management, as an example and reassurance to all employees.

Awareness of policies appears varied across the departments in Agency A which is predicted to occur in a large organization. But the variedness of awareness, especially in the departments which were designated as "fully using," was quite surprising as it was expected that awareness would generally be high. Hence, the departments which have made the decision to "go ahead" with social media have employees with higher awareness of the policy than those departments which are "fully using" social media. This is where the effectiveness of the organizational social media policy is needed as well as the dissemination of the policy and its repercussions in practice. As Bridgman and Davis (2003) suggest, there needs to be a bridge between technical expertise and policy domain. Organizational policy, which focuses on social media, needs to be developed with due diligence as employees are dependent on the policy to guide them.

Acknowledgment

The authors would like to thank Agency A for providing the research with the opportunity to interview and conduct short questionnaires with their employees.

Cross-References

- ▸ Legal Implications of Social Networks
- ▸ Social Media

References

Althaus C, Bridgman P, Davis G (2008) The Australian policy handbook, 4th Illustrated edn. Allen & Unwin, New South Wales, Australia

Bajwa DS, Lewis LF, Pervan G, Lai VS (2005) The adoption and use of collaboration information technologies: international comparisons. J Inf Technol 20(5): 130–140

Bridgman P, Davis G (2003) What use is a policy cycle? Plenty, if the aim is clear. Aust J Public Adm 62(3): 98–102

Colebatch HK (2006) What work makes policy? Policy Sci 39(4):309–321

Hrdinova J, Helbig N, Peters CS (2010) Designing social media policy for government: eight essential elements. Center for Technology in Government, University of Albany, Albany, New York, USA

Husin MH, Hanisch J (2011a) Social media and organisation policy (SOMEOP): finding the perfect balance. In: The 19th European conference on information systems – ICT and sustainable service development, Helsinki, 9–11 June 2011

Husin MH, Hanisch J (2011b) Utilising the social media and organisation policy (SOMEOP) framework: an example of organisational policy development within a public sector entity. In: The 19th European conference on information systems – ICT and sustainable service development, Helsinki, 9–11 June 2011

Husin MH, Swatman PMC (2010) Removing the barriers to Enterprise 2.0. In: 2010 IEEE international symposium on technology and society, UoW, Wollongong, pp 275–283

Johnston K (2007) collaborative filtering and e-Business: is Enterprise 2.0 one step forward and two steps back? Electron J Knowl Manag 5(4): 411–418

Mergel I, Schweik CM (2012) The paradox of the interactive web in the U.S. Public Sector. Public service, governance and web 2.0 technologies: future trends in social media. IGI Global, Hershey, Pennsylvania, USA

Moule B, Giavara L (1995) Policies, procedures and standards: an approach to implementation. Inf Manag Comput Secur 3(3):7–16

Onyechi GC, Abeysinghe G (2009) Adoption of web based collaboration tools in the enterprise: challenges and opportunities. In: 2009 international conference on the current trends in information technology (CTIT), Dubai, pp 1–6, 15–16 Dec 2009

Osimo D (2008) Web 2.0 in Government: why and how? Institute for Prospectice Technological Studies (IPTS), JRC, European Commission, EUR, vol 23358. Seville, Spain, p 57

Rogers E (2003) Diffusion of innovation, Paperback edn. Free Press, a division of Simon and Schuster, New York

Shumarova E, Swatman PA (2008) Informal ecollaboration channels: shedding light on 'shadow CIT'. In: 21st bled eConference: overcoming boundaries through multi-channel interaction, Bled, pp 371–394, 15–18 June 2008

Tapscott D, Williams AD, Herman D (2007) Government 2.0: transforming government and governance for the twenty-first century. New Paradigm White Paper

von Solms R, von Solms B (2004) From policies to culture. Comput Secur 23(4):275–279

Wergin JF (1976) The evaluation of organizational policy making: a political model. Rev Educ Res 46(1): 75–115

Woodford MD (2005) Central bank communication and policy effectiveness. SSRN eLibrary, Wyoming, USA

Social Media, Definition and History

Andreas M. Kaplan
Department of Marketing, ESCP Europe, Paris, France

Glossary

Ambient Awareness Awareness created through regular and constant reception and/or exchange of information fragments through social media

MMORPG Massively Multiplayer Online Role-Playing Game

Mobile Social Media Group of mobile marketing applications that allow the creation and exchange of user-generated content

UCG User-generated content

Definition

Social media are defined as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content" (Kaplan and Haenlein 2010, p. 61). Without any doubt, social media and UCG have become a reality for millions of individuals and corporations. If it were a country, the social networking site *Facebook* would be the third most populated one in the world, with its over 900 million users. The collaborative project *Wikipedia* with an impressive 22 million articles has over 370 million readers worldwide and is undeniable a key information provider in the online space. *YouTube*, probably the most known content community, is the second biggest search engine after the industry giant Google. The microblogging service *Twitter* (Kaplan and Haenlein 2011) with its half-a-billion active users generates 340 million tweets per day. Finally also virtual game and social worlds (Kaplan and Haenlein 2009) belong to the group of social media (cf. Fig. 1).

Many would probably identify the advent of Facebook, Twitter, and YouTube as the beginning of social media. But, contrary to this belief, the creation and exchange of user-generated content existed long before. The aim of this short essay is to provide a brief sketch of the key developments in social media history, its roots, and its future evolutions.

First Era – 1980s: Arrival of Social Media

The arrival of social media applications coincides with the Internet's first use by private individuals. In fact, a big part of the Internet started as nothing more than so-called newsgroups where individuals could view, discuss, and post bulletin board-like messages to numerous categories. Often these newsgroups were focused on technical issues but they also covered cultural topics such as science fiction or similar. *Usenet*, established in 1980 by Tom Truscott and Jim Ellis from Duke University, was the most popular discussion system at that time and can be seen as the direct forerunner of the category "Internet forum" which is similar to collaborative projects. These bulletin board systems quickly developed into real discussion groups by allowing individuals to create and exchange user-generated content with each other. Also the first virtual game worlds came up during this era of social media: in 1980, *Multi-User Dungeon*, the first so-called massively multiplayer online role-playing game (MMORPG) and precursor of virtual game worlds such as the World of Warcraft, was introduced by Roy Trubshaw and Richard Bartle from Essex University.

Second Era – 1990s: Fading of Social Media

During the second era of social media, user-generated content heavily lost in importance due to the fact that more and more companies started to make use of the Internet for their purposes. With industry giants such as Amazon or eBay arriving in 1995 and conquering the web with their corporate websites, the social media applications from the first era seemingly faded away. Despite the fact that social media went by unnoticed by the general public, more and more people started to have their own blogs during the second era and used them to publicly account of their personal lives. While the term "weblog" was introduced

		Social presence/ Media richness		
		Low	Medium	High
Self-presentation/ Self-disclosure	High	Blogs and Microblogs (e.g., Twitter)	Social networking sites (e.g. Facebook)	Virtual social worlds (e.g. Second Life)
	Low	Collaborative projects (e.g. Wikipedia)	Content communities (e.g. YouTube)	Virtual game worlds (e.g. World of Warcraft)

Social Media, Definition and History, Fig. 1 Classification of social media (for more details, see Kaplan and Haenlein 2010, p. 62)

by Jorn Barger not before the end of 1997, blogs existed already in the beginning of the 1990s. Its short form "blog," by the way, was coined by Peter Merholz, who jokingly broke the word weblog into the phrase we blog on his own blog in 1999.

Third Era – 2000s: Rising of Social Media

With the dot-com bubble bursting in 2001, social media came back into the game and started to recapture the virtual sphere. *Wikipedia* started on January 15, 2001, with the simple sentence: "Hello world. Humor me. Go there and add a little article. It will take all of five or 10 minutes." On February 4, 2004, Marc Zuckerberg launched *Facebook*, originally located at thefacebook.com, changing it to the current web address not before 2005. Founded on February 14, 2005, *YouTube*'s first video entitled "Me at the zoo" showed co-founder Jawed Karim at the San Diego Zoo and was uploaded on April 23 of the same year. And *Twitter*, launched on July 15, 2006, started out with its first tweet 4 months earlier on March 21 sent by cofounder Jack Dorsey typing "Just setting up my Twtr." All of these four social media applications lived an enormous success story and today belong to the top 10 websites worldwide.

Fourth Era – 2010s: Mobilizing of Social Media

The fourth era of social media is characterized by the arrival of so-called mobile social media (Kaplan 2012) such as Foursquare, i.e., social media accessed via a mobile device. These new mobile forms turn computer-based social media, despite their young age, already into traditional social media. Geolocalization and increased time sensitivity are two of the features offered by mobile devices. Both provide mobile social media applications with increased opportunities compared to computer-based ones. For example, with mobile social media one is aware not only of one's friends' plans but also of their current location and might just go and see them. Ambient awareness, defined as "awareness created through regular and constant reception, and/or exchange of information fragments through social media" (Kaplan 2012, p. 132), is an equally important concept within the area of mobile social media. Since this era just started, it is difficult to say more about its potential evolution for the moment. However, some futuristic, but not impossible, scenarios already arise on the horizon, e.g., facial recognition could make it feasible to take somebody's picture with a cell phone and compare it to social networking sites. A match could give the name and other details about this individual.

This brief sketch of the key developments in social media history showed that these applications started earlier than one would have thought, i.e., in the 1980s. Applications such as *Facebook* or *YouTube* can actually be seen as the Internet going "back to the roots" when the power was with the individual users instead of with big companies. Social media retransformed the Internet to what it was initially intended for – a platform to create and exchange user-generated content.

Cross-References

- ▸ Facebook's Challenge to the Collection Limitation Principle
- ▸ Flickr and Twitter Data Analysis
- ▸ Gaming and Virtual Worlds
- ▸ Location-Based Social Networks
- ▸ Virtual Goods in Social Media
- ▸ Wikipedia Collaborative Networks

References

Kaplan AM (2012) If you love something, let it go mobile: mobile marketing and mobile social media 4 × 4. Bus Horiz 55(2):129–139

Kaplan AM, Haenlein M (2009) The fairyland of second life: about virtual social worlds and how to use them. Bus Horiz 52(6):563–572

Kaplan AM, Haenlein M (2010) Users of the world, unite! The challenges and opportunities of social media. Bus Horiz 53(1):59–68

Kaplan AM, Haenlein M (2011) The early bird catches the news: nine things you should know about microblogging. Bus Horiz 54(2):105–113

Social Navigation

- ▸ Social Web Search

Social Network

- ▸ Human Behavior and Social Networks
- ▸ NodeXL: Simple Network Analysis for Social Media
- ▸ Social Engineering/Phishing
- ▸ Social Networking in Political Campaigns
- ▸ User Behavior in Online Social Networks, Influencing Factors

Social Network Analysis

- ▸ GUESS
- ▸ NetMiner
- ▸ Pajek
- ▸ Social Recommendation in Dynamic Networks
- ▸ Stability and Evolution of Scientific Networks
- ▸ Topology of Online Social Networks
- ▸ UCINET

Social Network Analysis and Company Linguistic Identity

Magdalena Bielenia-Grajewska
Cognitive Neuroscience Sector, Scuola Internazionale Superiore di Studi Avanzati in Trieste, Trieste, Italy
Faculty of Languages, Department of Translation Studies and Intercultural Communication, Institute of English, University of Gdansk, Gdansk, Poland

Synonyms

Corporate linguistic image; Corporate linguistic personae; Lattices and ties; Social grids

Glossary

Corporate Linguistic Identity A corporate identity created by means of linguistic tools and visible mainly in the selection of linguistic repertoire

Definition

The term company identity is used to encompass the areas covered by such concepts as corporate

identity and organizational identity. Analyzing target audience, corporate identity is connected with the way stakeholders perceive the company, whereas organizational identity encompasses employees' opinions about their place of work (Cornelissen 2008). Taking the process of identity creation into account, corporate identity mirrors the managerial contribution to effective communication, whereas organizational identity takes place during informal encounters between workers (Rughase 2006). Thus, company linguistic identity encompasses various linguistic tools used by managers, employees, and stakeholders, in both formal and informal settings. It covers the linguistic representation of corporate activities, visible in written and spoken forms, at both internal and external levels. Taking into account the importance of corporate networking in modern organizations, company linguistic identity can be presented through the prism of social networks, thus it is understood as the linguistic entity being formed by social networks, responding to social networks in the environment and forming social networks. These social networks are viewed through the linguistic and communicative functions they serve as well as by taking into account the linguistic tools they are created by.

Introduction

There are different reasons determining the application of SNA theories into the study on identities. One of them is related to the postmodern approach stressing changeability, interconnectedness, and relations. Thus, SNA focuses on such processual issues as knowledge flows, communication, and innovation. The other factor is the role of technology that stimulates the performance of social networks (Travica 1999). Technology becomes not only an important part of one's environment but also forms networks within the individual. It can be understood both in the direct sense, taking into account some technological devices inserted within a human body (e.g., artificial body parts) and, in a more general or even metaphorical way, by examining how technology shapes various networks within an individual (e.g., the networks related to one's biological functions, emotions, feelings, and social performance) (Michael and Michael 2008). Technology also creates and maintains linguistic networks at the personal and group level, determining communication styles, available communication channels and devices, as well as the selection of linguistic repertoire by individuals.

Key Points

The SNA approach can be used to study corporate linguistic identity from individual and social perspectives, examining the networks shaping the linguistic performance of a person and the language representation of an organization as such. Thus, this approach makes it possible to observe different levels of linguistic networks within an entity. At the individual level, linguistic networks can be viewed, e.g., through the prism of ethnic, national, and professional identities shaping the linguistic performance of a person. Examining the organizational domain, corporate linguistic identity can be studied by observing micro-(worker), meso-(company), and macro-(country/Europe) linguistic levels. Moreover, the mentioned domains can be researched individually, by studying the networks creating them as well as the network relation and dynamics among the constituting entities and areas (Bielenia-Grajewska 2010).

Historical Background

The first traces of research on social networks and social collectives go back to the nineteenth century, to the works of Auguste Comte, Ferdinand Tönnies, Norbert Elias, Emile Durkheim, and Gustave Le Bon. The role of networks in the life of individuals is also discussed in the works of George Simmel. In the 1990s, there was an expansion of interest in social network analysis, with the extensive publication of edited volumes, handbooks, and software packages and the active performance of network professional

associations (Gamper and Reschke 2010). It was also the time of linguistic turn in social network analysis, taking into account networks, discourse, and identity (Mische 2011). The interest in the discursive approach to organizations is visible in, e.g., the works of postmodernists who stress the role of changeability and fluidity in organizational communication, also from the linguistic perspective.

Proposed Solution and Methodology

Measuring corporate linguistic identity through the prism of social networks entails taking into account different notions, such as the characteristics of ties as well as the duration and frequency of contacts (Ferligoj and Hlebec 1999). Thus, such issues as the type of corporate relations (including vertical and horizontal contacts), the length of communicative acts, as well as their frequency determine the creation of organizational linguistic performance. For example, corporate relations and their impact on company linguistic identity can be studied by taking into account the tie approach that enhances the study on strong and weak ties shaping corporate communication and flows of information exchange among workers. The other view is connected with taking the boundary issue into account. One of the ways to study corporate frontiers is to apply the realist standpoint to show how actors view themselves. Another method is to rely on some formal divisions within the organizational setting to select and divide people into networks according to their professions or their positions in a company (Fombrun 1982). The methodology selected for analyzing corporate linguistic identity depends on the taken perspective. For example, individual linguistic identity can be examined by looking at both individual and societal factors. Attitude to the organizational culture, work-related factors, and language-related benefits belong to the personal dimension. Thus, the worker engages in a linguistic network if his or her participation is beneficial to him or her. Taking the social sphere into account, such issues as corporate language policy, corporate communication and corporate hierarchy determine network formation and network selection. The other network perspective that can be applied in the discussion on company online identity is the observation of how a company and its corporate lingo create individual performance and, simultaneously, how an individual worker and his or her way of communicating shape the organizational discourse. This study can be enriched by taking the environment into account, and consequently, studying micro-(individual), meso-(company), and macro-(environment) linguistic network levels separately as well as the interrelations among them (Bielenia-Grajewska 2010, 2013). The next possibility is to view corporate online identity through the prism of homogeneous and heterogeneous networks. Homogeneous linguistic networks mirror the relationships between people having similar language background, using similar professional genres or opting for common linguistic expressions that are gender, generation, or profession specific. Individuals decide to take part in these networks owing to relatively few problems with understanding the interlocutor. On the other hand, heterogeneous networks comprise individuals of diversified linguistic background and communicative methods (Bielenia-Grajewska 2012). Analyzing the methodology of social network analysis and sampling, various chain methods can be applied. For example, snowball sampling can be used since it relies on individuals eliciting themselves and other linguistic network members. This method turns out to be useful in the study on external organizational relations, to show how language determines contacts among various stakeholders.

Key Applications

The social network method can be used to study corporate relations from the linguistic perspective, by examining diversified networks underlying corporate communication and how these linguistic networks shape corporate linguistic identity.

Future Directions

It can be predicted that in the future, the network approach will be even more visible in the studies on the linguistic performance of modern organizations. The reason for this situation is the growing interconnectedness of companies and their dependence on other organizations as cooperatives, suppliers, customers, and also competitors.

Cross-References

▸ Cognitive Strategic Groups
▸ Collection and Analysis of Relational Data in Organizational and Market Settings
▸ Entrepreneurial Networks
▸ Inter-organizational Networks
▸ Intra-organizational Networks
▸ Learning Networks
▸ Managerial Networks
▸ Social Network Analysis in Organizational Structures Evaluation

References

Bielenia-Grajewska M (2010) The linguistic dimension of expatriatism-hybrid environment, hybrid linguistic identity. Eur J Cross Cult Competence Manag 1(2/3):212–231
Bielenia-Grajewska M (2012) Linguistic aspects of informal learning in corporate online social networks. In: Dennen VP, Myers JB (eds) Virtual professional development. IGI Publishing, Hershey, pp 93–112
Bielenia-Grajewska M (2013) Corporate linguistic rights through the prism of company linguistic identity capital. In: Akrivopoulou C, Garipidis N (eds) Human rights and risks in the digital era: globalization and the effects of information technologies. IGI Publishing, Hershey, pp 275–290
Cornelissen J (2008) Corporate communication: a guide to theory and practice. Sage, London
Ferligoj A, Hlebec V (1999) Evaluation of social network measurement. Soc Netw 21(2):111–130
Fombrun CJ (1982) Strategies for network research in organizations. Acad Manag Rev 7(2):280–291
Gamper M, Reschke L (2010) Soziale Netzwerkanalyse. Eine Interdisziplinäre Erfolgsgeschichte. In: Gamper M, Reschke L (eds) Knoten und Kanten: Soziale Netzwerkanalyse in Wirtschafts- und Migrationsforschung. Transcript Verlag, Bielefeld, pp 13–54
Michael K, Michael MG (2008) Homo electricus and the continued speciation of humans. In: Quigley M (ed) Encyclopedia of information ethics and security. IGI Global, Hershey, pp 312–318
Mische A (2011) Relational sociology, culture and agency. In: Scott J, Carrington PJ (eds) The Sage handbook of social network analysis. Sage, Thousand Oaks, pp 80–98
Rughase O (2006) Identity and strategy: how individual visions enable the design of a market strategy that works. Edward Edgar Publishing, Cheltenham
Travica B (1999) New organizational designs: information aspects. Ablex Publishing Company, Stamford

Social Network Analysis in a Digital Age

Mariann Hardey
Lecturer, Marketing, University of Durham, Durham, UK

Synonyms

Analysis; Consumer; Digital; Linkages; Networks; Relationships; Social anthropology

Glossary

Consumer An individual or organization that uses a commodity or service
Research Systematic inquiry or investigation
Social Network Facilitates communication, provided by a network of related linkages
Digital Data Digitally sourced and/or published statistics or items of information

Definition

The term "social network analysis" provides an increasingly overarching research context for scholars, as well as policy makers, industry, commercial organizations, and the public sector. It refers not only to an integrated set of theoretical concepts and analytic methods but intends to explain a whole set of relationships

and their variations as these are defined by "a specific set of linkages" and their "additional property" (Mitchell 1969: 2). One outcome of the burgeoning territory of digital and continuous development of ubiquitous connectivity has meant that new informational infrastructures and data production provide a rich and diverse series of network contexts, that have significant implications for scholars and digital researchers. Recognized as one of the first social network researchers, Barnes asks a pertinent question that remains as relevant today: when collecting data, on social relations that may not hold any obvious limits, where does the researcher set their boundaries? (1979: 414). The study of mediated social networks and increasingly digital spaces has a well-established concentration as a field of study in sociology (cf. Wellman 1983; Rainie et al. 2012). However, the capture of such data deserves further scrutiny that calls into question the acquisition and observation of such linkages that do not remain fixed into place, and are these equally accessible to all observers. The issue of technology, and especially digital and social media platforms, brings social network analysis into contact with debates about the nature of its technology, heritage, culture, and processes. These are debates that are particularly challenging in terms of the observations that we may conduct in society and analysis that any researcher would seek to adopt. Indeed the persuasive Actor Network Theory (ANT) provides an additional explanation of networks that are beyond the agency of the social actors alone (see Callon 1987; Latour 1996), where networks themselves become additional vehicles for social analysis and evaluation.

There are some important points for any researcher to discern and place into context for social network analysis. First, it is crucial to establish, at least at an individual level, the role that technologies have in the formal production of social practices and relationships. This is to scrutinize how, and in what way, technologies may surround research development and be deeply bound to political and economic products, as well as cultural artifacts (see Fine and Kleinman 1983). Second, the researcher must think about the social interpretation of network construction. This means the contextualizing of everyday and seemingly mundane social interactions within networks and groups of networks. Alongside these social relationships, consumption, and consumerism take on significance. We live, as Mats Alvesson notes, in a society where trendy jargon, media appeal, and "looking good" define the successes of individuals, groups, and organizations. It is natural to extend this as an escalation of expectations that are part of the "gilt edge of life" (Alvesson 2013: 188), into networks that provide a growing focus for individuals to pursue high-status employment, and to seek out socially prestigious others. Third, in understanding social networks and stripping away the elements of analysis, it is necessary to reflect on Barnes's question and to put in place some context for boundaries; whether these are a temporal positioning, or shaded by cultural, social, political, or commercial significance. Every researcher needs to justify such efforts as part of a theoretical and methodological account and to make explicit their instrument handling and data processes.

There are, no doubt, additional vulnerabilities and opportunities that are geared to the stimulation of social network analysis. The points made here offer an overall condition that is specific to digital-scientific interest that has been often over-celebrated without critical treatment and that may lead us to new interpretations and encounters with new forms of data.

Cross-References

- ▸ Arts and Humanities, Complex Network Analysis of
- ▸ Community Evolution
- ▸ Dark Sides of Social Networking
- ▸ E-Commerce and Internet Business
- ▸ e-Government
- ▸ Online Privacy Paradox and Social Networks
- ▸ Quality of Social Network Data
- ▸ Social Capital
- ▸ Social Network Datasets
- ▸ Spatial Networks
- ▸ Visualization of Large Networks

References

Alvesson M (2013) The Triumph of Emptiness: Consumption, Higher Education, and Work Organization. Oxford University Press, Oxford

Barnes JA (1979) Network analysis: orientating notion, rigorous technique or substantive field of study. In: Laumann EO, Marsden PV, Prensky D (1989) The boundary specification problem in network analysis. Research Methods in Social Network Analysis 61:87

Callon M (1987) Society in the making: the study of technology as a tool for sociological analysis. Soc Constr Technol Syst 550:83–103

Fine GA, Kleinman S (1983) Network and meaning: an interactionist approach to structure. Sym In 6(1)

Latour B (1996) On actor-network theory: a few clarifications. Soziale Welt 47:369–381

Mitchell JC (ed.) (1969) Social networks in urban situations: analysis of personal relationships in central African towns.

Rainie H, Rainie L, Wellman B (2012) Networked: the new social operating system. MIT, Cambridge

Wellman B (1983) Network analysis: some basic principles. Sociol Theory 1.1:155–200

Social Network Analysis in Organizational Structures Evaluation

Radosław Michalski and Przemysław Kazienko
Institute of Informatics, Wrocław University of Technology, Wrocław, Poland

Synonyms

Corporate hierarchy; Enterprise management; Organizational design; Organizational network analysis; Social network analysis in organizations

Glossary

HR Human resources

Organizational Chart A diagram representing the formal structure in the organization

ONA – Organizational Network Analysis The analysis of the organization which focuses on the relationship between the informal social network and the formal structures in the organization: organizational charts, process definitions, and others

OSN – Organizational Social Network An informal social network, which was created from the data collected within the organization like e-mail logs, phone call records, surveys, and others

MSN – Multilayered Social Network The social network which consists of multiple layers; every of them represents different type of information used as a source for creating the network layer

SNA Social Network Analysis

Definition

Although typical social network analysis (SNA) may bring interesting results while being applied in an organizational environment, it is a very promising to have these results compared with the organization itself in order to gain additional knowledge about the whole organizational environment. This is caused by the fact that each organization at the moment of performing social network analysis possesses a more or less structured hierarchy which regulates the information and workflow in the formal way. Simultaneously, members of the organization maintain the informal social network by contacting and collaborating with each other. It means that the comparison of formal and informal structures may enhance the knowledge about the employees by means of their role in both of networks. In other words, one may say that the goal of this comparison is to check the discrepancy between the visible (defined, official) and the invisible (unofficial social network) structures in the organization.

Results of such analysis may bring the answer on some of the following questions: are the organization members well placed in the organizational chart? Is the organization maximizing its information and decision flow efficiency by using the recent organizational design? Are there any substantial discrepancies in the employees' formal role in the organization and their placement in the informal social network?

This kind of analysis is a key part of ONA – organizational network analysis, which may be described as a framework for understanding formal organizations (Knoke 2001).

Introduction

The possibility to collect the formal structure of the organization and its communication logs or collaboration traces has enabled the researchers to create a new field of social network analysis. By performing the comparison of both social networks – the formal and informal ones – some findings from such analysis may be useful for organization managers and the HR departments.

While finding a chance to gain a competitive advantage, organizations are searching for the solutions that would enable them to beat their market opponents. It may be crucial to discover, among various ways of increasing company effectiveness, own potential, hidden in the social network of the organization. The knowledge derived from this capacity, if properly extracted and interpreted, may lead to various positive effects in organization management (Palus et al. 2010; Song and van der Aalst 2008). The general idea of SNA application in organizational structure evaluation is illustrated in Fig. 1.

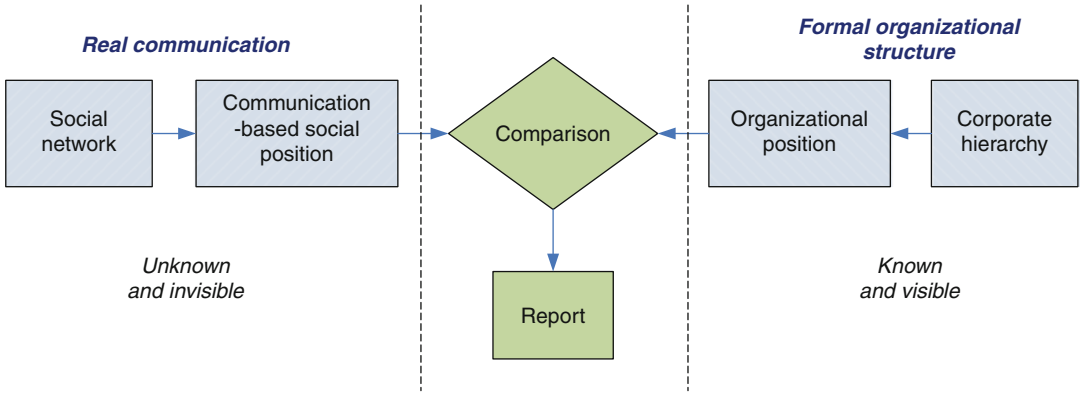
Managers may often ask the question about the proper alignment of their employees in the organization structure. The problem may be particularly important in fast-growing organizations, where medium-level management team may be chosen without prior adequate preparation and without the use of proper HR tools. Companies, in which some of the employees are awaiting retirement, are another example where the knowledge about real worker position may be vital. If a company decides to search internally for the successor or replacement of anybody, social network analysis may become helpful for such a task. It may also be helpful in extracting some prospective problems in the company, like managers avoiding communication with other employees within their units.

The problem of a proper organizational design is considered as a crucial one in corporate management (Daft 2009), because it strongly influences the information and decision flow. In that case, managers should permanently observe the internal structure of the company in terms of bottlenecks, overhead, or other possible problems. However, the task of performing such analyses manually becomes virtually impossible, due to the fact that the amount of information exchanged in the typical organization is too big even to be just observed. On the other hand, this process may be automated and more or less quantified by application of the social network analysis and using already existing information, such as organization charts and communication logs. Still, the final results should be treated individually, because some of the information may not be included in the analysis.

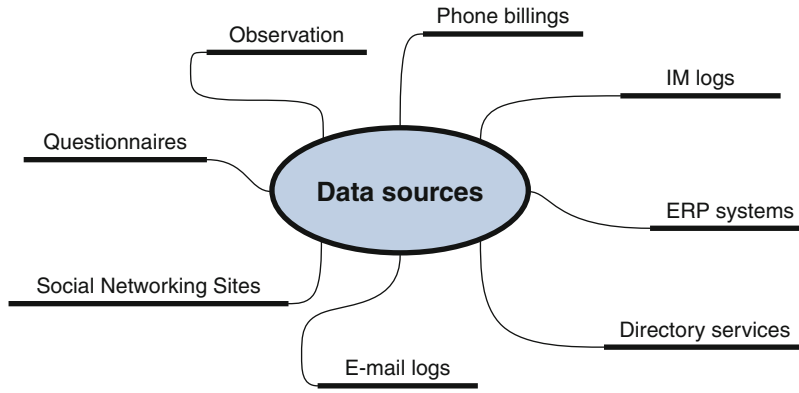
Key Points

To benefit from SNA applied to the organization, a number of steps have to be carried out. There is a need to obtain both formal and visible and informal and invisible structures of the organization; see Fig. 1.

Overall, in case of organizational structure evaluation, the most important information and the starting point is the organization chart. In the simplest scenario, the organization may use the functional design where every department is responsible for different tasks, each organization member belongs to only one department, and all the departments form a hierarchy, which starts from the management board (Daft 2009). However, nowadays, some more complicated scenarios also apply like process-oriented or matrix structures. These require some more effort while being analyzed, and they introduce additional limitations as well, which will be discussed later. The organization chart itself may also be represented by a graph that eases the further comparison of both structures. The informal part of the organization is often a multilayered social network, which was built by using variety of data sources available in the company. Some of these data sources were presented in Fig. 2.



Social Network Analysis in Organizational Structures Evaluation, Fig. 1 The idea of comparing formal and visible organizational structure with informal hidden social network based on real communication, based on Michalski et al. (2011)



Social Network Analysis in Organizational Structures Evaluation, Fig. 2 A choice of possible data sources for social network extraction in the organization (Michalski et al. 2011)

The process of evaluation of the organizational structure actually begins after performing some preliminary analysis understood as a feasibility study and conducting matching of entities of both networks, described later on. The overall result of the analysis is the report presenting the structural difference between formal and informal networks in the organization.

Summarizing, the whole analytical process consists of the following steps:

- Source data preprocessing
- Social network building
- Network measure calculation
- Social network and corporate hierarchy comparison

A complete process and framework that enables evaluating the organizational structure in the company by using SNA will be further depicted by using the example of the Enron company case.

Historical Background

The idea of performing the SNA in organizations is not completely new (Tichy et al. 1979); however, since the introduction of this idea, the general SNA measures and metrics were developed intensively (Wasserman and Faust 1994). At

the beginning, organizational network analysis (ONA) was more oriented to discovering key players in the organization without matching organizational social networks (OSN) to the organizational charts and finding the structural holes. Later on, some experiments related to uncovering the informal structure in the organizational social network were performed (Borgatti and Molina 2003), but the real enabler and accelerator in the research was the Enron case. It became especially famous worldwide in 2001 due to financial manipulation scandal. The Enron e-mail dataset was made public by the Federal Energy Regulatory Commission during its investigation (Klimt and Yang 2004). The Enron official hierarchy structure still remains publicly unavailable. However, there are some sources which can provide information concerning plenty of job positions of selected employees together with their department or division (Rowe et al. 2007). This led to a number of analyses which focused on the relationship between formal and informal structures (Rowe et al. 2007; Diesner et al. 2005; Hossain 2009; Borgatti and Molina 2003). Nevertheless, any new research is limited by the data availability – to fully evaluate the organization by means of the social network, as it was previously described, not only the social network data but also organizational details are needed. As a result, not all the organizations are happy to disclose this information limiting further research opportunity.

Analysis of Organizational Structures

A social network, which is built on the basis of employees communication logs, may be found useful in the evaluation of formal organizational structures existing in the company. The communication-based social network can provide information about social network leaders, communication gaps, and anomalies. However, the problem is what factors in the social network analysis results should be considered as important ones and useful in further company management decisions. Another problem is how to perform such analysis in order to ensure its acceptable meaningfulness and representativeness.

Although this sort of analyses is mostly performed for business companies, other types of organizations may benefit by performing such a study as well. Moreover, if a company has introduced some organizational changes in order to reach some goals that should also result in communication changes, these kinds of comparisons become essentially useful for validation of such changes.

The variety of communication forms, which are already used in organizations, such as e-mails, instant messaging systems, ERP systems, and landline and mobile phones, allow to build the social network as a multilayered one; each layer corresponds to another communication channel. All these data sources facilitate gathering more information about the whole communication in the organization and include more organization members. Additionally, by applying separate weights to these layers, some of them may be interpreted as more important or more *social* ones.

Although the formal structure influences all the communication between members of the organization, yet there still exists the space for the informal communication. However, there is no easy way to distinguish both types of communication without complex and resource-consuming analysis. But still, as it will be shown, such a distinction is not necessarily needed to perform meaningful and useful analyses.

In this entry, the typical social network analysis in organizational structure evaluation is presented. Especially, the role the preliminary part of the whole process – the feasibility study – is underlined and ideas regarding the analysis of the dynamics of the organizational network are introduced. To complete the study, selected limitations of the proposed approach are presented as well. The case study on the Enron company depicts usability of the proposed method.

Feasibility Study

As it was mentioned in the Introduction section, the first step of the process is the source data preprocessing which may be also treated as a separate feasibility study, because this part may even disqualify all source data or part of the company from further analysis.

There are at least five important factors to be taken under consideration while obtaining the data and performing data preprocessing:

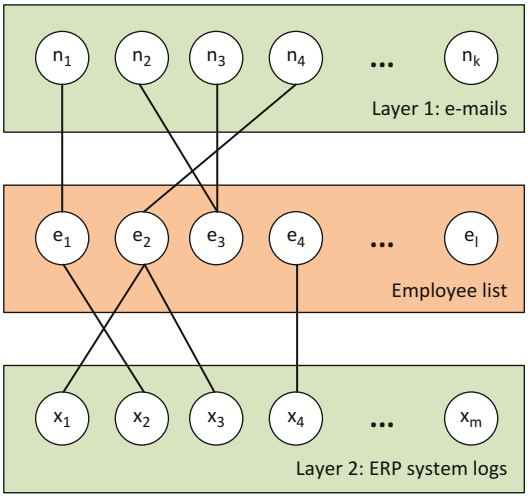
- The choice of data sources
- Adjustment of the period of the analysis
- Extraction of the official organizational structure
- Matching informal social network entities to employees
- Employees using external communication channels (or simply not registered by the organization)

To obtain the comprehensive communication social network, it is worth to consider multiple data sources: phone calls, e-mails, instant messaging systems, or even ERP and workflow management systems. These will be used to build a multilayered social network (MSN), a directed and weighted graph with weights representing the importance and intensity of relations between users. The chance to gather more valuable results grows up by using more than one data source, but not without consequences. First of all, if there are multiple data sources, the researcher should try to obtain the data of the same or at least similar time frame. It is caused by the fact that the social network evolves and the relations (in that case weights of the graph edges) change over time. It is nearly impossible to create a representative multilayered social network by using the data from different periods, so the nonoverlapping periods shall be cut off or only the common part of the datasets should be utilized.

When designing the study, a researcher faces the problem of choosing the period of the analysis. The shorter the period, the chance for matching most of the nodes in the graph decreases. However, for the longer periods (years), it is necessary to tackle with the problem of the probably smaller importance of old communication compared to the new one. Typically, it is solved by assignment of higher weights to newer communication (Kazienko et al. 2009). On the other hand, some other difficulties are partially overcome for the longer periods: holidays and longer illness absences. In general, the period of half a year should be considered as representative enough. However, it is also expected that the formal structure of the company would not change strongly over time of analysis. Otherwise, it would be hard to compare both the social network and the formal structure. Concluding, the above limitations clearly show that the selection of the most suitable time frame for the analysis may be challenging.

Yet one more problem is related to the organizational hierarchy. Paradoxically, such kind of relatively well-defined corporate relations, i.e., the organizational structure, can be hardly extracted automatically, because, depending on the size and profile, the company may have no need to maintain the full company structure drill-down from board through departments up to a single employee in their IT systems. That is why it may be necessary to convert organizational structure, taken from official documents, into a graph, where nodes represent employees and edges – employee-supervisor relations.

Another important problem is the need for finding the same entities in the datasets, as presented in Fig. 3. If we consider the employees as entities, there should exist a mapping between any entities in the other data sources to the employee. However, due to the nature of communication systems, such a mapping may not exist at all. To present one of the examples, let us assume that albeit every warehouse worker has got his individual e-mail account, all these workers may use the same one account in the instant messenger (IM) system. In that case, the researcher should decide whether the IM layer will be analyzed deeper, the IM layer will be discarded completely, or only the warehouse will be excluded from the analysis. Alternatively, the whole warehouse department will be treated as a single node, and only its relationships with other departments will be studied. Another source of potential problems is the fact that the company



Social Network Analysis in Organizational Structures Evaluation, Fig. 3 Mapping social network actors to employee list (Kazienko et al. 2011)

may use multiple aliases for a single employee, for instance, in the e-mail system – in that case they should be merged and mapped to the single entity in the employee list. Of course, it is also possible that an employee would not use some communication systems, which actually is not a problem for the analysis – a node may be isolated in some layers.

There might be some employees using external communication techniques (supervisory board, expats), and some employees may be represented by other ones, as it often happens for top-level managers – they are substituted by their assistants for writing and sending e-mails. This should be respected at the preprocessing stage.

The problems discussed above are to be found in most organizations. However, they may be overcome and there is still the chance to conduct reliable studies. However, some more limitations may be recognized in the organization itself and they were described in the Limitations section.

Building the Social Network and the Corporate Hierarchy

The process of building social network consists of choosing the graph type (directed or undirected) and weight calculation method for edges linking nodes. In general, various methods of relationship valuation may be applied including multilayered (multigraph) concepts, in which two nodes are connected by means of multiple edges. However, in the Enron dataset utilized in the use case, only one layer exists, so a single-layered social network has been built. A directed and weighted graph may be created from e-mail logs using the following formula for the weight of an edge between node i and j:

$$w_{ij} = \frac{\sum e_{ij}}{\sum e_i}, \tag{1}$$

where $\sum e_{ij}$ is the number of e-mails sent by node i to node j and $\sum e_i$ is a total number of e-mails sent by i. It means that weight w_{ij} focuses on the local neighborhood of an employee rather than on global network characteristic. As it was mentioned earlier, it is also possible to extend the above approach by applying the importance of correspondence in terms of time. Then, each email would not be counted as 1 (e_{ij}=1) but as a fraction of 1 depending on its time stamp – smaller values for older messages: $e_{ij} = 1/\lambda^k$, where $\lambda \in (0; 1]$ is constant, e.g., 0.8, and k is the period index (0 – for the newest period, 1 – for the previous one, and so on). Obviously, instead of numbers of e-mails exchanged, some other communication logs may be used like phone records or IM chats.

After building the social network, it is also required to obtain the organizational hierarchy. Paradoxically, sometimes it can be even more difficult than creation of social network. It refers especially informal and vague organizational structures. Especially for larger companies, their hierarchy may be extracted from internal phone books or other catalogues.

Introducing the Network Measures

A variety of measures were computed to reflect different aspects of the importance of the node in the social networks (Wasserman and Faust 1994; Scott 2000). They were combined into

a single value – social score, as presented in Rowe et al. (2007) and Palus et al. (2010) by including:

(a) E-mails count – the number of e-mails a user has sent and received.
(b) Average response time – the time elapsed between a user sent an e-mail and later received a response e-mail from that same person. The exchange of this nature is considered a "response" only if a received message succeeds a sent message within three business days.
(c) Response score – a combination of the number of responses and average response time.
(d) Number of cliques – the number of maximal complete subgraphs that the account belongs to.
(e) Raw clique score – a score computed using a size of the given account's clique set. Bigger cliques are worth more than smaller ones; importance increases exponentially with size.
(f) Weighted clique score – a score computed using the importance of the people in each clique, which is computed strictly from the number of e-mails and the average response time.
(g) Centrality degree – count of the number of ties to other actors (nodes) in the network.
(h) Clustering coefficient – likelihood that two associates of a node are also linked with themselves.
(i) Mean of the shortest path length from a specific vertex to all vertices in the graph.
(j) Betweenness centrality – reflects the contribution of a given node in all shortest paths connecting all pairs of nodes, i.e., how important is a node in linking other nodes.
(k) *Hubs-and-authorities* importance – refers to the algorithm proposed in Kleinberg (1999).

Above measures are then weighted and normalized to a [0, 100] scale, as presented in Rowe et al. (2007). Obviously, some other measures can be utilized in a given organization according to the needs and data availability.

Hierarchical Position

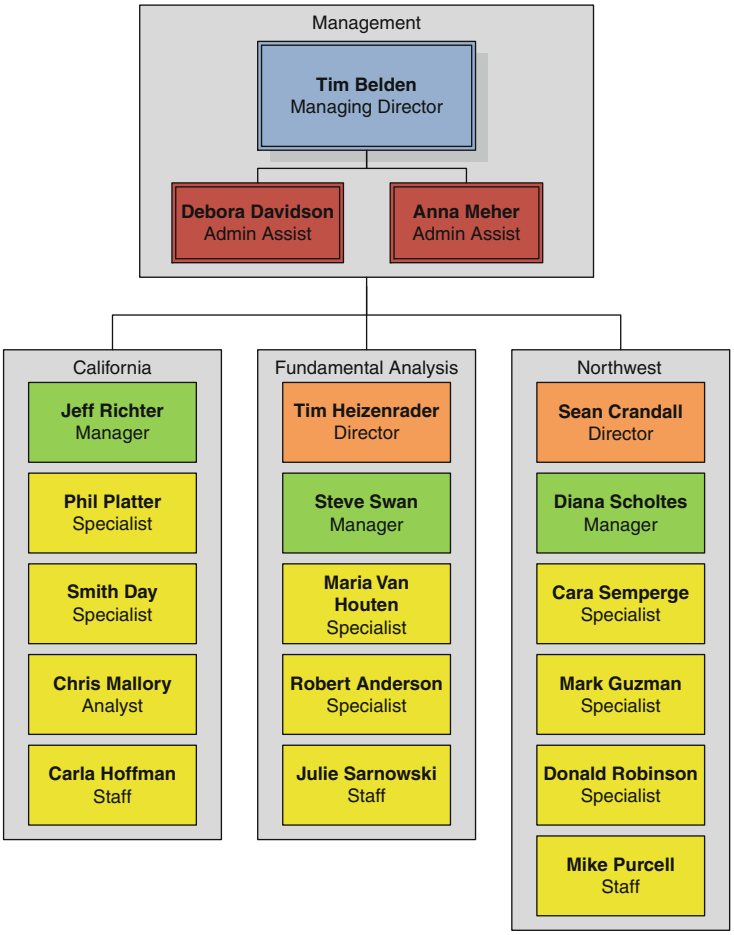
It is possible to find people who are higher or lower in the hierarchy for each employee in the corporate hierarchy. The Hierarchical Position (HP) is a measure that denotes the importance of an employee within the company (Kleinberg 1999). For each user i in a company C, there is a sum of hierarchical differences D between i and any other user j in the company normalized by the total number of other users.

$$HP(i) = \frac{\sum_{j \in C \wedge j \neq i} D\,(i, j)}{m - 1} \tag{2}$$

The hierarchical difference *D(ij)* is computed as follows:

$$D(i, j) = \begin{cases} 1, & \textit{if i is higher in the hierarchy than j} \\ 0, & \textit{if i and j are at the same level of the hierarchy} \\ -1, & \textit{if i is lower in the hierarchy than j} \end{cases} \tag{3}$$

At first, the Kendall's rankings comparison method was used to compare two rankings (Kendall and Gibbons 1990). It compares the nodes in pairs, i.e., the positions of pair nodes within both rankings. If the position of node A is related to the position of node B in both rankings monotonically in the same direction (lower or higher in the both hierarchies), then this pair is well correlated. It is assumed that when the level in hierarchy is the same within the pair, then it does not matter whether they are in different positions in the second ranking. Kendall's τ rank correlation coefficient is a value from the $[-1, 1]$ range, where 1 means that two rankings are perfectly correlated and -1 means that they are completely different (in the opposite order).



Social Network Analysis in Organizational Structures Evaluation, Fig. 4 A fragment of the Enron corporate hierarchy (Palus et al. 2010)

It is impossible to distinguish the importance of departments, e.g., whether the Director of Northwest is higher in the hierarchy than the Director of Fundamental Analysis; see Fig. 4. Thus, analyses in the Enron use case were not performed globally, but locally at department level.

Discovering the Organizational Level of Employees

The structural node measures within the social network can be utilized to predict hierarchy level of particular nodes – employees. Some of these structural features may be more while the other less correlated with the organizational level. As a result, the level of a given person in the organizational hierarchy may be discovered based only on this person's centrality measures in the social network.

Some of these typical centrality measures were compared for the Enron employees; see the use case described in the following section. The results are presented in Table 1.

The above analysis clearly shows that some measures like in-degree centrality and centrality eigenvector are able to identify the level of the employee with quite good accuracy and that there exists the general relation between the employees' social network position and the corporate hierarchy placement.

Social Network Analysis in Organizational Structures Evaluation, Table 1 The accuracy of management level matching while using various social network metrics (Michalski et al. 2011)

	Percentage of the management level employees matched	Percentage of regular employees matched
In-degree centrality	67	85
Out-degree centrality	50	77
Centrality betweenness	33	69
Centrality closeness	33	69
Clustering coefficient	17	62
Centrality eigenvector	67	85

Enron Use Case

The use case for evaluation of the organizational structure will be presented based on the Enron dataset that contains e-mail communication between employees. This e-mail corpus is extracted from mailboxes of 150 Enron employees, mostly senior management. In total, they contain 517,430 e-mail messages. Because this is the only available communication channel, only one layer in the social network, a single-layered social network, may be built (Klimt and Yang 2004).

Having the social network built, the organizational hierarchy must be identified. There is an Excel file with a list of over 160 employees and their job title available at Shetty and Adibi (2004). Many of them do not exist in the Enron Corpus, though. Using this list, four groups from Enron North American West Power Traders are chosen – it is possible to distinguish levels of hierarchy by matching them with job titles. Since only a part of hierarchy is available, the most complete part of it has been taken for further analysis. The extracted hierarchy is presented in Fig. 4.

The list of Enron employees sorted by their *social score* (see Introducing the Network Measures section) is presented in Table 2. The *HP* measure (Eq. 2) and *Position* column indicates official hierarchy structure. It can be seen very clearly that social scores of the management is far higher than the others.

The diagram of *Hierarchical Position* should be descending, but there are deep structural holes, as presented in Fig. 5.

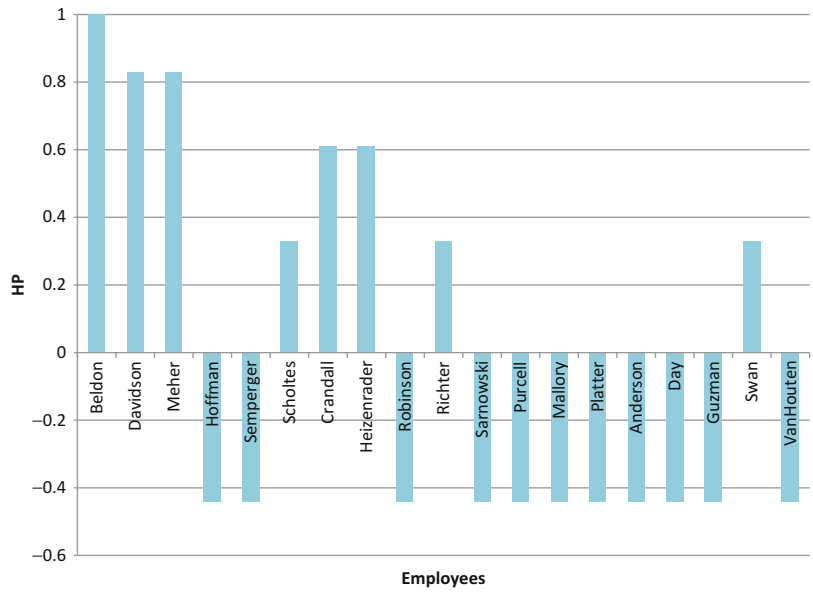
The summary of Kendall's correlation coefficient between the official hierarchy (ordered by *HP*) and the one derived from the social network (ordered by *social score*) for chosen departments is presented in Table 3.

The main problem with the Enron dataset is lack of information about direct hierarchy structure; only partial information was known. However, the analysis shows the rankings are very similar with Kendall's rank over 0.6 with management department perfectly identical (Kendall's rank of 1).

An interesting fact is that all employees who are lower in the hierarchy than who comes from the social network are women. There are 7 women among 19 analyzed employees, and there are 5 female workers in the top 6 of the social ranking, while 4 have been classified as lowest-level employees according to the hierarchy (these are marked green in Table 2). There are three possible reasons of such case. Firstly, a wrong assumption has been made while ranking job titles. Secondly, there can be a simple but important reason that women are underestimated and maybe should occupy higher company positions. The last, but not less probable, is that women may be more likely to gossip than men, and this fact is disrupting the process of proper social network extraction. It is possible that the real reason is the combination of these three.

Social Network Analysis in Organizational Structures Evaluation, Table 2 Social measures for Enron employees sorted by the social score (Palus et al. 2010)

Name	Surname	Position	Lvl	HP	Degree	Betweenness	Hubs	Clustering	Social score
Tim	Beldon	Managing Director	1	1.00	83	370.35	0.04	0.40	75.68
Debora	Davidson	Admin assist	2	0.83	66	278.35	0.04	0.41	63.51
Anna	Meher	Admin assist	2	0.83	62	260.94	0.04	0.42	62.84
Carla	Hoffman	Staff	5	−0.44	55	143.98	0.04	0.49	61.67
Cara	Semperger	Specialist	5	−0.44	63	82.96	0.03	0.52	53.68
Diana	Scholtes	Manager	4	0.33	45	21.44	0.03	0.70	53.31
Sean	Crandall	Director	3	0.61	42	40.04	0.03	0.62	43.64
Tim	Heizenrader	Director	3	0.61	33	19.45	0.02	0.71	35.56
Donald	Robinson	Specialist	5	−0.44	27	6.67	0.02	0.81	33.03
Jeff	Richter	Manager	4	0.33	25	12.80	0.02	0.74	32.53
Julie	Sarnowski	Staff	5	−0.44	28	25.94	0.02	0.63	32.14
Mike	Purcell	Staff	5	−0.44	24	5.02	0.02	0.79	30.36
Chris	Mallory	Analyst	5	−0.44	27	9.92	0.02	0.76	30.19
Phil	Platter	Specialist	5	−0.44	33	34.34	0.02	0.63	27.90
Robert	Anderson	Specialist	5	−0.44	8	0.15	0.01	0.96	20.06
Smith	Day	Specialist	5	−0.44	6	0.00	0.01	1.00	20.00
Mark	Guzman	Specialist	5	−0.44	18	6.84	0.01	0.75	19.97
Steve	Swan	Manager	4	0.33	9	0.20	0.01	0.93	19.55
Maria	VanHouten	Specialist	5	−0.44	7	0.11	0.01	0.95	19.44



Social Network Analysis in Organizational Structures Evaluation, Fig. 5 Hierarchical positions of Enron employees sorted by social score (Palus et al. 2010)

Social Network Analysis in Organizational Structures Evaluation, Table 3 Kendall's correlation coefficient for each department between official hierarchy and the social network (Palus et al. 2010)

Department	Kendall's correlation coefficient
Management (official vs. SN)	1.0
California (official vs. SN)	0.8
Fundamental Analysis (official vs. SN)	0.6
Northwest (official vs. SN)	0.6

Limitations

It must be clearly stated that the comparative analysis may be applied in the more effective way to companies with the stable (probably functional) organization design (Daft 2009), because other designs, such as matrix or horizontal ones, would not allow to create a hierarchy chart easy comparable with the social network ranks.

However, while performing such an analysis, there is also the need to consider ethical aspects of performing such studies inside the organization (Borgatti and Molina 2003). That is why every result of the evaluation should be individually interpreted and discussed. In particular, the access to organizational communication logs may be treated as violation of privacy protection restrictions. Sometimes, it may be necessary to obtain individual employee permissions to process the data.

It is also related to very important limitation: this kind of analyses require to process data, which may be considered as very sensitive for the organization. Even while applying anonymization procedures on the data, there is a big chance to map entities to real employees, especially while analyzing the organizational hierarchy. In fact, the organization must be trustful and convinced to share this kind of information with researchers.

Key Applications

The idea of matching organizational structure and the social network may be regarded as another possible way to improve overall company management. The idea focuses on the comparison of calculated node position ranks using chosen measures within the organization structure.

The positive results in comparison of formal structure with communication-based social network may mean that the similar level of managers and regular employees was properly assigned to their management levels. However, if the results differ significantly, some more sophisticated analysis might be needed to answer the question why real-life communication and hierarchy do not necessarily cover organization chart. The reasons may be different: (i) not the most important social network source has been analyzed; (ii) due to the profile of the company, communication between company members have nothing to do with their formal position; (ii) the relations change too fast to give stable point of view; or (iv) the company chose inappropriate persons to hold some management positions.

There might be also another usage of proposed concepts. The choice of new leaders in the organization can be supported by the application of the described set of methods, i.e., through recognition as prospective candidate for managers those employees who belong to the higher level of the management team in based on unofficial communication (having compared to the formal organization structure).

There is one more, more controversial, application field of the considered methodology. If someone wants to uncover organizational hierarchy, e.g., for crime groups, or at least wishes to know possible managers of this organization using available communication logs (phone records registered by the telecom company), they may discover organization managers in the easier, faster, and safer (passive) way. This may be used by the police in their investigations.

Despite all the techniques regarding core data analysis that may be very ambitious for SNA experts, the real challenge for companies is to properly interpret and make valuable use of the achieved corporate SNA results.

Future Directions

A very promising field in ONA is related to the dynamics of the organization. At monitoring the social network in organizations and calculating the HP values, one may discover that in some parts of the organization some problems, arise even before they will be officially mentioned.

The other usage of the above approach is the ability to observe how fast the just introduced organizational changes are influencing the communication social network. It could be used for validation of motivation programs.

Cross-References

▸ Anonymization and De-anonymization of Social Network Data
▸ Collection and Analysis of Relational Data in Organizational and Market Settings
▸ Managerial Networks
▸ Multilayered Social Networks
▸ Process of Social Network Analysis

Acknowledgments

The work was partially supported by fellowship cofinanced by the European Union within the European Social Fund, the Polish Ministry of Science and Higher Education, and the research project 2010–2013.

References

Borgatti SP, Molina JL (2003) Ethical and strategic issues in organizational social network analysis. J Appl Behav Sci 39(3):337–349
Daft RL (2009) Organization theory and design, 10th edn. Cengage Learning, Cincinati
Diesner J, Frantz TL, Carley KM (2005) Communication networks from the Enron email corpus "It's always about the people. Enron is no different". Comput Math Organ Theory 11(3):201–228
Hossain L (2009) Effect of organisational position and network centrality on project coordination. Int J Proj Manag 27(7):680–689
Kazienko P, Musiał K, Zgrzywa A (2009) Evaluation of node position based on email communication. Control Cybern 38(1):67–86
Kazienko P, Michalski R, Palus S (2011) Social network analysis as a tool for improving enterprise architecture.

In: KES-AMSTA 2011, the 5th international KES symposium on agents and multi-agent systems – technologies and applications. Lecture notes in artificial intelligence, LNAI 6682. Springer, Berlin Heidelberg, pp 651–660
Kendall MG, Gibbons JD (1990) Rank correlation methods, 5th edn. Edward Arnold, A Division of Hodder & Sloughton, London
Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632
Klimt B, Yang Y (2004) Introducing the Enron corpus. In: CEAS 2004, 1st conference on email and anti-spam, Mountain View
Knoke D (2001) Changing organizations: business networks in the new political economy. Westview Press, Boulder
Michalski R, Palus S, Kazienko P (2011) Matching organizational structure and social network extracted from email communication. In: BIS 2011, 14th international conference on business information systems. Lecture notes in business information processing, LNBIP 87. Springer Berlin Heidelberg, pp 197–206
Palus S, Bródka P, Kazienko P (2010) How to analyze company using social network? In: WSKS 2010, the 3rd world summit on the knowledge society, Corfu, Greece, 22–24 Sept 2010. Communications in computer and information science, vol 111. Springer, Berlin/Heidelberg, pp 159–164
Rowe R, Creamer G, Hershkop S, Stolfo SJ (2007) Automated social hierarchy detection through email network analysis. In: Proceedings of the 9th WebKDD and 1st SNAKDD 2007 workshop on Web mining and social network analysis WebKDD/SNAKDD 2007. ACM, New York, pp 109–117
Scott J (2000) Social network analysis: a handbook, vol 3(5). Sage, Thousand Oaks, p 208
Shetty J, Adibi J (2004) Ex employee status report. http://www.isi.edu/∼adibi/Enron/Enron_Employee_Status.xls
Song M, van der Aalst WMP (2008) Towards comprehensive support for organizational mining. Decis Support Syst 46(1):300–317
Tichy NM, Tushman ML, Fombrun C (1979) Social network analysis for organizations. Acad Manag Rev 4(4):507–519
Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge

Social Network Analysis in Organizations

▸ Social Network Analysis in Organizational Structures Evaluation

Social Network Analysis Packages

▸ Tools for Networks

Social Network Anonymity

▸ Privacy in Social Networks, Current and Future Research Trends on

Social Network Data

▸ Socio-Graph Representations, Concepts, Data, and Analysis

Social Network Datasets

Jérôme Kunegis
Institute for Web Science and Technologies, University of Koblenz–Landau, Koblenz, Germany

Synonyms

Network dataset; Social graph dataset

Glossary

Bipartite A network is bipartite when it contains two distinct node types, and all edges connect a node of the first type with a node of the second type

Directed A network is directed when each edge has an orientation, i.e., each edge explicitly goes *from* one node *to* another node

Timestamps When a network has timestamps, the creation time of each edge is known

Undirected A network is undirected when its edges do not have an orientation

Unipartite A network is unipartite when it contains a single node type

Weighted A network is weighted if its edges are labeled with edge weights, for instance, rating values

Definition

A social network dataset is a dataset containing the structural information of a social network. In the general case, a social network dataset consists of persons connected by edges. Social network datasets can represent friendship relationships or may be extracted from a social networking Web site (Kunegis 2013). Social network datasets are widely used, not only in the area of social network analysis but also in the areas of data mining, Web science, and network analysis as the basis for various kinds of research.

Introduction

In order to study social networks, social network datasets are necessary. Thus, the availability of social network datasets are of crucial importance in all disciplines covering social networks. Beyond the area of social network analysis, social networks are studied in such diverse fields as data mining, Web science, network science, recommender systems, and many more. In fact, the majority of research being performed in these fields takes the form of the analysis of a social network and its usage as the basis of further analyses. Therefore, an increasing number of social network datasets are used in the literature, of which more and more are openly available.

Since the success of social media platforms such as Facebook and Twitter with the general public, these social networks have been increasingly studied, and accordingly a high number of datasets of these sites are available.

Historical Background

Historically, sociologists and anthropologists have studied social networks either theoretically by constructing corresponding models, or have observed or conducted surveys and then assembled social network datasets by hand. As an arbitrary example, the article *Cultures of the Central Highlands, New Guinea* by Read (1954) lists, in the form of a table, the 55 relationships between 16 tribes of the Central Highlands in New Guinea.

While a network of 16 nodes is perfectly correct in that it faithfully represents that actual relationships between tribes, its 55 edges are too few for performing statistical analyses. For instance, a common network analysis tool is the *degree distribution*, in which the number of nodes having a specific number of neighbors (the degree) are counted, resulting in the observation of *power laws*. These power laws reflect the fact that the number of nodes with n neighbors is proportional to $n^{-\gamma}$, for some constant γ. When applied to the network of New Guinean tribes, power laws are not observed. This does not mean however that the tribal network is in any way special. It does just mean that the network is too small to observe power laws. In fact, the larger a network, the easier it is to make statistically significant observations. Thus, small networks assembled by hand are too small for data mining applications. Instead, larger networks must be used.

Large social network datasets have become possible with the World Wide Web. With the availability of online social networking sites, large social network datasets have become available for research and other purposes. Nowadays, a large selection of large network datasets can be used, although many datasets are still proprietary and only available to the research divisions of social networking companies. Additionally, social media is used to collect other types of networks, for instance, rating graphs, consisting of ratings by users of items, or communication networks, consisting of individual messages such as emails sent between users.

Social Network Datasets

A social network datasets is mathematically a graph, with optionally additional structure. In the simplest case, a social network dataset is simply a graph

$$G = (V, E)$$

in which the vertex set V represents the users and the edge set E represents the friendships. In these kinds of datasets, each edge $\{i, j\}$ is undirected as are the friendships on Facebook (www.facebook.com) rather than directed as is the *follow* relationship on Twitter (twitter.com). Also, these kinds of network datasets allow only a single edge between two nodes.

In the following subsections, we will first describe basic statistics and analyses that can be computed and performed with social network datasets, and then describe additional forms of structure associated with social network datasets. The possible structural features of network datasets are summarized in Table 1.

Network Dataset Statistics

Trivial statistics of a social network dataset are the number of nodes $|V|$, which is also called the *size* of a network, and the number of edges $|E|$, also called the *volume* of a network. The size of social network datasets range from a dozen for pre-Internet networks from anthropology and sociology to several hundred thousands for large social networking sites such as Facebook (Backstrom et al. 2012) and Twitter (Kwak et al. 2010). Figure 1 shows an overview of the network datasets from the KONECT project (Kunegis 2013), a collection of network datasets of typical sizes.

Other common statistics are described in the following. We must note that not all notations are established: While graphs almost universally written as $G = (V, E)$, the average degree, for instance, may be denoted by several symbols. The notation we use here represents a reasonable choice in symbols, although it is not universal.

The *average degree* is used as a statistic and ranges from 1 to about 100 in the most dense networks datasets. The average degree can be defined as

$$d = 2|E|/|V|.$$

The *fill* is the proportion of edges to the number of total possible edges. The fill can be defined as

$$f = |E|/(\frac{1}{2}|V|(|V| - 1)).$$

Both the average degree and the fill are sometimes called the *density* in the literature.

The size of the *largest connected component* is sometimes given as a social network statistic, although social network datasets are often connected, making this statistic equal to the network size.

The *clustering coefficient* c equals the probability that two friends of a single persons are themselves friends. The clustering coefficient is thus a number between zero and one. A high clustering coefficient in social networks is used as an indication that a network is a *small-world network* (Watts and Strogatz 1998).

The *algebraic connectivity* a is defined as the second-smallest eigenvalue of the social network's Laplacian matrix $\mathbf{L}$ (Fiedler 1973). The algebraic connectivity is zero when the network is not connected; otherwise, it is larger than zero. The algebraic connectivity is used to measure the connectivity of a network.

The *spectral norm* $\|\mathbf{A}\|_2$ of a network equals the largest absolute eigenvalue of its adjacency matrix. The spectral norm is used as a measure of the *size* of a network, complementing the volume.

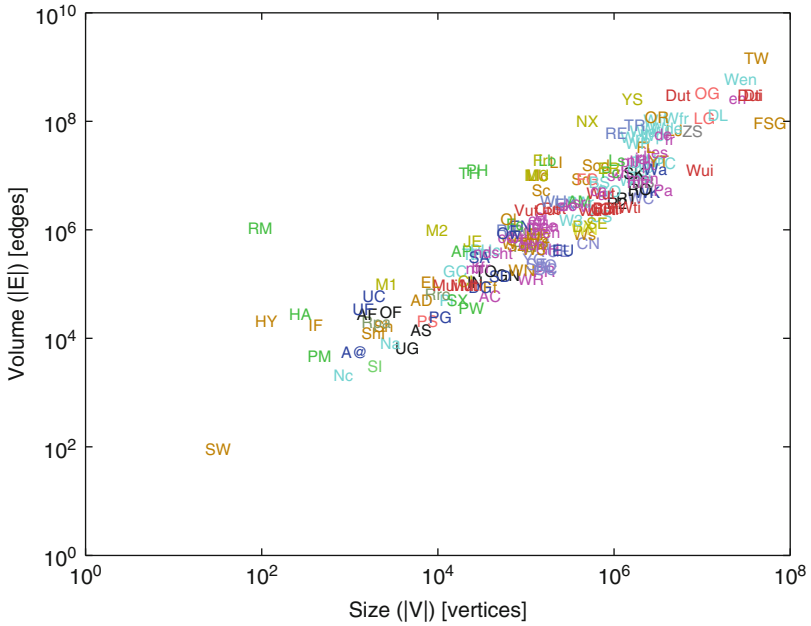
The *diameter* δ of a network equals the length of the longest path in the network. A small value of the diameter is used in conjunction with the clustering coefficient to characterize a social network as a *small-world network* (Watts and Strogatz 1998). As a robust replacement of the diameter, the following measures are often used:

- The 90-% effective diameter equals the number of edges one must take to reach 90 % of all nodes, on average.
- The mean path length is defined as the average of the distance between all node pairs.

Typical values for the diameter range from 4 to 6.

Social Network Datasets, Table 1 The possible structural features of social network datasets. Each of these features is described in one subsection

Feature	Description
Directed network datasets	Each edge is directed
Bipartite network dataset	There are two node types; each edge connects two nodes of different type
Network datasets with multiple edges	Multiple edges are permitted between any node pair
Signed network datasets	Edges can be positive or negative
Rating network datasets	Each edge represents a rating and is thus annotated with a rating value
Temporal network datasets	Each edge is annotated with an edge creation time, allowing the evolution of the network to be studied
Multirelational network datasets	Multiple edge types exist
Typed networks	There are multiple node and edge types



Social Network Datasets, Fig. 1 A typical collection of network datasets from social media, from the KONECT project (Kunegis 2013). Each *letter code* represents one network dataset

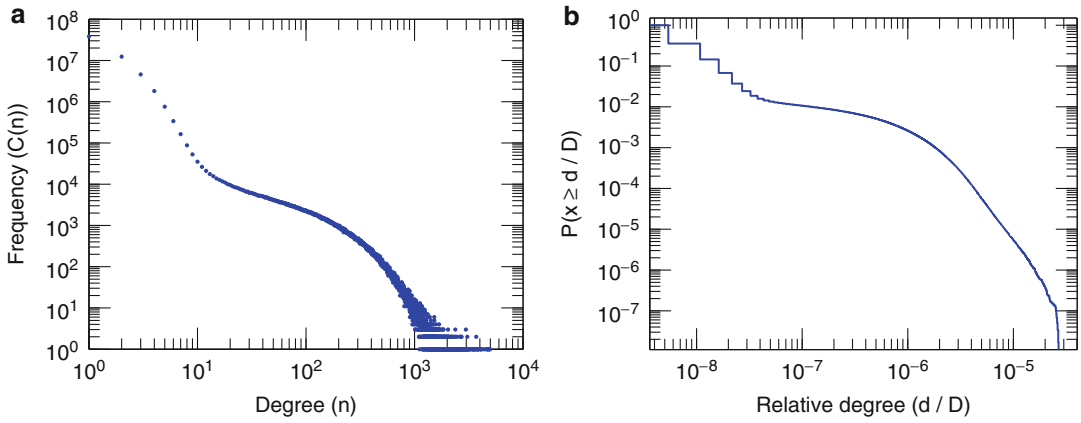
The list of social network statistics is much longer, and new statistics are constantly introduced in the literature.

Network Dataset Analyses

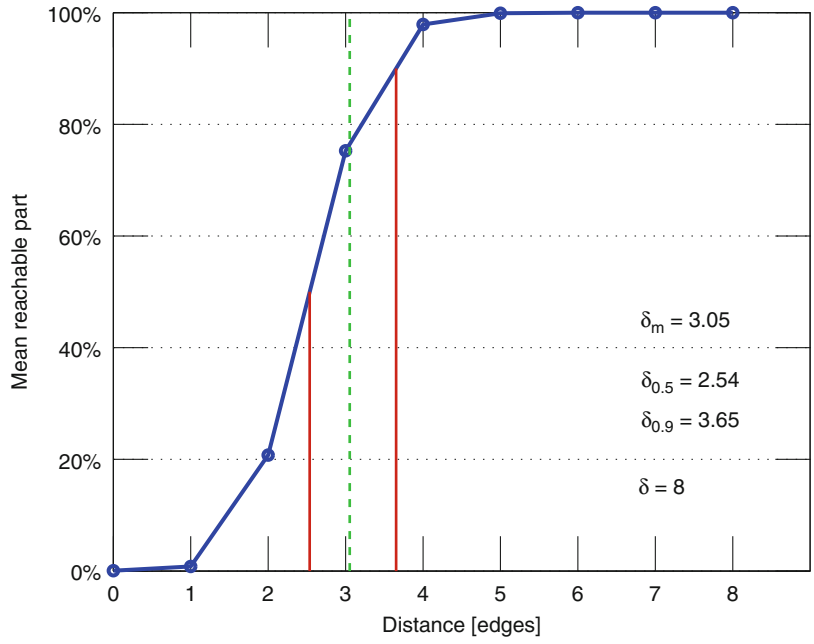
Social network datasets are used as the basis for a large number of analysis types. In this section, we review several very common types of analyses.

The *degree distribution* represents the distribution of the degree values, i.e., the number of neighbors in the social graphs, over all nodes in the networks. The degree distribution can be visualized in several ways, of which the most common is by far the simple degree distribution plot, and another is the cumulated degree distribution plot. Figure 2 shows the two types of plots for a subset of the Facebook social network (Gjoka et al. 2010). Both plots use a doubly logarithmic scale. Both degree distribution plots are typically used to point out a *power law*, i.e., the observation than the number of nodes with degree n is proportional to $n^{-\gamma}$ for a constant γ.

The second plot type we show is the hop plot. The hop plot shows, for each possible distance n, the average number of nodes at distance n from any nodes in the network. The hop plot can be used to read out the diameter, mean average path length, and 90-% effective diameter of the



Social Network Datasets, Fig. 2 The simple and cumulated degree distributions of a subset of the Facebook social network dataset from Gjoka et al. (2010). Both plots are shown on a doubly logarithmic scale. (**a**) Degree distribution. (**b**) Cumulated degree distribution



Social Network Datasets, Fig. 3 The hop plot of the online social network of users of an online community of students from the University of California at Irvine (Opsahl and Panzarasa 2009)

network. The plot can also be used to measure the median path length in the network, which corresponds to the 50-% effective diameter. The hop plot is expensive to compute. Figure 3 shows the hop plot of the online social network of users of an online community of students from the University of California at Irvine (Opsahl and Panzarasa 2009).

Directed Network Datasets

Some social networks have directed edges. An example are trust networks: The fact that person A trusts person B is independent of the fact the person B trusts person A. Thus, trust networks are directed and have directed edges.

Mathematically, directed networks are written as $D = (V, A)$, in which D stands for *digraph*

(an abbreviation of *directed graph*) and A is the set of arcs (or *directed edges*). A directed edge between nodes i and j is usually denoted (i, j) in contrast to the notation $\{i, j\}$ used for undirected graphs.

As an example for statistics specific to directed networks, the largest connected component can be extended to the largest strongly connected component. The strongly connected component of a directed network is defined as the largest set of nodes in the networks in which every node is reachable from every other node, using only directed paths.

In directed networks, two degrees are defined: the outdegree and the indegree. Thus, in addition to the usual degree distribution, the outdegree distribution and the indegree distribution can be defined. An example of an analysis in which both distribution behave differently are power laws: Indegree distributions follow much more often power laws than outdegree distribution.

Another key feature of directed networks are reflected in algebraic graph theory, i.e., those methods that represent the social network as a matrix. In an undirected social network, the adjacency matrix $\mathbf{A}$ defined as $\mathbf{A}_{ij} = 1$ when $\{i, j\}$ is an edge and $\mathbf{A}_{ij} = 0$ when otherwise is symmetric. In directed networks, the matrix $\mathbf{A}$ is not symmetric. Therefore, methods based on its eigenvalue decomposition must be modified. In an undirected network, the adjacency matrix can be decomposed as $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}$, in which $\mathbf{\Lambda}$ contains the real eigenvalues of $\mathbf{A}$. In undirected networks, this is not possible, and it is necessary to use either a non-orthogonal eigenvalue decomposition (leading to complex eigenvalues) or another matrix decomposition altogether.

Bipartite Network Datasets

A bipartite network is a network in which the set of nodes V can be partitioned into two sets V_1 and V_2 such that all edges connect a node in V_1 with a node in V_2. Social networks themselves are rarely bipartite. However, other networks extracted from social media are bipartite, for instance, user-item rating graphs or user-group inclusion graphs.

In bipartite networks, the clustering coefficient c is trivially equal to zero, since a bipartite network contains no triangles. Other network statistics and analyses must be extended to be used. In many cases, a statistic can be computed for the nodes in V_1 and V_2 separately. For instance, the average degrees of nodes in V_1 and V_2 can be defined. As another example, the largest connected component in a bipartite network contains a certain number of node from each of V_1 and V_2.

Network Datasets with Multiple Edges

In some social network datasets, multiple edges are allowed. An example is an email communication network, in which the nodes are the users and each edge represents a sent email. In these types of networks, also noted $G = (V, E)$, E does not represent a set but instead a multiset. Thus, two nodes, i and j, can be connected by multiple edges, for instance, denoting multiple emails that have been sent. Analogously, directed networks with multiple edges can be defined.

Most network statistics can be applied to networks with multiple edges without problem. For instance, the degree is defined as the number of edges adjacent to a vertex, counting multiple edges as such. The resulting degree distributions, as an example, can be tested for power laws.

When representing a social network with multiple edges as an adjacency matrix, the multiplicities are used as entries. In other words, the entry $\mathbf{A}_{ij}$ is defined to equal the number of edges between i and j, even when no edges connect the two nodes. The resulting adjacency matrix can be used in nearly all algebraic graph theoretical methods used for simple graphs.

Signed Network Datasets

Some social networks contain both positive and negative edges. An example are social networks with friendship and enmity links, such as the social network from the Slashdot technology news Web site, in which users can mark other users as *friends* and *foes* (Kunegis et al. 2009).

Mathematically, a network with positive and negative edges is modeled as a signed graph $G = (V, E, \sigma)$, in which σ is the sign function, mapping the edges in E to the set $\{-1, +1\}$.

The extension of social network statistics to signed graphs is not trivial. Using the example of the degree, each vertex of a undirected signed graph can be defined to have two degrees: the positive degree $d^+(i)$ counting the number of positive incident edges and the negative degree $d^-(i)$ counting the number of negative incident edges. Another way of defining degrees consists in subtracting the number of incident negative edges from the number of positive incident edges, giving the signed degree

$$d(i) = d^+(i) - d^-(i).$$

Analogous signed definitions can be given, for instance, for the clustering coefficient (Kunegis et al. 2009).

The adjacency matrix of signed graphs is typically defined as a $-1/0/+1$ matrix $\mathbf{A}$ defined as $\mathbf{A}_{ij} = \sigma(\{i, j\})$ when $\{i, j\}$ is an edge and $\mathbf{A}_{ij} = 0$ otherwise.

Rating Network Datasets

Rating network datasets are networks in which the edges represent ratings. As an example, if users of a dating site can rate other users, the resulting social network is a directed rating network between users. The much more common case however is that of bipartite rating networks, in which users rate items, for instance, movies (GroupLens Research 2006), songs (Yahoo! Labs 2011), or jokes (Goldberg et al. 2001). Ratings are typically numerical and given on a *rating scale*, the most common one ranging from 1 (dislike) to 5 (like).

To extend network statistics to rating networks, the ratings can be used as weights. However, care must be taken. In the example of the 1-to-5 rating scale, since the adjacency matrix is defined to contain the value zero for node pairs that are not connected, this would imply that a dislike of weight one counts as more than no rating of weight zero. Thus, it is typical to subtract the overall mean rating from all rating values and use the resulting numbers as weights in the adjacency matrix. Since the resulting matrix contains positive and negative values, a rating network can always be interpreted as a signed graph.

Temporal Network Datasets

A common type of study in social network analysis consists in observing the evolution of a network. In order to observe the evolution of a network, temporal information must be known. In the simplest case, edge arrival times are known for all edges, allowing one to reconstruct the network at any timepoint. All network statistics mentioned before in this article can be analyzed temporally, by computing them in function of time. The result can give insight into the processes of graph evolution. As an example, several network statistics which capture the notion of *diversity* of a network in different ways have been shown to decrease over time in a majority of social and other networks (Kunegis et al. 2012).

Multirelational Network Datasets

Signed and rating networks can be generalized to multirelational networks. In multirelational networks, any number of edge types are allowed (Greene and Cunningham 2009). For instance, edge types can be *friend*, *relative*, or *coworker*. Multirelational networks may be alternatively called *heterogeneous networks*.

Since the meaning of the edge types depends on the specific network, no simple generalization of network measures to multirelational networks is possible, beyond ignoring edge types. If the strength of each relationship type can be assessed, these values can be used as weights to compute the degree of nodes or as entries in the adjacency matrix. Another complication with multirelational networks are the structural properties of the various relationship types, which can vary. For instance, one relationship type can be directed, while another one is undirected.

Although multirelational social network datasets are available from various sources, only very few studies consider these types of networks generically. One example is given in Lippert et al. (2008).

Typed Networks

A further extension of multirelational networks are typed networks, in which in addition to multiple edge types, multiple node types are allowed. An example is given when a social network is

combined with a user–item rating network. Such a network contains users and items as nodes, and ratings and friendships as edges. Each edge type must thus connect two nodes of a given type. The edge and node types of a typed network can be summarized by an entity–relationship (ER) diagram. A bipartite network, for instance, can be modeled as a typed network in which the entity–relationship diagram consists of two nodes connected by a single edge. Typed networks are often used but seldom modeled as such. Examples of studies using typed networks are those combining social and collaborative recommenders (Adomavicius et al. 2005).

A further generalization of typed networks results in semantic networks, whose only constraints are that it consists of triples, and in which the fundamental difference between nodes and edges is removed at lower level and modeled as part of the network itself.

Practical Considerations

Several practical issue have to be dealt with when using social network datasets. First, a social network dataset that may be incomplete are biased due to the way it was aggregated. Then, legal considerations may be necessary for using datasets. Finally, varying data formats may affect usage.

Bias Due to Data Extraction

An ideal social network dataset is generated directly from the database of a social network company. Such a dataset is complete, and all statistics computed with it reflect that actual social relationships among the users. In practice however, most social network datasets are crawled by scientists from the social networking sites. Thus, they may be incomplete, corrupted, and reflect different parts of a social network at different timepoints. These biases can have a drastic effect in analyses performed on them. For instance, if degree distributions are studied in a social network where users with zero friends are excluded due to the way the data was crawled, the resulting average degree will be wrong. Other statistics however will not be affected, for instance, the diameter of the network.

Typical biases in social network datasets are the exclusion of nodes with small degree, the omission of everything except the largest connected component, and the fact that parts of the network were crawled at different times, resulting in a social network dataset that has never existed in that form at any timepoint.

Legal Considerations

Due to the sensitive nature of social networks, most social networking companies do not publish their datasets. Thus, datasets are usually crawled, putting the publication and usage of these datasets in a legal gray area. As an example of a large dataset of the Twitter social network which included user names was retracted from its Web site from the researcher that was involved, due to complaints from Twitter. Nevertheless, many social network datasets are available online, and many studies are performed on them. Well-known newly created social networks are crawled soon after they gain a sizable market share, as shown by the example of Google Plus (Schiöberg et al. 2012).

Data Formats

There is no unified data format for the publication of social network datasets. The formats that are used can be classified into those that try to be efficient, those that try to make it easy to combine the datasets with other datasets, and those that make it easy to access the dataset from a large number of programming languages and environments.

An example of an efficient format, both in terms of runtime and memory usage, is the binary format used by Boldi and Vigna (2004). An example of a format that makes it easy to combine a social network with other types of data is given by all social networks published as RDF. An example of social network datasets published in a format that is optimized for easy access from many programming languages is given by the tab separated value format used in KONECT (Kunegis 2013).

Key Applications

Applications of social network datasets are too numerous to cite and cover almost all aspects of data mining, information retrieval, recommender systems, Web science, and increasingly social sciences such as sociology.

Future Directions

New applications of social network datasets are published continuously. New network datasets are also published regularly. A trend in the recent years has been the aggregation of social network datasets into collections, for instance, in the Stanford Network Analysis Project (SNAP) (Leskovec 2010) and in the Koblenz Network Collection (KONECT) (Kunegis 2013). Another trend is the migration toward more interoperable formats, in line with the Link Open Data initiative.

Acknowledgments

All plots in this article are from the KONECT project (Kunegis 2013). The author of this work has received funding from the European Community's Seventh Frame Programme under grant agreement n° 257859, ROBUST.

Cross-References

- Linked Open Data
- Sources of Network Data
- Web Archives

References

Adomavicius G, Sankaranarayanan R, Sen S, Tuzhilin A (2005) Incorporating contextual information in recommender systems using a multidimensional approach. ACM Trans Inf Syst 23(1):103–145

Backstrom L, Boldi P, Rosa M, Ugander J, Vigna S (2012) Four degrees of separation. In: Proceedings of the web science conference, Evanston, pp 45–54

Boldi P, Vigna S (2004) The WebGraph framework I: compression techniques. In: Proceedings of the international world wide web conference, New York, pp 595–601

Fiedler M (1973) Algebraic connectivity of graphs. Czechoslov Math J 23(98):298–305

Gjoka M, Kurant M, Butts CT, Markopoulou A (2010) Walking in Facebook: a case study of unbiased sampling of OSNs. In: Proceedings of the conference on computer communications, San Diego, pp 2498–2506

Goldberg K, Roeder T, Gupta D, Perkins C (2001) Eigentaste: a constant time collaborative filtering algorithm. Inf Retr 4(2):133–151

Greene D, Cunningham P (2009) Multi-view clustering for mining heterogeneous social network data. Technical report, University College Dublin

GroupLens Research (2006) MovieLens data sets. http://www.grouplens.org/node/73

Kunegis J (2013) KONECT – the Koblenz Network Collection. konect.uni-koblenz.de

Kunegis J, Lommatzsch A, Bauckhage C (2009) The Slashdot Zoo: mining a social network with negative edges. In: Proceedings of the international world wide web conference, Madrid, pp 741–750. http://uni-koblenz.de/~kunegis/paper/kunegis-slashdot-zoo.pdf

Kunegis J, Sizov S, Schwagereit F, Fay D (2012) Diversity dynamics in online networks. In: Proceedings of the conference on hypertext and social media, Milwaukee, pp 255–264. http://userpages.uni-koblenz.de/~kunegis/paper/kunegis-diversity-dynamics-in-online-networks.pdf

Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? In: Proceedings of the international world wide web conference, Raleigh, pp 591–600

Leskovec J (2010) Stanford network analysis project. http://snap.stanford.edu/

Lippert C, Weber SH, Huang Y, Tresp V, Schubert M, Kriegel HP (2008) Relation prediction in multi-relational domains using matrix factorization. In: Workshop on structured input structure output, Vancouver

Opsahl T, Panzarasa P (2009) Clustering in weighted networks. Soc Netw 31(2):155–163

Read KE (1954) Cultures of the Central Highlands, New Guinea. Southwest J Anthropol 10(1):1–43

Schiöberg D, Schneider F, Schiöberg H, Schmid S, Uhlig S, Feldmann A (2012) Tracing the birth of an OSN: social graph and profile analysis in Google+. In: Proceedings of the web science conference, Evanston

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(1): 440–442

Yahoo! Labs (2011) KDD Cup from Yahoo! Labs. http://kddcup.yahoo.com/

Social Network History

▸ Networks at Harvard University Sociology

Social Network Mining

▸ Research Designs for Social Network Analysis

Social Network Privacy

▸ Anonymization and De-anonymization of Social Network Data
▸ Privacy in Social Networks, Current and Future Research Trends on

Social Network Randomization

▸ Privacy in Social Networks, Current and Future Research Trends on

Social Network Representation

▸ Socio-Graph Representations, Concepts, Data, and Analysis

Social Network Sites

▸ Privacy and Disclosure in a Social Networking Community
▸ Social Networking on the World Wide Web

Social Networking

▸ Social Networking in Political Campaigns

Social Networking for Open Innovation

Milan Stankovic, Saman Musacchio, and Philippe Laublet
Université Paris-Sorbonne, Hypios, Paris, France

Synonyms

Innovation crowdsourcing platforms; Open innovation social networks; Web problem-solving platforms

Glossary

Open Innovation A paradigm stating that companies can and should use both external and internal ideas to boost their innovation. This includes both "outside-in" (calling on external knowledge for use internally) and "inside-out" (when unused or under-performing internal knowledge is promoted outside company walls) approaches. The term has been attributed to Henry Chesbrough, professor at the University of California, Berkeley (USA)

Problem Solving on the Web Using Web user connectivity to collaborate (e.g., the 2009 Polymath Project) or answer open-problem challenges as individuals (e.g., P&G Connect, Innocentive, Hypios)

Crowdsourcing An approach that involves outsourcing tasks to a distributed group of people, both online or offline. This can involve the mass collaboration of thousands of individuals to accomplish one overall task (e.g., the Galaxy Zoo project, which recruited over 200,000 online volunteers to classify galaxies), or the competition of thousands of individuals in an open call for solutions (e.g., the X-Prize, Innocentive, Hypios)

Social Networking Engaging in social activities online involving but not solely based on connected platforms like Twitter, Facebook, LinkedIn, and Pinterest

Semantic Web A Web of interconnected self-describing data structures interpretable by machines. It represents an extension of the Web of connected pages, in which data resources are connected among each other with typed links thus forming a giant typed graph

Serendipity The discovery of relevant but unintended and unexpected facts, phenomena, and ways of thinking

Definition

Social networking for Open Innovation is the practice of using virtual social networks to identify and engage with participants, often external to a company, as part of a larger process to develop innovative solutions and products or acquire R&D.

The current methodology of online Open Innovation problem-solving platforms involves broadcasting problems to either undefined (e.g., the Web) or very specific communities (e.g., brand consumers, specific solver communities). Though this "push-out" method does produce results, it is found to generate excessive noise and limit the involvement of certain users.

However, coupling data collected on social networks (user profiles, comments) with various articles by the same users (publications, resumes) can allow the "crowdsourcer" – OI platform or company – to create an ad hoc global virtual community to address specific issues. This reduces noise, since only relevant solution providers will be identified, and increases resolution probability, as these individuals have not opted into specific communities and are wide ranging (in both areas of interest and geography).

Furthermore, and contrary to traditional "push-out" methods, these solution providers are personally contacted (via social networks, email, and sometimes phone calls) by a dedicated team to generate interest for the problem that needs to be resolved. Using this "pull-in" method, the overall success rate of problem resolution has increased significantly.

Introduction

It was in 2009 that Fields medalist Tim Gowers decided to use his blog to find a new combinatorial proof to the density version of the Hales-Jewett theorem – in other words, to solve a very complex mathematical problem. More of social experiment on his part, he decided to put the question out in the open and see how long it would take experts, collaborating online, to crack it. The Polymath project1, as it is now called, solved the problem in 37 days, with over 800 contributions from 27 people. It even led to two papers published under the name D.H.J. Polymath. It is a true example of collaborative innovation powered by online social networks.

Collaboration is in our DNA. From the study of animal groups to Georg Simmel's extensive research on social geometry, it is clear that social networks, from family units to tribes, to the nineteenth century's great urban centers, have been critical to all cultural, social, or scientific advancement. These networks are now more powerful than ever through the new tools of interconnectivity offered by the Web of the twenty-first century.

And these tools cater to every aspect of collaboration, from universal user-generated encyclopedias (Wikipedia) to sharing documents both outside (Google Docs) and inside (Sharepoint) company walls. Harnessing the power of effective online collaboration through blogs, forums, community networks, open problem-solving platforms, or social platforms is still a challenge today, even more so for large organizations, whose internal tools lack the connectivity enjoyed by their employees outside company walls. In fact, the flexibility, speed, and efficiency of temporary online collaborations – even those found in multiplayer games – are forcing companies to question internal processes and adapt from the outside in.

And the crux of the problem is the rapid identification of groups or individuals who are best qualified to solve any given challenge. In this paper we argue that a truly Open Innovation approach of using external social networks as basis for such ad hoc groups is key to successful

problem resolution. Yet to create the most efficient groups, existing social networks must be broken down to eliminate both homophily and propinquity. Adequate distance from the subject matter, bridges, and weak ties are highly important in promoting serendipity and guarantee that novel and often surprising solutions are submitted or that transferable processes already applied in different disciplines are quickly located.

We argue that the Semantic Web is best able to help discover the most relevant keywords for identifying such individuals on existing social networks, and from traces (papers, resumes, etc.) found online. The group can then be asked to collaborate or answer an open call for solution. This can be directly handled by the company or outsourced to a dedicated problem-solving platform.

Key Points

Using the Web for Open Innovation:

- Speeds up the process of innovation
- Gives access to external know-how
- Leads to unexpected results and solutions coming from unexpected domains
- Reduces the cost of acquisition of research
- Reduces the need for managing multiple contacts with several academic actors and gives one point of access to research

Historical Background

In the increasingly competitive market that characterizes the world economy today, the need to develop innovations quickly has become a Holy Grail for many companies. The Open Innovation (OI) model emerged as a response to the limitations of traditional innovation models, involving mainly internal research departments siloed in their respective areas of expertise. The traditional model was perceived as unsatisfactory mostly in terms of efficiency and heterogeneity of solutions considered. For Henry Chesbrough, who introduced the term in 2003, "open innovation is the use of purposive inflows and outflows of knowledge to accelerate internal innovation, and expand the markets for external use of innovation, respectively. The open innovation paradigm assumes that firms can and should use external ideas as well as internal ideas, and internal and external paths to market, as they look to advance their technology" (Chesbrough et al. 2006). According to existing literature, three key processes of the Open Innovation can be differentiated (Enkel et al. 2009): "outside-in" (the use of external resources), "inside-out" (the realization of profits from the commercialization of sleeping patents), and the "coupled" (co-creation with partners). This paper focuses on the outside-in processes completed by Open Innovation platforms.

Although these three classifications are perfectly in line with current company models, the idea that Open Innovation in its simplest form did not exist prior to 2003 is a fallacy. Recent literature argues that "open" practices have been applied before in companies in different ways (Trott and Hartmann 2009). Furthermore, open calls for solutions, for the benefit of organizations and governments, have been common throughout history, from the Longitude prize (1714) to the invention of canned foods (Napoleon's Food Preservation Prize 1795) or the development of submarines (The Confederate Prize for Inventions that Sink or Destroy Union Ships 1861).

The successful application of Open Innovation practices has been well documented by companies like IBM, P&G, Intel, Cisco Systems, DuPont, Lucent, or Philips (Sari et al. 2007). What follows are a few examples of how one open innovation practice in particular, crowdsourcing, is adopted by some of today's leading firms.

1. **Branded Platforms with Corporate Needs**
   Probably the most famous example of such a platform is Procter & Gamble's (P&G) Connect & Develop site. The company formulates specific needs and posts them on its website in order to invite innovators and researchers to submit potential solutions. One challenger posted on P&G's platform was the need to develop a lipstick that would glow for 4 h, much longer than today's standard lipsticks.

While this is not a problem that anyone in cosmetics would find surprising, the company with such a solution would gain an essential edge over its competitors.

2. **Corporate Communities for Discussion of Questions and Needs**
   Another approach is the one taken by Clorox with Clorox Connect: building a platform where innovators and researchers can sign up to discuss issues with employees of the company, in a forum led by a corporate community manager. The downside to this type of platform is that it competes with specialized science wikis, forums, or specialized social networks like Research Gate, Academia, or university intranets.
3. **Communities for Customer Co-Creation**
   As opposed to (2), this type of platform addresses customers or fans of a company's products. An example of this was Lego Mindstorms, which during its lifetime led to the commercialization of several products. While this kind of initiative tends to lead to highly motivated and (nearly) self-driven communities, it is not applicable to every company.
4. **Corporate Idea Boxes**
   Shell's Gamechanger exemplifies this type of platform. While it is rather successful, all ideas that have led to products and process improvements have come from inside the company.
5. **Platforms That Centralize Open Problems**
   As opposed to corporate platforms, websites like Innocentive or Hypios list problems from a number of companies. This attracts individuals who are generally interested in solving problems, wherever they may arise. For a researcher who wants to maximize the chances of finding a problem that he/she can solve, such platforms are highly attractive. These emerging open innovation platforms are trying to leverage Web technology and most notably its social aspects to help innovation occur faster and more efficiently. Those services rely on social networks to diffuse innovation challenges, engage with experts, and boost collaboration.

Web Technologies for Open Innovation

Hypios, a Web-Based Marketplace for Solutions

Hypios, a French solution marketplace launched in 2009, best exemplifies the latest methods available for Web-based innovation. Companies with R&D problems (called seekers) use Hypios to externalize their problems to an ad hoc group of experts (called solvers), who then submit novel and often unexpected solutions. Karim Lakhani, HBS professor and leading academic expert on the subject, calls this method "problem-broadcast." R&D departments usually have expertise in a specific area and approach problems from a certain perspective. Yet it is evident that across the world – thus somewhere on the Web – there are people with different perspectives who can approach the problems differently and suggest truly novel solutions. The goal of a marketplace for solutions is to ensure that R&D problems reach the right people on the Web. One of the initial observations made was that companies constantly reinvented the wheel, simply because they didn't know where to look for existing plans for wheels – or because they were too scared that their competitors could find out that they were working on the wheel. Yet the truth is that most of their competitors are also working on the wheel. In the words of Kevin McFarthing, who implemented Open Innovation at Reckitt Benckiser: "R&D problems that would surprise your competitors are very rare."

The people-centric approach of Hypios makes it possible to identify explicit solutions (e.g., in publications or patents) as well as "incorporated solutions," ones that have not been made public but that can be provided by individuals if you ask them. The ability to find such "sticky" and implicit knowledge is a key advantage of identifying people rather than existing explicit solutions.

Semantic Web Technologies for Open Innovation on the Web

Although there have been studies on online search (Parkes 2007), the work done in relation to the use of social networks and new technologies

by innovation intermediaries are far and between. What is of particular interest to our research is that the Social Web and Semantic Web are powerful tools that can be used for building and maintaining relationships of dispersed social communities and thus create and expand networks to produce synergies through combined interactions of users (Breslin et al. 2009). While the Social Web has already been introduced, the Semantic Web, on the other hand, should be clearly defined.

The Semantic Web is an extension of the current Web in which any content is tagged with a more precise meaning to enable machines to process it and thus better answer human queries.

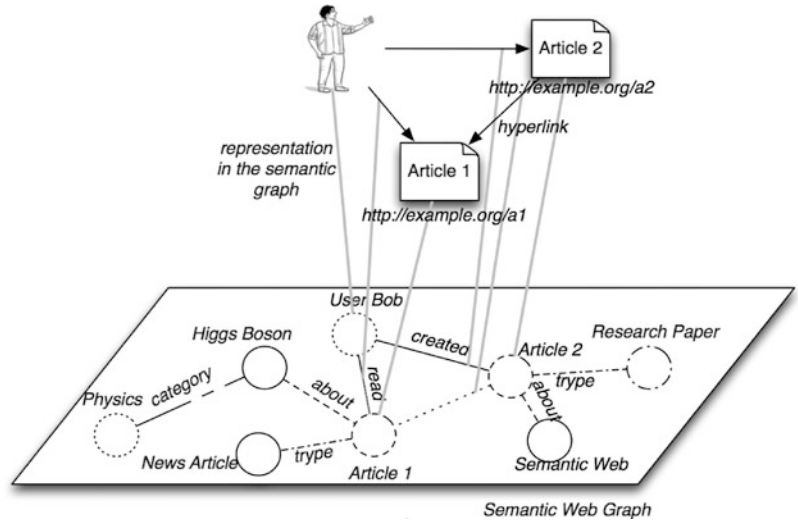
In addition to the plain textual, visual, and additive content that dominated the initial Web, the Semantic Web creates structured, self-describing articles that are published and interconnected on the Web. In this section, we will discuss how Semantic Web technologies, as a complement to the Social Web, can be used to enhance the problem-solving processes of Open Innovation platforms.

Expert Identification

The possibility of using the Web as a source for identifying experts has already been explored in literature. Web resources that users create or interact with have been used to assess expertise for such tasks as human resource management and finding assistance in e-learning scenarios.

Recently, a new trend has emerged in regard to how data is published on the Web: Linked Data (Bizer et al. 2009), which is now seen as an integral part of the Semantic Web. In contrast to representing data in the form of regular Web pages, Linked Data publishes information in a more structured format with a semantic overlay. Linked Data publishing is increasingly widespread, and even large data providers like Facebook are turning to specific forms of semantic data representation standards (http://developers.facebook.com/docs/opengraph/). Several possibilities of currently available user data in Linked Data form have already been explored to identify experts (Stankovic et al. 2010), and further improvements will be made once data publishers accept richer forms of expressing expertise-related data (Aleman-Meza et al. 2007). The benefits of the Linked Data formats are primarily in the rich structure of typed nodes and links. When user activities, such as interaction with Web content, are represented as Linked Data, the result is a rich structure of traces, which clearly identify the users' interests and knowledge. In the example represented in Fig. 1, we can see a user interacting with two articles on the Web. The user read Article 1 and created Article 2. Modern Social Web applications using Semantic Web standards would store and selectively publish data about those activities in the form of a Semantic Web Graph. In this graph, a node representing a user would be connected to nodes representing Web content, which are further connected to nodes representing topics of interest. The relationship of the user with the content would determine the strength and the nature of his relationship with the topics of the content. Given the diversity of those links, it is possible to weigh the importance of certain topics for a given user differently in different situations. For instance, in our Fig. 1, when searching for user qualifications, we would consider the topic "Semantic Web" more important, because the user created content on this topic, and when searching for his interests, we would likely pick "Higgs Boson," because the user read content on this topic. In our previous research, we have constructed a system Hy.SemEx (Stankovic et al. 2011) that relies on the diversity of link types to deliver a better expert identification engine, adapted to different needs and different situations.

Identifying experts for Open Innovation practices (especially when dealing with the Hypios platform) is different than for simple HR needs or similar queries. For OI, it is essential to find potential problem solvers that are not necessarily the best-ranked experts, with rich expertise in the given problem area (Jeppesen and Lakhani 2009). It is therefore especially important to adapt the way experts are selected in order to create a broad base of individuals with relevant, yet distant areas of expertise from the context of a specific innovation problem.



Social Networking for Open Innovation, Fig. 1 User activities and their traces in the Semantic Web Graph

This is how Semantic Web-based expert identification technology can be used to reach the outside solvers, including the ones on the margin of the area in question, and to encourage the transfer of knowledge between those fields. In doing so, it enables "cross-sectorial problem solving." Likewise, by identifying experts in peripheral fields, this technology helps determine the graph of social behavior between relevant solvers on the Web and thus identifies "weak ties."

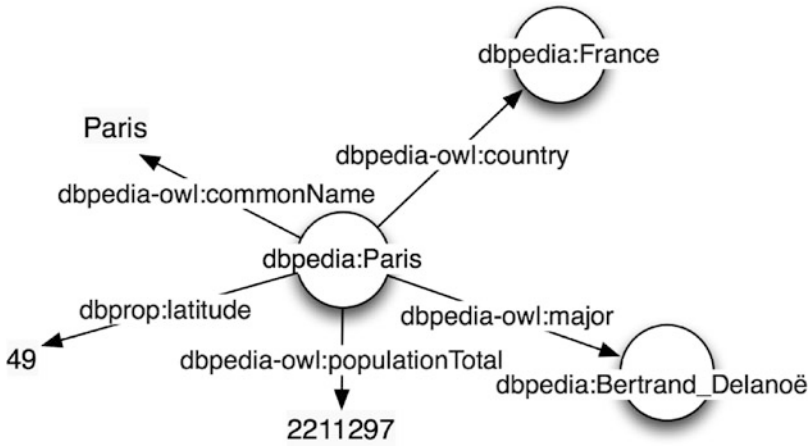
Semantic Keyword Discovery

Semantic keyword discovery extends the standard matching of documents by keywords, with a notion of semantic proximity of keywords. By going beyond exact matches, it enlarges the space of possibilities. This particular property of semantic keyword matching fills a real need in Open Innovation models.

Different communities use different words to express the same or similar concepts. Thus coming from one community of practice and using one's own words to express an innovation problem heavily limits its reach in different areas. Present technologies exist to find synonyms and words of similar meaning, based on taxonomies of concepts (Ziegler et al. 2006) and word co-occurrence (Cilibrasi and Vitanyi 2007). Such existing approaches have limitations, however, as they focus on providing relevant suggestions and often neglect the need for serendipity and discovery that are essential to OI scenarios. Novel approaches that use Linked Data sources, such as DBPedia.org, to make meaningful connections between concepts in the area of music (Passant 2010) and enable the discovery of unexpected, but relevant concepts give hope that such sources might also serve to establish a notion of semantic proximity of concepts that would be more open to serendipity.

When using a problem description to identify profiles of potential experts, semantic matching makes it possible to find experts who not only work in the exact discipline of the problem but also in areas that are semantically relevant. One of the primary motivations for using a broader matching approach is to decontextualize the problem from its context and thus from the language prevalent in a particular area of expertise and increase the diversity of submitted solutions promised by a truly OI process.

For example, let us imagine an innovation problem related to detecting cable joints underground. Solving the problem requires expertise that would allow one to construct a device capable of precisely detecting cable joins by scanning the surface of the ground. A standard approach for finding experts capable of solving this problem would be to consider different keywords



Social Networking for Open Innovation, Fig. 2
A part of DBpedia graph

related to cables, electricity, and metal detection. However, a semantic approach to keyword discovery might provide less expected keywords, for instance, those related to the bone and vascular joints in the body. Those topics might seem unexpected, but in fact experts working on medical scanning equipment might have knowledge that can be transferred to the cable joint problem.

In order to deliver the functionality of semantic keyword discovery for its problem-solving platform, Hypios has developed a unique solution based on Semantic Web data structures, called Hy.Proximity (Stankovic et al. 2011). Hy.Proximity uses DBpedia, a semantic version of Wikipedia, to find concepts related to a number of initial concepts of interests. DBpedia is composed of a rich structure of concepts and their typed links. In a small part of the semantic graph featured in Fig. 2, the concept "Paris" is connected to the concept "France" with the link of type "country." When calculating the proximity score of two nodes, Hy.Proximity takes advantage of the different types of links that connect them and gives them different weights. In addition to structural futures of the graph, the semantic nature of links and nodes open numerous possibilities for constructing fine-tuned recommender systems. Hy.Proximity exploits those possibilities to deliver concept recommendations that are both relevant and unexpected for the user, encouraging the discovery of then unknown, but relevant knowledge.

We have compared the performance of our system against state-of-the-art keyword recommendation approaches. While an exhaustive evaluation has already been made (Damljanovic et al. 2012), we present here a couple of results and examples to illustrate the usefulness of the Semantic Web-based approach. For instance, we compared our system, Hy.Proximity, with the AdWords tool (https://adwords.google.com/o/KeywordTool) for keyword suggestion used to help advertisers better design their online promotional campaigns. A main difference of the Semantic Web-based approach used by Hy.Proximity is that AdWords applies a statistical approach, looking for words that co-occur in search queries and Web documents. In our evaluations, the two systems performed similarly on relevance of their proposed keywords, but in terms of unexpectedness of relevant suggestions, Hy.Proximity outperformed AdWords.

To illustrate the usefulness of suggestions that Hy.Proximity provides, what follows is an example of keyword suggestions obtained from our system and from Google AdWords. We have run both systems to obtain keyword suggestions that would help us advertise an innovation problem to an audience of experts, potential problem solvers. The problem in question deals with Kaolin extraction and issues with current mining techniques. The initial keywords we used for suggestions were Kaolinite, Drying,

Social Networking for Open Innovation, Table 1 Keyword suggestions obtained from Hy.Proximity and AdWords

Hy.Proximity	AdWords
Drying	Dry eyes
Induced gas flotation	Dry cleaners
Souders-Brown equation	Dry shampoo
API oil-water separator	Chem dry
Dissolved air flotation	Dry tortugas
Froth flotation	Dry scalp
Aqueous two-phase system	Flights to Brazil
Gas separation	Text mining
Adduct purification	Dry-erase board
Liquid-liquid extraction	Mining companies
Acid-base extraction	Filter press
Spinning cone	Dry rot
Vapor-liquid separator	Cheap flights to Brazil
Settling	Brazil holidays
Flotation process	Brazil travel
Sublimation apparatus	Mining jobs Australia
Filter paper	Salvador, Brazil
Azeotrope	Dry ice blasting
Supercritical fluid extraction	Dry eye syndrome
Fluid extract	Dry suit

Mining, Separation process, Settling, Filter press, Brazil, Mill (grinding), Tailings, and Redox. The suggestions provided by the two systems are given in Table 1. While the keyword obtained through the treatment of the semantic DBpedia graph (Hy.Proximity) concerns specific topics likely to be used by experts and includes diverse topics (some topics related to mining and some related to similar processes used in other industries, such as filtering using filter paper), the topics provided by Google AdWords represent combinations of terms often used together on the Web. Their utility is more in reminding the user of known notions than in enriching him with unknown concepts. The Semantic Web structures thus play an important role in opening the audience to unexpected disciplines from which knowledge transfer can be expected – a key feature of the Open Innovation approach.

Key Applications

Using a semantic graph for identifying relevant, unexpected, or distant links can have applications in every industry – from advertising to recommendation algorithms. Yet our interest here is how useful this technology can be for solving complex innovation and R&D problems and thus accelerating research and reducing time to market.

The main problem with existing OI open problem-solving platforms is that the target group of solvers has either too much heterogeneity or too much homophily. In the first case, broadcasting a problem to random solvers will not necessarily ensure the positive and rapid resolution of a problem. Furthermore, no matter the group size, it will be limited to solvers who have opted into one of these platforms. The second case exemplifies what already plagues internal research departments. Broadcasting a problem in airplane aerodynamics to experts in the field will seldom lead to novel, unexpected solutions and discourage any technological cross-pollination.

A semantic approach for identifying experts, coupled with an outreach process, solves both problems at once. By using relevant keywords to identify experts, an ad hoc group of solvers – a network – can be created from profiles anywhere in the world, in real time, and for every specific problem. This group will have enough heterogeneity – *relevant* heterogeneity – to offer novel solutions, limit noise, and encourage the "systematic serendipity" of ideas.

Future Directions

Social Web technologies for Open Innovation have so far mostly addressed the field of open problem solving. However, the field of Open Innovation is much wider. Different actors in the OI world still remain to be connected and their collaboration facilitated by Social Web technologies. For instance, apart from the need to connect with problem solvers, companies

that adopt OI approaches also need to connect to peer companies with whom they could codevelop innovative products. Furthermore, there is also a need for better interactions within their ecosystem of suppliers, consultants, and partners. Social networks can still play a role to encourage a paradigm shift at this level.

Social networks may also prove critical in the design of novel ways of engaging with consumers and guiding the companies' innovation towards impulses coming from social networks. Many brands already maintain an online presence via social networks to promote their products and control their image. A stronger connection between innovation research and consumer content on these social networks could prove useful in the future, as the use of social networks by customers becomes increasingly ubiquitous.

Acknowledgments

The work of Milan Stankovic has been partially funded by ANRT (French National Agency for Research and Technology) under the grant number CIFRE N 789/2009.

Cross-References

- ▸ Collective Intelligence for Crowdsourcing and Community Q&A
- ▸ Creating a Space for Collective Problem-Solving
- ▸ Linked Open Data
- ▸ Online Communities
- ▸ R&D Networks

References

Aleman-Meza B, Bojars U, Boley H, Breslin JG, Mochol M, Polleres A, et al (2007) Combining RDF vocabularies for expert finding. Lecture notes in computer science, vol 4519. Springer, p 235. Retrieved from http://www.springerlink.com/index/p6u10781711xp102.pdf. Accessed 1 Feb 2013

Bizer C, Heath T, Berners-lee T (2009) Linked data – the story so far. Int J Semant Web Inf Syst 5:1–22. (Special issue on Linked data)

Breslin J, Passant A, Decker S (2009) The social semantic web. Springer, Heidelberg

Chesbrough HW, Vanhaverbeke W, West J (eds) (2006) Open innovation: researching a new paradigm. Oxford University Press, Oxford

Cilibrasi RL, Vitanyi PMB (2007) The Google similarity distance. IEEE Trans Knowl Data Eng 19(3):370–383. doi:10.1109/TKDE.2007.48

Damljanovic D, Stankovic M, Laublet P (2012) Linked data-based concept recommendation? Comparison of different methods in open innovation scenario. In: Proceedings of extended semantic web conference (ESWC 2012), Heraklion

Enkel E, Gassmann O, Chesbrough H (2009) Open R&D and open innovation: exploring the phenomenon. R&D Manage 39(4):311–316

Jeppesen LB, Lakhani KR (2009) Marginality and problem solving effectiveness in broadcast research. Organ Sci 20. Retrieved from http://dash.harvard.edu/handle/1/3351241. Accessed 1 Feb 2013

Parkes DC (2007) Online mechanisms. In: Nisan N, Toughgarden T, Tardos E, Vijay VV (eds) Algorithmic game theory. Cambridge University Press, Cambridge, pp 411–439

Passant A (2010) dbrec – music recommendations using DBpedia. In: Patel-Schneider PF, Pan Y, Hitzler P, Mika P, Zhang L, Pan JZ, et al (eds) Proceedings of the 9th international semantic web conference – ISWC 2010, Shanghai, vol 1380. Springer, pp 1–16

Sari V, Pekka S, Marko T (2007) Implementation of open innovation paradigm cases: Cisco systems, DuPont, IBM, Intel, Lucent, P&G, Philips and Sun Microsystems. Research report 978-952-214-478-2, 189, Lappeenranta

Stankovic M, Wagner C, Jovanovic J, Laublet P (2010) Looking for experts? What can linked data do for you? In: Proceedings of linked data on the web 2010, on WWW 2010, Raleigh. Retrieved from http://events.linkeddata.org/ldow2010/papers/ldow2010_paper19.pdf. Accessed 1 Feb 2013

Stankovic M, Breitfuss W, Laublet P (2011) Discovering relevant topics using DBPedia. In: Proceedings of the web intelligence conference (WI2011), Lyon

Stankovic M, Jovanovic J, Laublet P (2011) Linked data metrics for flexible expert search on the open web. In: Proceedings of 8th extended semantic web conference (ESWC 2011), Heraklion. Springer

Trott P, Hartmann D (2009) Why 'open innovation' is old wine in new bottles. Int J Innov Manage 3(4):715–736

Ziegler C-N, Simon K, Lausen G (2006) Automatic computation of semantic proximity using taxonomic knowledge categories and subject descriptors. In: Proceedings of the 15th ACM international conference on information and knowledge management, CIKM'06, Arlington. ACM, New York, pp 465–474. Retrieved from http://doi.acm.org/10.1145/1183614.1183682

Social Networking in Political Campaigns

James D. Ponder[1], Paul Haridakis[1], and Gary Hanson[2]
[1]School of Communication Studies, Kent State University, Kent, OH, USA
[2]School of Journalism & Mass Communication, Kent State University, Kent, OH, USA

Synonyms

Facebook; Internet; MySpace; Online; Social media; Social network; Social networking; Twitter; YouTube

Glossary

SNS Social networking site

SNS and Politics

In their short history on the political scene, social networking sites (SNS) have had a dramatic impact on how political campaigns function. For instance, in 2004, US Democratic presidential hopeful Howard Dean used a diverse network of bloggers and donors to rise from a relative unknown to a front-runner for the nomination in only a few months. In fact, the Dean campaign was hailed by political and media scholars as the first digital campaign (Hindman 2005). Dean's willingness to relinquish control over his campaign empowered Internet opinion leaders to support and strengthen his campaign. Still, the Dean campaign presented an enigma to political scholars as his vast success in the early stages of the Democratic primary failed to result in the Democratic nomination.

In 2008, Barack Obama used SNS to develop a grassroots effort that raised over $750 million in campaign funds (Bradley 2008) and organized over 8 million volunteers (Smith 2008) on the way to becoming the 44th president of the United States, breaking records for fundraising and volunteers along the way. The sheer amount of money raised by the Obama Campaign more than doubled previous fundraising efforts. This was especially surprising given the vast amounts of donations (6 million of the 6.5 million total donations) were less than $100 (Vargas 2008).

In their short history on the political scene, SNS also have had an impact on how people acquire and share political information with each other. Prior to the twenty-first century people relied ostensibly on the mass media for political information. Moreover, the flow of this information typically moved from political elites such as the news media and political parties/candidates to the general population (Haridakis and Hanson 2009). However, the rise of SNS has provided new outlets for information to flow in multiple directions. Most notably, it has enabled audience members to share and disseminate information with professionals and amongst themselves. With such changes, the Internet has become an important gateway connecting users to the larger world.

Although politics is not the driving force behind the use of SNS, people do use their social networking pages for politics and to watch political videos online. Candidates also use social media to bypass mainstream media and reach voters directly, and YouTube has become a major source of campaign videos and other politically related fare (for a more in-depth explanation, see Haridakis and Hanson 2009). In fact, during an average month during the 2008 US presidential campaign, more than 81 million unique viewers used SNS for political information and watched politically related YouTube videos (Ramirez 2008). SNS also provide people with additional avenues for exchanging information. People with access to SNS are not limited to more one-way directional mass communication for political information. They are not limited to face-to-face discussion with those with whom they have strong ties and weak ties. They can blog, tweet, text message, tag videos to share with others, and/or become Facebook friends with similar others.

People have also used SNS to institute political changes in and around the globe, serving as a primary means to connect people and coordinate their efforts in the uprisings and revolutions in Algeria, Egypt, Syria, Tunisia, Libya, and Yemen (Jowett and O'Donnell 2012). With such a vast array of applications and outcomes in the political arena, SNS have become a central component of the political process, serving as venues for candidates to connect with their constituents as well as places for citizens to engage in political activities. In this entry, we discuss the role of SNS in politics. While the use of social media is a global phenomenon, in this entry we use examples largely from the United States, where major social networks such as Facebook and Twitter first emerged. The focus will be on three applications of SNS: the use in campaigns, the use by and effects on voters during political campaigns, and the use by citizens to effectuate social and political change.

Social Networks and Political Campaigns

The role of social networks and use of media for political information in political campaigns has been an area of interest for many years. For example, in a study of the 1940 US presidential election, Lazarsfeld et al. (1994) set out to study how and why voters made up their mind about the candidate for whom to vote and the information sources they used to do so. They found that people in their social networks were more influential sources of information than were the media. They termed certain influential interpersonal sources as opinion leaders. These opinion leaders were people who tended to obtain information directly from the media and then shared it with others in their interpersonal networks. This social networking process was termed the "two-step flow of communication." These findings led mass communication researchers to give greater weight to social ties among people in the political communication process.

In later years, researchers found that the flow of communication within political and other social systems was more complex than a simple "two-step" flow. Later investigations expanded on the complexity of the flow and diffusion of news and political information and the role of the media and interpersonal communication in the diffusion of information and ideas. We have come to understand the media are particularly effective at getting information to people, but real attitudinal and behavioral changes occur through interpersonal and group influence. The advent of SNS has provided new channels for effectuating that interpersonal and social connection, thereby functioning not only as a viable venue for attitude and behavior influence but also as a place where people can easily access information.

Other investigations have considered the role of social networks in non-mediated settings. These early explanations of the relationship between membership in social networks and political involvement include arguments that the membership stimulates a collective interest in politics (Schlozman et al. 1995), makes people available to elites for mobilization (Leighley 1996), and helps people learn skills that make participation easier (Schlozman et al. 1995). More recently, scholars have found that social networks are a rapid way to disseminate innovative information and values in a society (Gibson 2001).

One of the key tenets of these investigations into social networks is that social interaction exposes people to a different set of politically relevant information and stimuli than they possess individually. Since individual understanding, information, resources, and ability are inherently limited, this means that social interaction provides people with other opportunities to accumulate resources, such as information, that lower the barriers to political participation. Consequently, participation in social networks supplement (rather than supplant) the person's resources and abilities that make participation likely (McClurg 2003). McClurg (2003) found that social interaction has a twofold influence on likelihood to participate in politics (i.e., vote). First, he found that social interaction in these networks exerts a positive and statistically precise effect on participation, but only when it is politically relevant. Second, this effect exists even after controlling for membership in organized groups, which indicated that formal and informal social

interaction have theoretically distinct effects on involvement. While we do have a rather thorough understanding of non-mediated social networks, we have yet to fully understand how online social networks influence the voting public. So how does the advent of the Internet and the different structures inherent in it influence these social networks?

Online Social Networks

The advent and growth of the Internet enhanced the ability of people to maintain and expand their social ties with others. For example, as early as the 1980s, before the Internet was widely diffused, interactive communities emerged on the Internet. Some, like the California city of Santa Monica's PEN project, were designed for interactive exchange among citizens and between citizens and city officials about issues including political matters. Boyd and Ellison (2007) defined social network sites (SNS) as

> Web-based services that allow individuals to (1) construct a public or semipublic profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system. The nature and nomenclature of these connections may vary from site to site (p. 211).

Not surprisingly, this definition covers a broad array of various online sources (for a more in-depth discussion of the different types of SNS, see Boyd and Ellison 2007) including Facebook, MySpace, LinkedIn, and even YouTube. In terms of politics, these sites provide users with a variety of different options. Individuals can form various groups that support particular candidates or issues, seek out political information, engage in online discussions with others about issues or candidates, blog about political issues, and even share videos (Boyd 2008; Brown et al. 2007; Xenos and Foot 2008).

Even prior to growth in popularity of SNS such as YouTube and Facebook, social media such as blogs had become important sources of political information (Sweetser and Kaid 2008). However, the effects of blogs and the Internet generally have been debated. One major debate has been whether people use online communities such as those fostered by SNS to become exposed to new ideas, change attitudes and beliefs, or reinforce existing attitudes and beliefs (for a more in-depth discussion, see Boyd 2008). The answers to such questions probably are contingent on how SNS are used. SNS can be used by small groups of homogeneous people. They also can be used for mass communication to reach large groups of people. Regardless of how they are used, SNS provide users the ability to generate content, share content, and serve as a portal to a variety of print and video sources.

Online social networks also have been shown to have a consistent, prosocial impact on individuals in terms of politics. For instance, Brown et al. (2007) found that individuals who participate in online social environments such as SNS are likely to experience a sense of understanding, connection, involvement, and interaction with others who participate in these environments. Also, individuals who belong to online civic-political groups report higher levels of civic participation, both online and offline, as a result of that participation in online civic-political groups (Kavanaugh et al. 2007). While most individuals engage in dialogues with homogenous others (Boyd 2008), these interactions have been shown to increase civic engagement (Xenos and Foot 2008). Xenos and Foot (2008) suggested that the unique aspects of online social networks allow individuals a communicative, creative, and social freedom to explore their position on a variety of political issues that appeals to younger adults. Therefore, SNS not only provide opportunities for individuals to seek out information and engage in meaningful conversations about political issues, but these sites also appeal to a younger population.

Social Networking Sites and Political Activity

Previous researchers have shown that people who use SNS for information about the candidates or to discuss politics are more likely to engage in civic and political activities online such as signing petitions (Abroms and Lefebvre 2009)

and donating money (Baumgartner and Morris 2010; Valenzuela et al. 2009). Moreover, the current evidence suggests these online activities can lead offline activities such as volunteering for a candidate (Abroms and Lefebvre 2009), talking with others about candidate preferences, and voting (Valenzuela et al. 2009). For instance, Valenzuela et al. (2009) found a positive relationship between the intensity of SNS use and students' life satisfaction, social trust, civic engagement, and political participation. However, the contribution of SNS was relatively small and this led the authors to suggest that online social networks are not the most effective solution for youth disengagement from civic duty and democracy. They did suggest that participation in Facebook groups tended to increase political participation more than participation in other aspects of Facebook (e.g., blogging, friending political candidates). Conversely, other scholars have found that this relationship is inconsistent at best (for a more in-depth explanation see Baumgartner and Morris 2010). It is important to note that while there is evidence suggesting a link between using SNS and political activity, there has yet to be a study that tries to create a causal linkage between SNS use and subsequent political activity.

While the Internet and SNS can enhance participation, whether greater political participation is necessarily healthy for a democracy or for political stability is debatable. For example, some have argued that highly active population acting through targeted social networks can be disruptive to established order and can advance multiple and sometimes conflicting political agendas (Jowett and O'Donnell 2012). Internet and SNS use can allow for greater fragmentation. There also remains a digital divide. Some have argued that the Internet and SNS can widen the gap not just between those who have access or not but also between those who are more politically active and those who are less active, because the former are the ones most likely to take advantage of these technologies. However, it is important to note that while SNS may have this effect, TV remains, at least in the United States, the most used medium for political information (Smith 2011). Thus, the potential of SNS for robust public participation and dialogue still has not supplanted the more traditional media structure, though it has complemented it.

SNS Use and Attitudinal Effects

Scholars also have examined the relationship between SNS use and its effects on a person's attitudes toward politics. For instance, Vitak et al. (2011) found a positive relationship between a person's levels of political interest and engagement in political activities on Facebook.

Valenzuela et al. (2009) found that SNS use was positively related to higher levels of social trust. Contrary to previous research that found a positive relationship between traditional media use and cynicism (for a more in-depth discussion see Pinkleton and Austin 2001), Hanson et al. (2011) found a negative relationship between a person's level of political cynicism and their use of SNS. They suggested that these sites offered their users the ability to interact with others, thereby reducing people's levels of cynicism. Further, they proposed that these venues offered their users the ability to interact with other like-minded people to better understand politics and to increase feeling of political self-efficacy. However, without further investigation, it is impossible to know if this is a sustainable relationship, an artifact of a new technology, or a reaction to the messages used by the candidates of this particular election.

Once again, these links suggest a relationship between SNS use and attitudinal outcomes. Due to the relatively recent rise of SNS in political campaigns, the causal linkage between SNS use and its effects on attitudes has not yet been examined sufficiently.

Social Networking Sites and Political and Social Change

The ubiquitous nature of the Internet and the growth of satellite communications and mobile technologies such as cell phones have increased connection among citizens and global communication. The role of the Internet in politics has garnered a great deal of inquiry from scholars,

the public, social movements, and popular press. Its growth in use, when coupled with increasing globalization, has highlighted the potential of the Internet for political use (Jowett and O'Donnell 2012). Its potential for broadening participatory democracy has long been lauded. At the same time, some governments around the world have taken great pains to limit citizen access to the Internet in efforts to curb such broader political participation and change and to curb its potential social and political impact. For example, Myanmar shut down access to the Internet in 2007 to curb demonstrations. In China, YouTube, Twitter, and Facebook have been blocked, and China even shut down access to the Internet in one region during the fight between Uighurs and Han in 2009 (Jowett and O'Donnell 2011). The search engine Google has had to acquiesce to China's demands to censor content in order to operate there. China also has blocked content from mobile phone text messaging.

Part of the appeal and political potential of the Internet is due to the attributes of the medium which include its interactivity, low cost, lack of control by nations, and potential for anonymity. The emergence of social media sites in the first decade of the twenty-first century has expanded the Internet's political potential. The growth in the types of mobile communication technologies from which the Internet can be accessed such as cell phones and iPads has expanded opportunities even further.

In recent years this potential for political action and participation via social media and mobile technologies has been very visible in their use for communication during uprisings and protests around the globe. Social media such as Facebook and Twitter and text messaging via mobile phones have been used for political protests and/or uprisings in Ukraine in 2004 and Moldova in 2009 (Jowett and O'Donnell 2012). Cell phone cameras and other mobile devices have been used to upload images to social media sites and effectively spread the images captured. A well-recognized example occurred in 2009 when Iran cracked down on demonstrators after Mahmoud Ahmadinejad was elected. In the face of censorship of mainstream media, cell phone cameras and text messages were used to get pictures and descriptions of brutality and violence of police in cracking down on those who were marching in protest.

Mobile devices and social media were used at an unprecedented level to both mobilize and get images communicated around the world during the "Arab Spring" revolutions that spread across North Africa and the Middle East in 2011. In fact, by 2011 there were more than 27 million Facebook users in the Middle East (Al-Momani 2011). While a number of factors led to protests and events that toppled several authoritarian regimes and forced others to institute significant political reforms, social media such as Facebook, YouTube, and Twitter provided large numbers of people a vehicle for communication before, during, and after the protests and uprisings. Today a combination of mobile media (such as mobile cameras and text messaging) and SNS (such as Facebook and Twitter) is used to spread political ideas and foster and advance social and political movements and change. Therefore, it is hard to argue that revolutions during Arab Spring were social media-generated, when much of the planning and orchestration occurred offline.

In addition to those who challenge mainstream parties or those currently in power, SNS are also used by mainstream parties and campaigns to advance their agendas. At times, political candidates use SNS to go around mainstream media channels to get information from and communicate directly with constituents, supporters, or potential supporters (Jowett and O'Donnell 2012). Whereas traditional Internet sites provide voters with an avenue for getting political information on candidates at their discretion, SNS such as Twitter and Facebook give candidates an opportunity to connect with voters in a way much different than just setting up their own websites to be found when voters want it. SNS permit candidates to become part of voters' social networks and communicate their messages directly to them, making campaigns potentially much more interactive (Gilmore 2011). Gilmore (2011) suggested that such media may help disadvantaged groups and their candidates who are less well established compete with those who

are more entrenched and have significantly more resources for campaigning.

Like candidates, citizens, and interest groups are using social media in progressive movements such as the environmental movement, United for Change and Occupy Wall Street. In the United States, the Tea Party movement and Tea Party leaders have used SNS to encourage others to join the movement as well as pressure politicians on certain issues such as government spending and taxation (Barrow 2012). These movements use SNS to disseminate ideas amongst their members, coordinate movements, and connect with mainstream media outlets. However, when broad-sweeping change does occur, it is usually due to older methods such as marches, pressure from constituents, mainstream media coverage, and rallies. Previously those pushing for societal change had to rely on print media such as flyers, manifestos, and pamphlets. SNS have provided new vibrant forms of mass, group, and interpersonal communication to spread ideas more quickly and efficiently. Clearly, SNS do provide new forms of connection.

However, because there is little empirical research on the uses and effects of SNS for political communication specifically, we do not really understand all of their positive and negative effects at this time.

Conclusion

Newer social media, like media before them, are tools for accessing and sharing information. They have only been around for less than a decade, but they are now part of the media landscape. How they fit within that landscape, generally, and their role in politics, specifically, is still being explored. They are being used in social and political movements and by candidates, interest groups, campaigns, and advertisers to reach potential voters. They can be used by voters to satisfy their needs and desires for political information and political entertainment and to help them make up their minds and engage in their own political activities. Researchers are exploring these and other uses and effects of social-mediated communication. Nonetheless, more research is needed on individual differences and desires of users and the social and political contexts in which SNS use occurs. Until more is known about such use and effects, it is difficult to hypothesize or generalize about the effects of SNS use and/or the functions of different platforms for accessing SNS sites.

Understandably, there is still little empirical research demonstrating their effects on voter turnout, candidates selected, and impact on voters. That which does exist has tended to focus on campaigns in more developed nations where social media are widely used and diffused. Many of these nations also have somewhat stable democratic governments. More research has to consider variables such as sociopolitical factors that influence SNS use and effects; these include the extent of a population's access to SNS, the power structure of the different countries in which the SNS are used, the extent of government control exerted, the homogeneity of the populace using social media, the type of political activity, and a host of other sociopolitical factors.

Future Directions

It has been argued that real activism requires strong ties among people within a social system/network. Groups who use social media for real activism need large numbers of people beyond their more intimate social networks to muster the political will and numbers to effectuate change. Gladwell (2010) argued that SNS really build weak, not strong, ties and may not foster long-term relationships necessary to govern once change has occurred. However, with so many candidates and political parties focusing on these sites as venues for engaging their constituents, it is important to understand if SNS use is effective at sustaining these changes.

SNS and other Internet sites also can be used by group to hide their true identities, making it hard for those who access them to make informed decisions about the credibility of sources. This may make SNS potentially strong tools for spreading misinformation and

propaganda. Therefore, future scholars should not only examine the effects of SNS use but also examine the accuracy and truthfulness of the information presented to SNS users.

SNS use may assist people in engaging in more political activities (e.g., voting, protesting) and becoming more polarized. However, as of yet, it is impossible to determine if SNS are responsible for increased political activity and polarization or if people who are already politically active and polarized use these sites to reinforce such behavior and attitudes.

Future research also needs to explore the extent of fragmentation in SNS. One of the key aspects of SNS is that they can connect people together in groups. However, in the process of connecting with others in a group, people also may disconnect from those not associated with the group. Therefore, future scholars should examine how this connection and disconnection influences the public. Specifically, does this connection lead to negative view of those not associated with the group? Can it lead to greater in-group bias and out-group derogation?

Perhaps one of the most negative effects of SNS use is selective exposure. Specifically, when a person strongly identifies with a particular political party or movement, and he or she acquires most of his or her information through trusted (and more likely like-minded) political channels, does that have an impact of his or her own perceptions of reality? In particular, is he or she more likely to believe messages from politically similar people or organizations (in-group members) and disbelieve messages from politically dissimilar people or organizations (out-group members)?

Additionally, many of those who use SNS in political campaigns may be more likely to be politically active. However, what is unknown is if this activity is inherently good for democracy. Scholars should examine SNS use and determine if these sites are good for democracy and lead to productive change, if the use of these sites may foster partisanship and polarization and other unproductive changes, and/or if there is a mixture of both productive and unproductive changes as a result of using SNS.

Cross-References

- ▸ Community Evolution
- ▸ Election Forecasting, Scientific Approaches
- ▸ Political Networks
- ▸ Social Media
- ▸ Social Networks and Politics

References

Abroms LC, Lefebvre RC (2009) Obama's wired campaign: lessons for public health communication. J Health Commun 14:415–423

Al-Momani M (2011) The Arab "Youth Quake": implications on democratization and stability. Middle East Law Gov 3:159–170. doi:10.1163/187633711X591521

Barrow B (2012) Tea party evolves, achieves state policy victories. Retrieved from: http://news.yahoo.com/tea-party-evolves-achieves-state-134945533.html. Accessed 24 Oct 2012

Baumgartner JC, Morris JS (2010) MyFaceTube politics: social networking web sites and political engagement of young adults. Soc Sci Comput Rev 28:24–44. doi:10.1177/0894439309334325

Boyd D (2008) Can social network sites enable political action? Int J Media Cult Polit 4: 241–244. doi:10.1386/macp.4.2.241_3

Boyd DM, Ellison NB (2007) Social network sites: definition, history, and scholarship. J Comput Mediat Commun 13(1):210–230. doi:10.1111/j.1083-6101.2007.00393.x

Bradley T (2008) Final fundraising figure: Obama's $750M. Retrieved from: http://abcnews.go.com/Politics/ Vote2008/story?id=6397572&page=1&singlePage=true. Accessed 24 Oct 2012

Brown J, Broderick AJ, Lee N (2007) Word of mouth communication within online communities: conceptualizing the online social network. J Interact Mark 21(3):2–20. doi:10.1002/dir.20082

Gibson JL (2001) Social networks, civil society, and the prospects for consolidating Russia's democratic transition. Am J Pol Sci 45:51–68

Gilmore J (2011) Ditching the pack: digital media in the 2010 Brazilian congressional campaigns. New Media Soc 14:617–633. doi:10.1177/1461444811422429

Gladwell M (2010) Small change: why the revolution will not be tweeted. The New Yorker. Retrieved from: http://www.newyorker.com/reporting/2010/10/04/101004fa_fact_gladwell. Accessed 24 Oct 2012

Hanson G, Haridakis P, Wagstaff A, Sharma R, Ponder J (2011) The 2008 presidential campaign: political cynicism in the age of Facebook, MySpace and YouTube. Mass Commun Soc 13:584–507. doi:10.1080/15205436.2010.513470

Haridakis P, Hanson G (2009) Social interaction and co-viewing with YouTube: Blending mass communication reception and social connection. J Broadcast Electron, 53:317–335
Hindman M (2005) The real lessons of Howard Dean: reflections of the first digital campaign. Perspectives 3:121–128
Jowett G, O'Donnell V (2012) Propaganda and persuasion, 5th edn. Sage, Thousand Oaks
Jowett G, O'Donnell V (2011) Propaganda and persuasion (5th ed). Sage, Thousand Oaks, CA
Kavanaugh AL, Zin TT, Rosson MB, Carroll JM, Schmitz J, Kim BJ (2007) Local groups online: political learning and participation. Comput Support Coop Work 16:375–395. doi:10.1007/s10606-006-9029-9
Lazarsfeld PF, Berelson B, Gaudet H (1944) The people's choice. Columbia University Press, New York
Leighley J (1996) Group membership and the mobilization of political participation. J Polit 58:447–463
McClurg SD (2003) Social networks and political participation: the role of social interaction in explaining political participation. Pol Res Q 56:449–464. doi:10.1177/106591290305600407
Pinkleton BE, Austin EW (2001) Individual motivations, perceived media importance and political disaffection. Pol Commun 18:321–334. doi:10.1080/10584600152400365
Ramirez J (2008) The big picture. Newsweek. Retrieved from: http://www.thedailybeast.com/newsweek/2008/11/09/the-big-picture.html. Accessed 24 Oct 2012
Schlozman KL, Verba S, Brady H (1995) Participation's not a paradox: the view from American activists. Br J Pol Sci 25:1–36
Smith SD (2008) How many volunteers did Obama have? Message posted to https://my.barackobama.com/page/community/post/trishaifw/gGxZYv/commentary. Accessed 24 Oct 2012
Smith A (2011) The internet and political news sources. Retrieved from Pew Internet & American Life Project website: http://pewinternet.org/Reports/2011/The-Internet-and-Campaign-2010/Section-2.aspx. Accessed 24 Oct 2012
Sweetser KD, Kaid LL (2008) Stealth soapboxes: political information efficacy, cynicism and uses of celebrity weblogs among readers. New Media Soc 10:67–91. doi:10.1177/1461444807085322
Valenzuela S, Park N, Kee KF (2009) Is there social capital in a social network site?: Facebook use and college students' life satisfaction, trust, and participation. J Comput Mediat Commun 14:875–901. doi:10.1111/j.1083-6101.2009.01474.x
Vargas JA (2008) Obama raised half a billion online. The Washington Post. Retrieved from: http://voices.washingtonpost.com/44/2008/11/obama-raised-half-a-billion-on.html. Accessed 24 Oct 2012
Vitak J, Zube P, Smock A, Carr CT, Ellison N, Lampe C (2011) It's complicated: Facebook users' political participation in the 2008 election. Cyberpsychology, behavior, and social networking 14(3):107–114. doi:10.1089/cyber.2009.0226
Xenos M, Foot K (2008) Not your father's Internet: the generation gap in online politics. In: Bennett WL (ed) Civic life online: learning how digital media can engage youth. MIT, Cambridge, pp 51–70

Social Networking in the Telecom Industry

Laurent-Walter Goix[1] and Fabio Luciano Mondin[2]
[1]Telecom Italia S.p.A., Milano, MI, Italy
[2]Telecom Italia S.p.A., Torino, TO, Italy

Synonyms

Federated Social Web; Rich communication suite; Telco operators

Glossary

ENUM E.164 NUmber Mapping
FSW Federated Social Web
GSM Global System for Mobile communications
GSMA GSM Association
IETF Internet Engineering Task Force
IM Instant Messaging
IMAP Internet Message Access Protocol
IMS IP Multimedia Subsystem
IP Internet Protocol
MMS Multimedia Messaging Service
OMA Open Mobile Alliance
OTT Over-The-Top
POP Post Office Protocol
PSN Personal Social Network
RCS Rich Communication Suite
SIP Session Initiation Protocol
SMS Short Message Service
SMTP Simple Mail Transfer Protocol
SN Social Network
SNEW Social NEtwork Web
VAS Value Added Services
W3C World Wide Web Consortium

Definition

Soon after the rise of the Social Web on the Internet, pervasiveness, in particular mobile access, has been fostering the adaptation – and evolution – of the entire telecom industry. In this essay we illustrate how mobile devices, and consequently telco networks, have been tremendously evolving to support this trend through various approaches from competition to cooperation, backed up by examples of massive societal usage. We also report the continuum flow of technical activities ongoing in that area currently moving to the concept of federation of social networks, or interoperability, through standardization efforts, specification work from the industry and the Web community, and the deployment of early solutions and open-source projects.

Social Networks (SN) have introduced a new paradigm of communication/content exchange between users that have tremendously boosted the telecom industry over the last years through the massive adoption of smartphones and the explosion of broadband and mobile Internet accesses worldwide.

However, although widely used from mobile devices (e.g., 50 % of monthly Facebook users), SN services are not using mobile assets efficiently. Together with the many over-the-top mobile applications available nowadays, they are harming mobile network infrastructures due to heavy signalling traffic.

The related ecosystem is growing fast, driven mainly by the Web/enterprise industry, and moving towards standardization and regulatory institutions. Federation of Social Networks is the future of the Social Web that is expected to create brand-new business opportunities based on interoperable communities of all kinds.

This evolution can also be seen as a concrete potential opportunity for operators to leverage (back) their customer base relying on the phone number as a trusted user identity and on its reputation to protect customer privacy and ensure "data portability." At large scale, the success of such a service depends on the federation/peering amongst operators, at least at national level, as for the GSM service.

Introduction

While chat rooms (remember IRC) and instant messaging were the first services that gave birth to users and identities (although usually fake) within the Internet, their real-time constraints have been superseded by the advent of the current popular social networks and their chronological stream of activities. The epoch-making turn was made when Facebook launched the news feed in September 2006 (Marshall 2006) where users could see what their friends were doing at latest. . .

With a new (tele)communication paradigm that would sweep away real-time communication patterns (although still used to some extent) allowing people to keep in touch with friends anytime anywhere, with no need for contemporaneity, the "always-on" reachability of the wall de facto created a virtual representation of each user.

Since then the "wall" has become for its owner a history of private/public activities giving control of their outreach in an implicit manner: with respect to traditional (telco) systems, target audience is not defined one by one but grouped together in a list or circle, which is resolved by the central (dispatching) entity.

It appears clear that Social Networks have been facilitating many-to-many communications with respect to telco messaging systems (e.g., SMS). Besides, considering that posting a message on a Social Network has little to no cost one can understand why this mechanism allowed the massive widespread of information at a worldwide level. In some cases a single post can reach millions of users able to interact with each other up to the point of drastically impacting the society, like for earthquake prevention or popular revolutions. Such scenarios are later described in this essay.

Key Points

The knowledge user relationships, together with the real user identity, have become over the years the most valuable artifacts of the Social Web and

has led to many battles between telcos and OTTs on how to leverage, or obtain, this information.

Furthermore the operator's assets such as network-based user authentication and location or mobile push mechanisms are key elements for which applications, and more recently device operating systems, have been designing and implementing alternative solutions that in some cases still are suboptimal and harm the operator's network infrastructures and the device battery through heavy signalling, awaiting standards more friendly to the telecom industry.

In parallel the explosion of walled-garden Social Networks has fragmented, and in some cases replicated, users and their relationships based on their interests, unavoidably calling for interoperability (also called "federation") of social networks to avoid isolation (and the failure of *closed* social network tentatives of telcos). This same popular trend of vertical social networks has created privacy concerns by users in trusting their service providers, which may not have proven tracking records and which has further led to some self-regulation principles by regulation authorities.

Historical Background

It is widely believed that communication can be considered a primary need for human beings, such as eating or sleeping, and that is probably why they keep searching for better ways to (tele)communicate.

In the Beginning Was the SMS…

Even if social networking is quite a new concept, telecommunications systems exploiting the same principles have been used for years in many different ways and contexts. The Short Message Service (SMS), which was first used in 1993 (MobilePronto 2010), despite of its simplicity can be considered the first forefather of (mobile) social networks. Somehow SMS clearly showed the need of a direct, short, effective, and asynchronous way to communicate with friends, which can be found in its closest relatives. The SMS usage exploded also as a (near) real-time service and incentivized the specification of the MMS (Multimedia Messaging Service) standard, introduced in 2002 (Mobile Phones Uk 2012) to support multimedia content including photos and animations, which never reached the same success due to early interoperability problems on devices and high costs for users.

In the meantime on the Web side, Internet Relay Chat (IRC), allowing many users to exchange text in real time in a chat room, led to instant messaging, thus making another step towards Social Networks. Instant messaging reduced consistently the number of contacts and the audience per single message, but increased their quality and, most of all, introduced the concept of a "presence" status, a virtual real-time "marker" of our online availability in the instant messenger.

The Blogger's Dream…

As the Internet grew (Internet World Stats 2012), IM clients such as ICQ before and MSN and Yahoo! Messenger later went extremely popular, leading to a variety of specifications aimed at standardizing and interconnecting those types of systems.

Meanwhile, the Web community was experimenting different forms of communications. Bulletin boards became the best way of discussing about very specific topics, while weblogs (later called *blogs*) evolved from their initial idea of "online diary" to something more related to opinion and journalism, turning each blogger into a potential Pulitzer winner. Besides the "illusion" of blogs, some things became very clear: users were becoming content "prosumers" initiating the "Web 2.0" era, but their audience was smaller than the one they imagined, maybe due to the missing link with contacts, messaging, and audience/privacy control.

The Dawn of Social Networking…

Social Networking is the form of communication fitting best this need: a profile, in which users can put their content and show with whom they are in contact.

The way to Facebook, by far the most successful social network as of writing, was started in 1997 by SixDegrees and passed through

Friendster, Myspace, and Linkedin. The aces in the hole for Facebook were probably the insertion of the Facebook Wall, which somehow overtook the concept of online "presence" with an always-available "virtual presence" concept, together with the "Facebook Platform," opening the social network to third-party applications.

Diverging Interests

When looking at what is happening between telcos and over-the-top service providers (OTT), one could consider it an "epic battle between Ying and Yang." On one side OTTs are dedicated and structured to offer innovative services but may not have proven track records in managing personal information, while on the other side telco operators control the network and its assets and have a deep knowledge of their real users and their relationships but lack of rapid processes as their strategy is still focused on optimizing their network.

OTT services are usually offered over an untrusted domain by recent lightweight start-ups with very efficient and restricted process: users get to know about the specific service offered by that company, quickly subscribe, and start to use it, but there is no certainty about the real user's identity. This may also explain why Facebook, Google, and other OTT services are trying to get more user data such as the actual full name of their users or their phone number (Smith 2012). Indeed the competition is not really centered around the number of subscribers or active users to a specific service, rather to the quality of these users. Being able to profile users has become the key success factor for service providers who often request additional personal data and permissions to perform social data mining on profile and communication data. In a world in which online services are free of charge, revenues come mainly from advertising and their value increases the more it fits with the profile the advertising target.

Furthermore, OTTs count much on network capabilities that are not under their control. Usually their services benefit much of "always-on" users and have further contributed to accelerate the deployment (and subscription) of broadband and mobile Internet devices and infrastructures over the last decade (OMT 2012), also due to a viral effect amongst users & their friends.

Instead telco operators know their users very well: they get personal data when customers subscribe their contracts and also know the most active contacts of a user through voice or text communications, but usually have regulatory restrictions to leverage this data for any other purpose. Additionally, telcos are traditionally large companies (especially incumbent operators) still getting their revenue mainly from the voice service (Patuano 2012) and used to complex (thus slower) processes to accommodate high availability of their network together with regulatory compliance.

Which Solutions?

It appears clear that OTTs and telcos are nearly pulling in opposite directions. OTTs often perceive operators as "carriers": from their perspective operators should offer high-quality data connectivity to their customer and should not compete with them on services, which is typically what telcos want to avoid as the price of mobile Internet connection is lowering (also due to regulatory agencies) and so are revenues; offering high-quality, affordable services and exploiting user phone number and identity are something mobile operators perceive as an opportunity, a way to escape from the dreaded "bit pipe" fate.

One possible solution could be taking this competition to a higher level. Both OTT and telcos seem to be aware of the importance of user profiling (and of the related privacy issues), and both know very well how a complete user profile should largely exploit the user relationships. This can explain the "raison d'être" of Facebook's social graph and be even more evident looking at some used cases and applications: recently many OTT applications helped by the device evolution are trying to catch information from the user's address book. This is, for example, the case of WhatsApp (2012) in which users are identified by their phone number (requested at sign-up) and

buddies are automatically discovered through the address book. The WhatsApp intuition is to use the address book as a social network: one is in contact with many people in many different social networks, but the people one really keeps in contact with is probably a part of her/his address book.

On the other side, telcos, who easily have access to the user's phone number, are trying to exploit it socially although unsuccessfully. Many operators have tried to build their own social networks, e.g., Vodafone360, Orange Pikeo, and Telefonica KeTeKe, but none of these became truly popular, probably due to the absence of integration with other social networks and from the cold start problem of achieving critical mass.

The ace in the hole could be in making social network a commodity and move the competition from the mere "existence" of the social network service to the quality of the provided service.

New Approaches to Social Networking

Over the last years new approaches have emerged from the telecom industry to position themselves with respect to OTT. This is also related to the tremendous evolution of smartphones and data plans that fostered the wide adoption of social networking on mobile. According to Microsoft, in March 2011, 91 % of mobile Internet access was to socialize, and over 1/3 of FB's 600M users used it from mobile (Microsoft 2011). In June 2012, 57 % of Facebook's 950M users were mobile according to Facebook itself.

A first cooperative approach is in recognizing SNs as the owners of the user identity, contacts, and social interactions. This approach can vary from a plain "proxying" (e.g., through aggregation) to contractual partnership (sometimes exclusive). The aggregation (or gateway) functionality is nowadays a popular feature provided by telco operators or embedded within device operating systems themselves that do not have a strong relationship with a specific SN, but rather offer their own customers to connect to their favorite SN.

Social Network Aggregation services are popular entry doors to the social activities of users having multiple accounts over the Internet. They acquire messages, status feeds, content, and friends from various stand-alone SNs and aggregate all information in one point (device and/or server). Some specific – valuable – applications/features can cause users to migrate from isolated Social Networks to an Aggregation Site/Service. For example, some of them also offer cross-posting capabilities to simultaneously update all user accounts. This has become very popular on smartphones as well where the most popular Social Networks are integrated in operating systems such as iOS or Android to offer these features as native capabilities to users. Some tentatives have also emerged to design "social smartphones" that are explicitly focused on SN interactions (Inqmobile 2012). In some cases a business alliance is established with a specific SN to facilitate access from mobile devices (such as KDDI with GREE) in Japan (Fujimura 2011), although such a tighten relationship could eventually lead to fragmentation in "isolating" those mobile users from their friends on other operators that partnered to an alternative SN. This can become particularly risky in case of homegrown SN that is more popular within a single country than global players such as Facebook or Twitter.

Yet in other cases telcos have adopted even stronger strategies, by buying an existing SN to internalize know-how. This happened with Spain's Telefonica buying the national Tuenti in 2010 (Butcher 2010) to target "local" mobile SN services for youngsters (where telco can help) and leverage an already-popular and well-established SN (contrary to other tentatives of building one from scratch) to grow further. This challenging approach aims at growing and merging the SN user base with the telco customer base (potentially also abroad) while closing it de facto to "external" users and creating isolation.

Interestingly some telcos have realized the need to evolve these approaches of providing their own SN and open to noncustomers. DoCoMo in Japan has open its community service to other Japanese operators (Akimoto 2010), and more recently Telefonica announced the global

availability of Tuenti (Lunden 2012). These strategies are clearly tentatives to keep the community alive and overcome the isolation created by the customer-only approach at the time where mobile users, further assisted by number portability, are attracted by many offers to keep switching operators. While the validity of this approach still remains to be assessed as regards the long term, it clearly calls for interoperability and walled gardens end.

Indeed alternative paths have been studied over the last years for a long-term solution beyond establishing or partnering with walled-garden SNs. Such paths aim at defining standard specifications for the popular service that has now become SNs. Indeed in the meantime, SNs evolved from a niche youngsters playground to a must-have service for the telco industry now nearly to a commodity and a new global societal way of communicating and exchanging content. The telco standard community through the GSMA association has for several years been defining the RCS (Rich Communication Suite) specifications (RCS 2012) based on the SIP-based IMS infrastructure that focuses on real-time communications such as chat, file, & video sharing with other RCS-enabled users. Recently RCS was universally branded "Joyn" (Joynus 2012) and is being deployed and offered commercially by some telcos in Germany and Spain mainly. In this context, collaboration and the simultaneous launch of the interoperable service by multiple operators within the same country is essential for its success, following the lessons learned by the GSM cellular communication standard (in the positive sense) and the failure of stand-alone attempts of SNs.

In parallel, the Web community has also been moving towards standardization: while the telco industry could leverage its standardization and interoperability experience in communication services, the Web industry has been inventing the SN paradigm and improving it through various initiatives. This naturally led some large companies in the field as well as self-initiated initiatives to start building a "Federated Social Web" based on well-known Web technologies where the telecom industry is already active. This approach however is targeting the "wall-based" asynchronous & implicit communication paradigm, which is slightly different – and actually complementary – to the RCS-based communication scenarios. Eventually both these worlds will merge and some telco standardization initiatives already have been working in that sense (e.g., OMA Social Network Web). By participating in the standardization activities, the telco industry, manufacturers & telcos, can also improve the architecture and protocols to be optimized for networks and over-the-air communications by leveraging well-known assets.

The Way to Standardization (and Regulation)

Why?

While the Web is becoming increasingly social, social networking itself is heavily fragmented due to the multitude of disparate services, implementing a "walled-garden" approach as reported above. This limits interaction & sharing between users belonging to different Social Networks (S-N).

Furthermore privacy problems arise as global SN providers reside in different countries than their users and, besides legal implications, may not have proven track records in managing personal information. Users are requesting to have more control on sharing their own data or for the "right to be forgotten."

Besides consumers, businesses rely on popular SN (e.g., Facebook, Twitter) to promote themselves through a Social Media strategy, in the form of pages, advertising, and other initiatives. This ensures popularity but provides limited control over the community itself, to customize, manage, or animate it, or to get statistics, besides moving users away from the enterprise's official website.

Alternatively, creating their own user community as a stand-alone website typically results in being isolated from those SNs and remains a niche with little profit expectations.

Such enterprises, but also public administrations, are demanding for self-managed communities that can maximize brand awareness

and allow users to join while still be connected to their friends and other SNs.

Federation (or interoperability) is a proven solution for this type of issues and also a natural evolution of popular societal trends set by a few stand-alone competing initiatives that eventually need to collaborate. In the recent Internet history, the email communication system is a track record of such an evolution from proprietary systems (RFC808) to global standards (POP, IMAP, & SMTP to cite a few).

But interoperability is also a native concept within telecom industry (operators are interconnected for telco services, including IP-based, e.g., MMS, with well-defined procedures for global routing of phone numbers including ENUM).

The Evolution of the Social Web Ecosystem

Between 2008 and 2010 many initiatives within the Social Web have been dedicated to aggregation as a way to limit market fragmentation: FriendFeed. Such an approach is now showing strong limitations.

Starting 2010 this community promoted SN interoperability (or federation), similarly to email systems, to overcome silos and provide users back in control of their own identity & personal information. Some commercial platforms (e.g., Ning) allow users to "easily" set up their own SN in a hosted environment.

More recently, various initiatives ranging from stand-alone projects (Diaspora, Vodafone OneSocialWeb) to community based (OStatus, OpenSocial) or even commercial platforms (SocialEngine) now provide solutions to self-create & host one's own SN.

The Benefits of Interoperability

It is reasonably foreseeable that Federated Social Networks are the future of the Social Web. In this context users can communicate with each other across domains through global identifiers (whose syntax is similar to email addresses) without the need for replicating accounts. User data portability becomes easier so that users can choose their favorite social network and migrate. From a systemic perspective, such a distributed approach also provides major scaling & robustness of the overall Social Web avoiding single points of failure. For the telecom industry, such interoperability is also a benefit, besides an opportunity. For telcos it allows to leverage their existing customer base to offer SN communication paradigm, letting their subscribers interact with friends across different SN/operators similarly as with calls/SMS. By being involved in the definition of such specifications, it also allows to leverage mobile assets and ensure network optimization. For example, it can enable users to reuse their phone number as social identity or for authentication, which is seamlessly recognized and asserted by the operator's network. Furthermore by standardizing the core interaction features of the social networking communication paradigm, the migration across platforms provided by vendors should become seamless and further allow telcos to differentiate by providing specific rich features (e.g., games) beyond the "basic" interoperability. On the other hand device manufacturers can provide smartphones that can seamlessly connect to any social networking service irrespective of their provider, thus allowing users to easily switch devices & SNs.

The Current Standardization Landscape

As anticipated above, the Web community has the leading expertise on the SN world driving most of the specification work.

In particular, large enterprise software players such as IBM are leading the OpenSocial specification work and its reference open-source implementation work (Apache Shindig project), mostly targeting enterprise social containers. Similarly, Google (who initiated the OpenSocial work), Facebook, and others are either coauthors or early adopters of some specifications related to social data models or federation protocols.

In parallel, most of the biggest Web players are involved in the related standardization bodies or industry fora such as W3C, IETF, and the OpenSocial Foundation. While the latter has long-term expertise in designing client-server specifications for Social Networking, the IETF is currently focused on refining discovery protocols

and social network global identity. Within W3C several Community Groups were activated in 2012 that act as large discussion forums mainly targeting the "federation" aspects, anticipated in 2005 by the Social Web incubator group that started to investigate privacy concerns and a distributed approach (SW 2005).

Regarding federation specifically in 2010, OStatus created a Web-based specification targeted to interconnection of social networks by combining together several other draft specifications related to protocols and data models for exchanging social information. This created a de facto early reference for initial implementations from the open-source community and for the upcoming standards.

In the telecom standardization landscape, "Mobile Social Networking" (SNEW 2011) is viewed as a bridge between the SN Web community and the mobile world. OMA has been recently working at a specification called SNeW (Social Network Web) that targets this bridge with an end-to-end vision from the customer perspective.

Indeed current "mobile" version of SNs suffers from lack of mobile specificities on various aspects: frequent usage of polling instead of push notifications, no reuse of mobile identity/authentication, poor user experience in case of loss of connectivity or roaming (differed delivery not possible), and no integration with SMS/MMS or other traditional communication mechanisms.

In addition, most of the current open specifications are not addressing an end-to-end approach: OpenSocial or OStatus are in fact focused only on a specific type of interactions (respectively client-server and server-server) with a lack of consideration for interworking of such specifications.

Towards Regulation

As described above, standardization initiatives, and the Web industry, are focusing on solutions (protocols, data models, & architectures) for social network interoperability. In this context, increasing care is given to tackle data privacy issues from a technical perspective, in particular with respect to discovery, sharing, and deletion of users' data.

Over the past years, various legal cases have been targeting SNs on leaks and breaches in managing user's data privacy, typically under the jurisdiction of the SN's home country that may bypass institutions or even violate local laws of their users.

Since 2008 the European Commission has been working with SN providers on a concept of self-regulation to overcome the duration of a European legislation process in that field. The basic idea is for SNs to self-declare their compliance with "safe principles" that target young people protection. Most of the current popular (and mostly non-EU) SNs have provided such a declaration, further explaining how they implemented it (EU-selfreg 2011). Such declarations have been assessed periodically (latest in 2011) by the European Commission through an independent assessment on nine social networking sites (EU-report 2012).

In January 2012, Viviane Reding, Vice-President of the European Commission, EU Justice Commissioner, has further announced her/his commitment to give back users the control over their personal data (EU-dataprotection 2012):

> You will have an effective "right to be forgotten" so that you can remove your personal information from any site if you so wish;
> Web operators must provide 'privacy by default'. The default settings for all services should be the most privacy-friendly;
> You will have the right to know how your personal data will be used and where your consent is required, you must give it explicitly;
> You will be able to move your personal data from one service provider to another more easily ("data portability");
> Organizations processing your personal data must inform you as soon as possible if your data has been compromised;
> Your personal data will enjoy the same level of protection if it is transferred outside the EU as applies within the EU - vital in this age of instant global data flows.

Although not yet effective, this statement is clearly attempting to relaunch the debate in overcoming the current limitations of the privacy laws in place in most countries regarding digital identity & related data privacy.

Key Applications

The Potential of Mobile Social Networking

It is a fact that SNs are more and more influencing our daily life as they are powerful and independent sources of information, so powerful to be used for earthquake prevention, and so independent to help in spreading news and organizing protests during the Arab spring.

Surveys (Huang 2011) showed that nearly 9 in 10 Egyptian and Tunisian used Facebook and Twitter to organize protests and get news. In such a context, rapidly evolving and changing in which the main media were under control by the government, SNs got a key role, due to their speed and independence. It has been shown (Stepanova 2011; Ellis 2011) that the Twitter updates were faster than the media updates (and widespread because of low flat costs for mobile Internet, starting $8 in Egypt).

Twitter's speed is being exploited also by another application, aiming at reducing the number of victims caused by earthquakes. It has been seen (Sakaki et al. 2012) that it is possible to use Twitter to detect target events such as earthquakes by using each Twitter user as a sensor revealing data in real time. Such an earthquake reporting system has been really developed in Japan where the earthquakes are more frequent.

Social Network Analysis of Telecom Data

The idea to consider social network services as a field of convergence for services has been already taken into account by many players. A proposal is to identify social networks over the Telco Networks (Galindo 2008). Each communication media can be the starting point for a network of people, and discovering and exploiting this information can be a valuable opportunity for telco operators.

Social Network discovery can be performed by analyzing user's call and SMS history (Tomar 2010) in order to understand which people in our address book users are more in touch with. The basic idea standing behind this approach is to discover the social graph underneath the network and exploit this to empower the provided services.

Future Directions

Nearly related to the concept of FSW stands the idea of Personal Social Networks (PSN): once a technology is able to offer users interconnected social networks, there is theoretically no constraint on the dimension of the social network.

The idea standing behind PSN is to have a trusted environment for user's data. The user publishes his/her data on the personal social network, and the federation becomes a way to share data with users belonging to different SNs (personal or not). The advantage is that users can publish their data on a system, which is under their direct control and thus are free to turn off at any time, a technology that could comply to the EU Directives about digital oblivion and data portability. Besides the directives about privacy already mentioned earlier in this essay, the European Commission has shown growing interests about this topic which is standing behind projects such as di.me (Di.me consortium 2010), related to personal services, and Societies (ICT-Societies.eu), related to community smartspaces.

In particular, di.me also relates to semantics, a popular research topic beyond social networks, where a precursor can be seen in SMOB (Passant 2008) as early semantic microblogging tool. The basic idea is to have any social information semantically described in a machine understandable language (such as RDF). This gives the possibility to augment content with external content (e.g., provided by Linked Open Data) and thus to provide users the content they are really searching for through enriched semantic queries (Rodriguez 2012).

Cross-References

▸ Actionable Information in Social Networks, Diffusion of
▸ Community Evolution
▸ Connecting Communities
▸ Exchange Networks
▸ Extracting Individual and Group Behavior from Mobility Data

▸ Futures of Social Networks: Where Are Trends Heading?
▸ Inter-organizational Networks
▸ Linked Open Data
▸ Location-Based Social Networks
▸ Location-Based Status Updates and Camera Phone Apps in Social Networks
▸ Misinformation in Social Networks, Analyzing Twitter During Crisis Events
▸ Mobile- and Context-Aware Applications of Social Networks
▸ Mobile Communication Networks
▸ Multiple Social Networks, Data Models and Measures for
▸ New Intermediaries of Personal Information: The FB Ecosystem
▸ Online Privacy Paradox and Social Networks
▸ Personal Networks: The Intertwining of Ties, Internet and Geography
▸ Privacy and Disclosure in a Social Networking Community
▸ Privacy and Disclosure in a Social Networking Community
▸ Privacy in Social Networks, Current and Future Research Trends on
▸ Privacy Issues for SNS and Mobile SNS
▸ Semantic Social Networks
▸ Social Communication Network, Case Study
▸ Telecommunications Fraud Detection, Using Social Networks for
▸ Trust in Social Networks

References

Akimoto A (2011) Japan No.1 cellphone carrier's official social network now opened to other two. asiajin.com, 26 Apr 2011. http://asiajin.com/blog/2010/04/26/japan-no-1-cellphone-carriers-official-social-network-now-opened-to-other-two/. Last Access 20 Dec 2012

Butcher M (2010) Tuenti looks like it will go to Telefonica for $99 million. techcrunch.com, 4 Aug 2010. http://techcrunch.com/2010/08/04/tuenti-looks-like-it-will-go-to-telefonica-for-e75-million/. Last Access 20 Dec 2012

Di.me consortium (2010) http://www.dime-project.eu/en/home/dime/project/contenido.aspx. Last Access 20 Dec 2012

Ellis W (2011) The role of information and communication technologies in shaping the Arab Spring. POLI-340, Nov 2011. http://www.scribd.com/doc/75905411/The-Role-of-Information-and-Communication-Technologies-in-Shaping-the-Arab-Spring. Last Access 20 Dec 2012

EU-dataprotection (2012) Protection of personal data. http://ec.europa.eu/justice/data-protection/index_en.htm. Last Access 20 Dec 2012

EU-report (2001) Implementation of the safer social networking principles for the EU, Sept 2001. http://ec.europa.eu/information_society/activities/social_networking/eu_action/implementation_princip_2011/index_en.htm. Last Access 20 Dec 2012

EU-selfreg (2011) Safer social networking: the choice of self-regulation. http://ec.europa.eu/information_society/activities/social_networking/eu_action/selfreg/index_en.htm. Last Access 20 Dec 2012

Fujimura N, Yamaguchi Y (2011) Gree, KDDI Sue DeNA Amid Japan social-network competition. Bloomberg.com, 21 Nov 2011. http://www.bloomberg.com/news/2011-11-21/gree-kddi-sue-dena-amid-japan-social-network-competition-2-.html. Last Access 20 Dec 2012

Galindo LA et al (2008) The social network behind telecom networks. Position paper for W3C workshop on the future of social networking. http://www.w3.org/2008/09/msnws/papers/telefonica-business-operator.pdf. Last Access 20 Dec 2012

Huang C (2011) Facebook and Twitter key to Arab Spring uprising: report. The National, June 2011. http://www.thenational.ae/news/uae-news/facebook-and-twitter-key-to-arab-spring-uprisings-report. Last Access 20 Dec 2012

ICT-Societies.eu, http://www.ict-societies.eu/. Last Access 20 Dec 2012

Inqmobile (2012) http://www.inqmobile.com. Last Access 20 Dec 2012

Internet World Stats (2012) Internet growth statistics. http://www.internetworldstats.com/emarketing.htm. Last Access 20 Dec 2012

Joynus (2012) http://www.joynus.com/. Last Access 20 Dec 2012

Lunden I (2012) Tuenti, Telefonica's answer to Facebook and Twitter, opens up to users worldwide. techcrunch.com, 11 July 2012. http://techcrunch.com/2012/07/11/tuenti-telefonicas-answer-to-facebook-and-twitter-opens-up-to-users-worldwide/. Last Access 20 Dec 2012

Marshall M (2006) Facebook launches "News Feed" and "Mini Feed" – as YouTube invades turf. venturebeat.com, Sept 2006. http://venturebeat.com/2006/09/05/facebook-launches-news-feed-and-mini-feed-as-youtube-invades-turf/. Last Access 20 Dec 2012

Microsoft (2011) Mobile Stats 2011. Microsoft Tag, Mar 2011. http://tag.microsoft.com/Libraries/Blog/mobile-marketing-and-advertising-landscape.sflb.ashx. Last Access 1 June 2012

Mobile Phones Uk (2012) What is MMS? http://www.mobile-phones-uk.org.uk/mms.htm. Last Access 20 Dec 2012
MobilePronto (2010) The history of SMS text messaging. http://www.mobilepronto.org/en-us/the-history-of-sms.html. Last Access 20 Dec 2012
OMT (2012) Social media vs smartphone usage in Europe. Online marketing trends, Feb 2012. http://www.onlinemarketing-trends.com/2012/02/social-media-vs-smartphone-usage-in.html. Last Access 20 Dec 2012
OStatus (2010) http://ostatus.org/. Last Access 20 Dec 2012
Passant et al (2008) Microblogging: a semantic web and distributed approach. In: SFSW
Patuano M (2012) Telecom Italia 1H 2012 results. Telecom Italia Webcasting. http://telecomitalia.webcasting.it/1H2012-ondemand/files/SlideMarcoPatuano1H2012.pdf. Last Access 20 Dec 2012
RCS (2012) http://www.gsma.com/rcs/. Last Access 20 Dec 2012
RFC808 (1982) Postel J, Mar 1982. http://www.rfc-editor.org/rfc/rfc808.txt, Appendix A. Last Access 20 Dec 2012
Rodriguez RO et al (2012) LODifying personal content sharing. In: EDBT conference, Berlin
Sakaki T, Okazaki M, Matsuo Y (2012) Tweet analysis for real-time event detection and earthquake reporting system development. IEEE Trans Knowl Data Eng. IEEE Computer Society Digital Library. IEEE Computer Society
Smith G (2012) Now Facebook wants your mobile number: social network to ask 900 million users for phone details to, 'prevent' Linkedin-style hack. The Daily Mail. http://www.dailymail.co.uk/sciencetech/article-2159672/Facebook-ask-900-million-users-phone-details-prevent-LinkedIn-style-hack.html. Last Access 20 Dec 2012
SNEW (2011) OMA brings interoperability of social networks to the mobile world. http://www.openmobilealliance.org/comms/technical/msn_overview.htm. Last Access 20 Dec 2012
Stepanova E (2011) The role of information communication technologies in the "Arab Spring", May 2011. Ponars Eurasia. http://www.gwu.edu/~ieresgwu/assets/docs/ponars/pepm_159.pdf. Last Access 20 Dec 2012
SW (2005) A standards-based, open and privacy-aware social web. http://www.w3.org/2005/Incubator/socialweb/XGR-socialweb-20101206/. Last Access 20 Dec 2012
Tomar V et al (2010) Social network analysis of the short message service, TICET. http://www.ee.iitb.ac.in/~karandi/pubs_ dir/conferences /vikrant_himanshu_karandikar_vinay_swati_prateek_ncc10.pdf. Last Access 20 Dec 2012
WhatsApp (2012) http://www.whatsapp.com. Last Access 20 Dec 2012

Social Networking on the World Wide Web

Qingpeng Zhang[1], Dominic DiFranzo[1], and James A. Hendler[2]
[1]Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY, USA
[2]Computer and Cognitive Science Departments, Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY, USA

Synonyms

SNS; Social network sites; Social networking services; Social networking sites

Glossary

Collective Intelligence Shared or group intelligence that emerges from the collaboration, collective efforts, and competition of many individuals and appears in consensus decision making (http://en.wikipedia.org/wiki/Collective_intelligence).

Crowdsourcing The practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people, and especially from an online community, rather than from traditional employees or suppliers (http://www.merriam-webster.com/dictionary/crowdsourcing).

Human Flesh Search, HFS Is the phenomenon of distributed researching using Internet media such as blogs, forums, microblog, etc. (Zhang et al. 2012).

Microblog A broadcast medium that exists in the form of blog, with content that has a typically smaller size.

Social Computing, Computational Social Science Computational facilitation of social studies and human social dynamics, as well as the design and use of information and communication technology technologies that consider social context (Wang et al. 2007).

Social Media The forms of Web interactions among people through which Web users create, and share information, personal messages, ideas, etc. in virtual communities and SNS.

Social Networking Sites, SNS A platform for Web users to build and maintain social networks and share interests, activities, and/or real-life connections.

Web 2.0 The websites that use advanced technology beyond the static pages to enhance the social networking applications.

Web Science The socio-technical science of understanding the complex, cross-disciplinary dynamics driving development on the Web (http://tw.rpi.edu/web/concept/WebScience).

Introduction

The rising popularity and use of social computing technologies has not only connected people in new and interesting ways but has also generated vast amounts of data on human crowd behavior. This is allowing researchers to view and study crowds and communities at a scale never before possible. The growth of social networking sites (SNSs) has dramatically changed the way people communicate, collaborate, and maintain their social connections. Social networking on the Web has also enabled the emergence of crowdsourcing and collective intelligence sites, allowing for new economies and workflows to develop. In the past decade, SNSs have played an important role in current movements all around the world. This article reviews the history of social network sites on the Web and summarizes research on SNSs. This article also includes a review of Chinese SNSs, which has not been fully taken into consideration previously because of the isolation of SNSs in the Chinese Mainland.

Social Networking on the World Wide Web

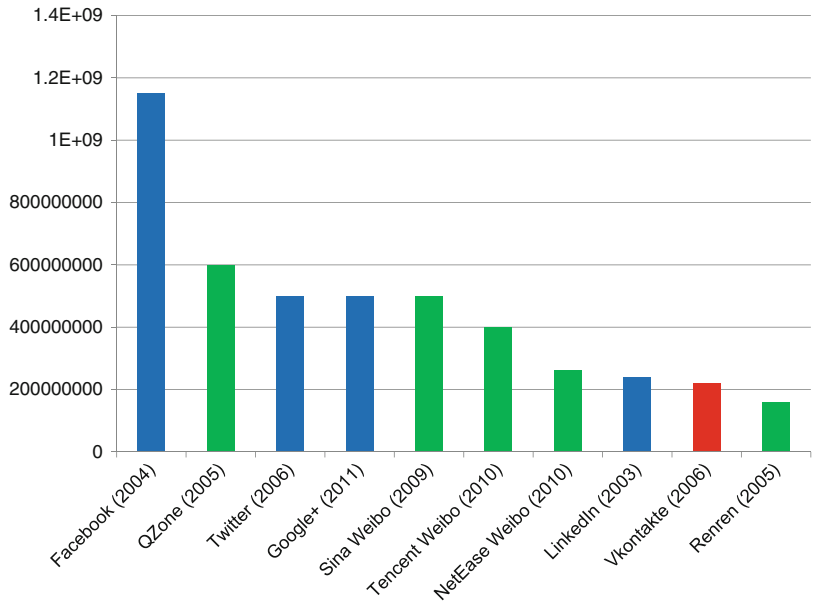
Global SNSs at a Glance

Social networking sites have become predominant in the age of the World Wide Web. The burst of social networking sites (SNSs) has dramatically changed the way people communicate, collaborate, and maintain their social connections. SNSs provide platforms and interfaces that enable people to follow and communicate with their friends, families, and other social connections. The sizes of SNSs have been growing rapidly. Boyd and Ellison reviewed SNSs in 2006 and defined SNS as "web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system" (boyd and Ellison 2007). We feel this definition is still appropriate to describe the phenomena and covers newer SNSs (especially microblogging sites like Twitter). Figure 1 and Table 1 show the top ten largest SNSs according to the number of registered users (date of count and resources listed in Table 1). In Table 1, we use "SNS" to refer to traditional SNSs as defined in boyd and Ellison (2007) and use "microblog SNS" to annotate Twitter and other equivalent services. Figure 2 is from the June 2013 edition of the World Map of Social Networks as reported in the Vincos Blog (http://vincos.it/world-map-of-social-networks/). It shows a map of the most popular social networking sites by country, according to Alexa traffic data (http://www.alexa.com/).

As shown in Figs. 1 and 2, and Table 1, Facebook is the largest SNS in the world, with over a billion registered users. Following, Twitter is the third largest SNS and the largest microblog site, with over 500 million registered users. These two US-originated SNSs are dominating the social networking services all over the world, except in a small number of countries, which either have very strong SNSs of their own or have limits to the access of Facebook and Twitter.

Chinese SNSs

In this article, we use China to refer Chinese Mainland, which does not include Taiwan, and China's Special Administrative Regions of Hong Kong and Macau. The censorship of Internet in



Social Networking on the World Wide Web, Fig. 1 Ten largest SNSs (with launched year) according to the number of registered users, as in late 2012 to early 2013. Color represents the origin. *Blue*, USA; *green*:, China; *red*, Russia

Chinese Mainland is stricter. As of July 2013, the number of Internet users in China is nearly 600 million, which consist over one fifth of the global Internet users (CNNIC 2013; ITU 2013). Among them, 78.5 % primarily use mobile phone to surf the Internet (CNNIC 2013). The very large number of Chinese Internet users has enabled the birth of many SNS giants in China (five of the top ten largest SNSs in the world). As the access to several SNSs (including Facebook and Twitter) has been limited in China (most require use of a virtual private network to visit) since 2009 (Facebook was also blocked for a small time period several times before 2009), the Chinese SNSs do not directly compete with American SNSs. It is worth noting that the restriction to visit oversea SNSs is not the only reason for the growth of Chinese SNS giants. The biggest SNS in China, QZone, was founded in 2005, just after Facebook was born. In fact, a lot of foreign Internet services (like ICQ, the first Internet-wide instant messaging service in the late 1990s) were defeated by their Chinese equivalents (which were usually improved and optimized for Chinese users) due to a variety of reasons (which will be discussed later in this paper). In addition, Chinese Internet users use online forums extensively. There are also some novel and unique Chinese SNSs which have no equivalents to any foreign services (like douban.com, launched in 2005). The more inherent cultural factors of this phenomenon are yet to be analyzed.

The History of SNSs

As there is a comprehensive review of SNSs prior to 2006 (boyd and Ellison 2007), we briefly summarize the history of SNSs before 2006 and then concentrate on a more recent activity.

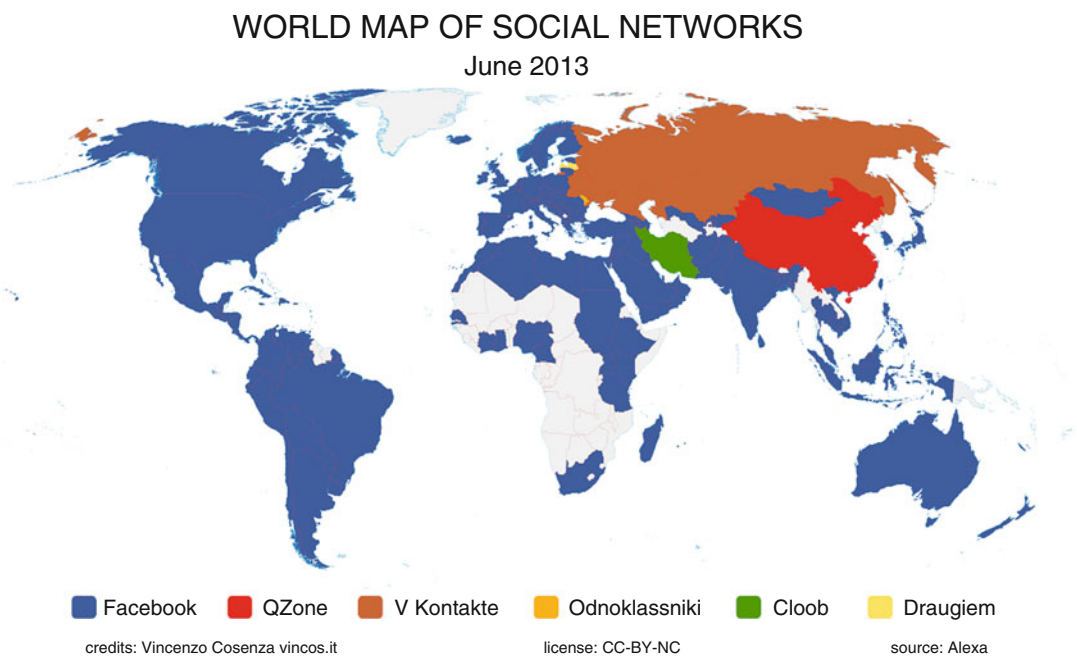
Before 2000 (Early Days)

Launched in 1997, SixDegrees.com is usually cited as the first recognizable SNS. It contains the basic functions of SNS, including listing friends and maintaining personal profiles. There were millions of users in SixDegrees in the late 1990s, before it was closed in 2000. The founder of SixDegrees thought that, at that time, Internet users did not have many friends online and people usually did not want to meet strangers (boyd and Ellison 2007). The next important SNS is LiveJournal, which was launched in 1999.

Social Networking on the World Wide Web, Table 1 Ten largest SNSs

Name	Registered users (in million)	Time counted	Year launched	Original country	Type	Source
Facebook	1150	Mar 2013	2004	USA	SNS	http://investor.fb.com/releasedetail.cfm?ReleaseID=761090
QZone	610	May 2013	2005	China	SNS	http://www.tencent.com/en-us/content/at2013/attachments/20130515.pdf
Twitter	500	Mar 2013	2006	USA	Microblog SNS	http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html
Google+	500	May 2013	2011	USA	SNS	http://googleblog.blogspot.com/2012/12/google-communities-and-photos.html
Sina Weibo	500	Dec 2012	2009	China	Microblog SNS	http://news.xinhuanet.com/tech/2013-02/21/c_124369171.htm
Tencent Weibo	400	Dec 2012	2010	China	Microblog SNS	http://www.techweb.com.cn/internet/2012-04-24/1183131.shtml
NetEase Weibo	260	Oct 2012	2010	China	Microblog SNS	http://tech.163.com/12/1018/18/8E49Q121000915BF.html
LinkedIn	238	May 2013	2003	USA	SNS	http://press.linkedin.com/about
Vkontakte	220	Aug 2013	2006	Russia	SNS	http://vk.com/catalog.php
Renren	160	Aug 2012	2005	China	SNS	http://life.renren.com/

Around the same time, SNSs also emerged in Asia, for example, Cyworld in South Korea (launched in 1999, with SNS features added in 2001) and Tencent QQ, which was launched in 1999 as a Chinese equivalent of ICQ, and added SNS-type features known as QZone in 2005. Due to Tencent dominating the instant messaging field in China (about 800 million active accounts), QZone, the SNS service for QQ users, has been one of the biggest SNSs in the world since its birth in 2005 (QZone was once the largest SNS before Facebook bypassed it) (boyd and Ellison 2007). QZone has integrated a lot of features, first as a blog site, and then music and photo sharing, and even some microblog-similar features before microblogging became popular in China. However, these features did not bring overwhelming success for QZone in either SNS or blogging, despite the very large number of users. Tencent even produced separate SNS (Tencent Pengyou) and microblog (Tencent Weibo) services to compete with other Chinese SNSs. Many of these services failed, but some eventually became popular. For example, Tencent Weibo, a microblog service, is now #6 largest SNS in terms of the number of registered users.



Social Networking on the World Wide Web, Fig. 2 The most popular SNSs by country, according to Alexa traffic data (from June 2013 edition of the World Map of Social Networks; http://vincos.it/world-map-of-socialnetworks/; accessed September 5, 2013)

2001–2004 (Burst of SNSs)

Ryze.com was launched in 2001 as a business-driven SNS. Following Ryze, Friendster was launched in 2002 as a social networking complement to Ryze, aiming to help build friend-of-friend connections. Friendster was the top SNS until 2004, when it was overtaken by MySpace. Friendster found a second home in Asia, as a social gaming site, with over 90 % traffic coming from Asia. After the success of Friendster, a large number of SNSs were launched. Among them, LinkedIn, an SNS for professional connections and networks, became one of the most successful and largest SNSs. At the same time, media-sharing sites like YouTube and Flickr started to incorporate more SNS-type features. MySpace was launched in 2003 and because of a variety of both technical and societal difficulties Friendster was facing (one of the major difficulty was with the ill-equipped databases), many Friendster users migrated to MySpace and other SNSs (boyd and Ellison 2007). MySpace employed the (musical) bands-and-fans dynamic to attract both bands and fans to join in and communicate. This strategy fostered marketing in SNS. Several years later, a similar and more well-designed strategy was successfully adopted by Sina to promote their microblog service in China, Sina Weibo (will be discussed later). MySpace also allowed users to add and modify HTML elements into their profiles to generate more personalized MySpace pages. In 2004, there were more and younger Internet users joining MySpace. They joined MySpace mainly because they would like to connect with their favorite bands. The "word-of-mouth" effect quickly spread in the teens' world, and MySpace grew very fast during this period. During this time period, IT giant Microsoft also released their SNS service in 2004, MSN Spaces, which became very popular outside the USA (e.g., China). However, MSN Spaces closed in 2011.

Since 2004 (Facebook and Its Equivalents)

In early 2004, Facebook was launched as an SNS only for Harvard students. It opened to other college students and then high school students

and corporate networks in 2005. Eventually, Facebook moved to an open signup (to users older than 13) in 2006. The social ties in Facebook are mutual. Users have their profile pages and news feeds for their home pages to highlight the updates of users' activities. Each user has a "wall," which summarizes updates of his or her friends. Facebook also added the "like" feature so that users can express that they like others' content. Facebook also has messaging functions, and the mobile app of Facebook enables users to communicate with their Facebook friends without having to sit in front of a computer. The major difference between Facebook and MySpace is that Facebook requires users to give their true identity (at least before the open signup). In addition, Facebook allows developers to produce "applications" and "games" for Facebook to allow users to personalize their pages and have more fun with their friends. Facebook also revealed its "Facebook Platform" (with "Graph API" as a core) in May 2007 to allow developers to read and edit the data of Facebook, especially the social graph. Facebook overtook MySpace in April 2008 and eventually became the largest SNS worldwide. It is also the most popular SNS in English-speaking countries. Facebook had 500 million users in July 2010 and quickly doubled it to 1 billion users in October 2012. Today, Facebook is not only an SNS for many people. It is an integrated social platform for almost everyone in many countries around the world (see Fig. 2).

Because Facebook requires users to use their real names, privacy has been a big concern. In November 2007, Facebook implemented its advertising system, Beacon. It used the data of Facebook users and advertised to friends of users using the history of purchases they made, causing a backlash of criticism. It was shut down in a month. In 2009, Facebook enabled users to choose which parts of their profile can be viewed by everyone, though the name and profile photo are always accessible to public.

Facebook is not only the largest SNS in America but also the largest SNS in Europe. However, Facebook is not dominant in Russian-speaking countries. VKontakte (VK) was launched in Russia in 2006. It was first only for college students and then opened to public. It quickly grew and became the second largest SNS in Europe.

In 2011, Google launched its SNS, Google+, after the failure of Google Buzz. Google+ has been described as a combination of Facebook and Twitter, with an aim to attract users from both sites. Google+ has its unique "circles" for users to organize their friendship information. "Circle" enables Google+ to have "social layers" and enhanced a major property that Facebook and Twitter lacked, making users' updates and messages visible to only a subgroup of their contacts, instead of pushing their information to everyone connected to them (Facebook and Twitter have since added their own variants of this capability). Google+ reached 500 million users in May 2013, making it the fourth largest SNS worldwide. However, there are many reports saying that Google+ is a "ghost town," with a large number of registered users but few activities (Gonzalez et al. 2013). Gonzalez et al. conducted a comprehensive empirical study of Google+, looking at its topological properties and evolution patterns. They found that the stable connectivity features of Google+ network were very similar to Twitter and different from Facebook, indicating that the use of Google+ was more like the messaging propagation in Twitter, rather than pairwise relations in Facebook. They also found that the user is not actively engaged in Google+ network, as compared with Twitter and Facebook (Gonzalez et al. 2013). More research on Google+ has focused on the privacy issue, taking a closer look at its "circle" function.

During the same time period, Renren (formerly known as Xiaonei, literally "on-campus network") was launched in China in 2005. It is widely known as the Chinese equivalent of Facebook. Similar to Facebook and VKontakte, Xiaonei was first only open to college students. In August 2009, Xiaonei was renamed to Renren (literally "everyone's network"), in order to expand its user size. Renren has been competing with Kaixin001 since the latter was launched in 2008. Kaixin001 first aimed at "white collars" (educated people performing professional,

managerial, or administrative work in office) and then changed their strategy to compete with Renren for all users. Both SNSs have their own user groups. They have been losing active users since the birth of Sina Weibo, a microblog service. As it now stands, Sina Weibo and other microblog services are losing their active users (CNNIC 2013), with the emergency of WeChat (we will discuss Chinese SNS later in this paper in more detail).

The burst of SNSs has also attracted the attention of researchers. The rich data generated by SNSs provided ideal test-beds for research. In 2007, Wang et al. revisited the term "social computing" and gave a new definition to refer the research on both the design of social software (which was coined as "social computing" in 1994 by Doug Schuler) (Schuler 1994) and the study of social systems using computational science methodologies (Wang et al. 2007). In 2009, another similar term "computational social science" appeared in Science Magazine (Lazer et al. 2009). Computational social science refers to the second part of social computing, and both terms became popular and widely used among researchers studying SNSs.

Besides fostering the birth of social computing and computational social science, the bursts of SNSs also facilitated the growth of several other domains, including young fields like network science (Barabási 2013) and Web science (Shadbolt et al. 2013) and mature fields like data mining (Han et al. 2006), machine learning (Bishop and Nasrabadi 2006), and natural language processing (Manning and Schütze 1999). In particular, the study of various social networks formed by SNSs has been one of the most active research topics following pioneering work defining properties such as scale-free and small world networks (Barabási 2013). This research on SNSs expanded earlier small-scale (tens or hundreds of nodes and edges) survey-based social network analysis to very large scale, usually from thousands to millions of nodes and edges. The nodes in these social networks were typically a unique user ID in SNS, and the edges between nodes represented different types of social connections/interactions, including directed or undirected friendship, message exchange, and comment and reply, which normally indicate the social structure and the information propagation in SNS. During the past decade, researchers have studied almost every popular SNS, including the blogosphere, Facebook, Google+, Renren, various media-sharing SNSs, and Q&A SNSs (please refer to Recommended Reading section for typical publications of these SNSs). The social network analysis (SNA) studies revealed many interesting aspects of the social systems and dynamics of SNSs. We summarize a few typical research results. (We summarize the results briefly below. For more details, please refer to Recommended Reading section for source publications of the results).

- In most SNSs, people found that a small portion of the users were controlling the communications and information spread in SNSs, and people are easily connected with each other via "travelling" through those key users, known as hubs.
- Users were clustered around different topics, and in certain events (like political elections and revolutions), users were polarized into two or more big clusters, with few interactions in between.
- Researchers also conducted temporal and spatial analysis on the conversations in SNSs.
- There are many successful algorithms being developed to discover the subcommunities in SNSs based on social networks.
- In addition, topic models and other probabilistic models have been employed to further explore the implicit subcommunities.
- Furthermore, the privacy and trust issues in SNSs have also been studied.
- Researchers have conducted empirical studies of the use of SNSs in social movements and performed experiments of using SNS for social mobilization.
- For behavioral and social science researchers, various theories in social network can be validated with the "big data." Among them, balance theory ("the enemy of my enemy is my friend") was one of the most intuitive and early studied theories, and it was found to hold in most SNSs.

- The strength of weak ties and the relevant structural holes theory have also been validated in SNSs, showing that users in the broker position of social networks formed by SNS have the advantage to be more innovative and productive because they have access to various fresh ideas.

For details and a more comprehensive review of state-of-the-art research on SNSs, please refer to other chapters of the Encyclopedia of Social Network Analysis and Mining.

Since 2006 (Twitter and Its Equivalents)

The birth of Twitter in 2006 changed the cyberspace again. Twitter created a new form of SNS named a microblog, in which users post short messages (up to 140 characters) via the Web, smartphone apps, email, mobile phones, and instant messages. Different from other SNSs, the relationship in Twitter is not reciprocal, meaning that a user can follow other users, and a user can be followed by others without following them. The followers of a user in Twitter can view the messages (named as "Tweets") from the user. The followers can reply or retweet this user's tweets. Twitter users use @ to mention a Twitter user and hashtag # to represent a topic of the tweet.

Since its launch, Twitter quickly became one of the most visited websites as "the SMS of the Internet" and the largest microblog site worldwide (though its Chinese equivalent is close). As compared to traditional blogs and SNSs, microblogging is a faster method to communicate, share quick thoughts, and report news. In addition, the frequency of updating a microblog is usually much higher than traditional blogs and SNSs. These features made Twitter and other microblogging services distinct. In a recent review, Murthy describes Facebook as to "keep ties between users active and vibrant," while Twitter is used to seek the "accumulation of more and more followers who are aware of a user's published content" (Murthy 2013).

The use of Twitter in China is limited. Twitter was not popular in China before being blocked. The first Chinese microblog was Fanfou.com, which was launched in May 2007. The number of Fanfou users was around one million in 2009. Largely because of riots that happened in certain parts of China, Twitter and Facebook were blocked in July 2009 and have been limited in access since then. Fanfou and some other microblogs were also blocked for a while in July 2009. Chinese IT giant Sina.com grasped this opportunity and launched Sina Weibo in August 2009 (1 month after Twitter was blocked). "Weibo" means "microblog" in Chinese and Sina registered weibo.com. Therefore, people usually use Weibo to refer to Sina Weibo. Sina had its unique marketing strategy – Sina invited celebrities to sign up to Sina Weibo and communicate with their fans. This strategy worked very well and Sina Weibo quickly became the largest microblog service in China. Within a year, Twitter's other Chinese equivalent, Tencent Weibo, NetEase Weibo, and Sohu Weibo, started to grow along with Sina Weibo. Sina Weibo's competitors also tried to pay some celebrities so that these celebrities would only use their service to post microblogs. However, the Sina Weibo community had already grown to a large number of users, who had also connected to their friends and families and constructed their networks and thus did not want to turn to another platform. Some celebrities even flew to Sina Weibo to be more visible. Therefore, although other Chinese microblogs have successfully built their own communities (which are also large scale), they do not really threaten Sina Weibo, which is still dominating the Chinese microblog world.

In the West, Facebook still has been growing since Twitter was born. People are using Facebook and Twitter for different purposes. However, in China, traditional SNSs quickly lost active users, and many of Chinese SNSs became ghost towns after Sina Weibo's launch. There is a sign that it may also happen for Sina Weibo 3 years after its birth. Tencent (the company who produced QQ and QZone) launched WeChat in 2011. WeChat was first a multimedia (text, voice, video) messaging software. However, Tencent soon added its SNS features "Moments" into WeChat. Moments is a user timeline similar to Facebook. WeChat now has over 400 million active users, and many Weibo users moved to WeChat. Although Moments of WeChat

is growing very fast, currently most WeChat users are still using it solely for messaging purposes. Therefore, we do not include it in the ranking of SNSs (Fig. 1 and Table 1). According to a report by GlobalWebIndex in January 2013, the number of active Weibo and traditional SNS (like Renren) users decreased significantly in 2012, when Twitter, Facebook, and Google+ were still increasing (http://www.pingwest.com/twitter-the-fastest-growing-social-platform/). This decline is likely attributed to the changing dynamic between WeChat and its competitors.

An early and highly cited empirical study of the topology and intention of Twitter was published in 2007, finding users use Twitter to talk about their daily activities and to seek/share information (Java et al. 2007). Since 2008, due to these unique features, Twitter and other microblog services have quickly become the key social media and SNS for not only in daily conversation and chats but for news reporting (i.e., discussing breaking news, report news, political elections), business (i.e., marketing, advertising), emergent events (i.e., disasters, protests, and terrorist attacks), and social movements (i.e., Occupy Movement, Arab Spring, civil wars) as well. The recent research on SNSs has largely focused on microblogs. Another reason that microblogs are now the key datasets for research is because it is easier to retrieve data as compared to other SNSs like Facebook and Renren. The two biggest microblogs Twitter and Sina Weibo both have open APIs that allow people to retrieve all kinds of data, usually with limits in the volume of data to be retrieved or the number of requests to the server. Kwak et al. analyzed a Twitter network of 41.7 million users, 1.47 billion social relations, and 106 million tweets with 4,262 topics and conducted a series of quantitative analyses on the data to reveal the difference between the Twitter network with other SNSs (Kwak et al. 2010). The research that has been done on traditional SNSs like Facebook and MySpace has been repeated with Twitter data, and more novel research has been conducted to answer many interesting research questions that could not be answered before. People have explored whether the information diffusion seen in Twitter was due to social connections or external resources, the roles of Twitter in information diffusion, the formulation and organization of groups in protest and revolutions, emerging distributed group chats on Twitter, and so forth (please refer to Recommended Reading section for corresponding publications). Currently, Twitter is the most frequently used data for researchers in social computing and computational social science, and Sina Weibo is playing the same role in Chinese academia.

Since 2004 (Crowdsourcing and Collective Intelligence)

Collective intelligence is defined as the intelligence emerged from the communication, collaboration, and competition of a group of individuals. The term was first coined by sociologists, who studied the swarm intelligence of insects, birds, mammals, bacteria, etc. (Lévy and Bonomo 1999; Bonabeau 2009). With the advances of SNSs, massive collaboration among a large number of users around the world has become a reality. People can collaborate online to work on the same task and solve problems. For example, Wikipedia is "a collaborative edited, multilingual, free Internet encyclopedia supported by the non-profit Wikimedia Foundation" (http://en.wikipedia.org/wiki/Wikipedia). The 30 million articles in 287 languages of Wikipedia were written by volunteers all over the planet. Anyone has access to edit almost every article of it (Glott et al. 2010). Wikipedia has been one of the top ten most popular websites according to Alexa (http://www.alexa.com/).

Online forums were the first big platform for collective intelligence. The use of online forums for collective intelligence ranges from small-scale Q&A systems (Zhang et al. 2007) to very large-scale "human flesh search" (a Chinese translation, in which "human flesh" refers to human empowerment; it has another name as crowd-powered search) (Wang et al. 2010; Zhang 2012; Zhang et al. 2012), in which a large number of voluntary Web users formed groups to collaborate on a single task. In 2006, Howe coined the term crowdsourcing and gave a definition of crowdsourcing as "the

act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in an open call" (Howe 2006). This definition covers most collective intelligence applications (particularly those for business), but it is a little too narrow to cover the large, voluntary, and loose organized crowd behaviors, like human flesh search, or crowdfunding sites like Crowdfunder and Kickstarter. Tarrell et al. reviewed 135 crowdsourcing-related articles from January 2006 to January 2013 (Tarrell et al. 2013). They found that research on crowdsourcing has been growing steadily. Researchers from computer science (CS) and information systems (IS) are the major contributors to this field. There are different focuses of CS and IS researchers. Generally speaking, CS researchers are mainly interested in modeling the collaboration of crowdsourcing and the design of better crowdsourcing systems (Zhang et al. 2007, 2010; Jurca and Faltings 2009; Pickard et al. 2011; Bozzon et al. 2013; Difallah et al. 2013), while IS researchers are more interested in the topics related to traditional IS research, like knowledge management, knowledge sharing, and incentives of contribution (Moon and Sproull 2008; Olivera et al. 2008; Mannes 2009; Bothner et al. 2011; Boudreau et al. 2011; Bayus 2013). There is also a trend of researchers from both sides are joining together to collaborate on crowdsourcing studies.

Since 2009 (Isolated Chinese SNSs)

As mentioned previously, as the country with the largest number of Internet users, China has blocked the access to some major SNSs including Facebook and Twitter (for consistency, we use "China" to refer Chinese Mainland only. Facebook and Twitter are popular in Hong Kong, Macau, and Taiwan, though a portion of people from these regions also use Chinese SNSs). However, Renren, Weibo, and other regional SNSs successfully took over the roles. In fact, these blocked SNSs did not perform very well in Chinese Mainland (as compared to their popularity in Taiwan) before they were blocked.

Although there is censorship upon Chinese cyberspace, and the "real-name policy" was recently applied to SNSs, SNSs are still among the freest platforms for Chinese Internet users to express their opinions (http://www.theatlantic.com/china/archive/2013/03/why-chinas-real-name-internet-policy-doesnt-work/274373/). Sometimes, the topics and keywords of users' discussions are seen censored and deleted automatically by SNSs, but users could generally find an alternative way to express the same meaning. There are countless "juicy stories" being generated by SNS users, in particular, Weibo users. The topics of their discussion are not quite the same as Twitter. Business people and brands have unique ways of marketing on Weibo and WeChat. For example, they create WeChat groups and push multimedia advertisements to users and communicate with users directly using the WeChat account. Rumors abound in the community. People collaborate to conduct "human flesh search" (Lu and Qiu 2013). The topics of users' arguments and fights can range from a tiny statement made by a celebrity or a TV program to serious economic or political issues.

Here, we present one example of the human flesh search (HFS) against corruption that aroused on Weibo. In 2012, a government official was photographed smirking after a tragic traffic accident. It enraged Chinese Internet users and this official's life quickly became under scrutiny. The HFS against him was started right away. Several photos of him started spreading in Weibo next day, and people quickly discovered 11 pricey watches he was wearing from these photos. Weibo users thought that there was no way that he could afford these watches on an honest government official's salary. The discoveries from HFS made the government start to investigate whether he was a corrupt official. Eventually, he lost his job and political career and is now under further investigation by judicial departments. The above story is an illustration of the Chinese Internet users using SNSs to do HFS for anti-corruption purposes. However, there are also some other examples, in which people violated the personal privacy of others. A complete analysis, in English, of HFS can be found in Zhang (2012).

There are also "Internet water armies" (paid Internet commentators) on Weibo (Zheng et al. 2011), for example, groups advertising for

a brand and attacking other brands, groups doing HFS and being HFSed, and groups criticizing or defending the government (there are mainly two major groups: (a) those who mainly criticize the government and would like a change and (b) those who defend the government and prefer a more stable society rather than a radical change). It was reported that there are government-/institute-/organization-funded Internet commentators trying to steer the public opinions towards the policies of governments (both within China and overseas) (http://news.bbc.co.uk/2/hi/asia-pacific/7783640.stm). Internet users called those defending government as "(RMB) 50 Cent Party" and those attacking the government (sometimes with fake rumors) as "(USD) 50 Cent Party," because the two "parties" are paid RMB 50 cent by the local government or USD 50 cent by a foreign government or institutes. These groups have been fighting each other on Weibo, and some groups were making up fault rumors to attack others and try to attract more Weibo users to support them (Fossato 2009; Bremmer 2010).

To regulate people's fights and control the existence of rumors, Sina Weibo proposed a credit system. In this system, each user has a credit score, people can sue others if they intentionally spread fake rumors, insult others, violate others' personal privacy, etc. If a user's score is low, he or she will be marked as a "low-credit user." A lot of such interesting things are happening in Chinese SNSs. However, most research on Chinese SNSs repeated the study of Twitter. How to distill interesting and unique research questions based on Weibo data and to properly answer them is a strongly promising and needed research.

Conclusion and Future Directions

In this article, we briefly review the history of SNSs worldwide. In particular, we describe the use of SNS in China, which has not been well covered by the literature in the West. SNS is still a rapidly evolving area, with new types of SNSs emerging and novel research directions being explored. Despite numerous powerful quantitative analysis methodologies developed, there are still a large number of unanswered research questions from theoretical social sciences. The link between computational sciences and social sciences could be much stronger with solid research, which answered key research questions derived from social theories. Another future research topic that we anticipate is the cross-cultural analytics of SNSs. Most researches to date have been focused on popular SNSs in the West, with datasets that mostly came from one SNS and a single country or language. What are the differences across different SNSs? How were multiple SNSs linked together? Are there any cultural differences in the behavior of people using SNSs? These research topics are expected to not only fill the holes of current literature, but also to shed light on an in-depth understanding of the use in different cultures. We hope that our review and discussions can help researchers and practitioners to get a brief overview of SNS to date and gain an outlook of future research directions on SNS.

Acknowledgments

This work was supported in part by a grant from the US Defense Advanced Research Project Agency Social Media in Communication Project (SMISC) and the US Army Research Laboratory Network Science CTA.

Cross-References

▸ Classical Algorithms for Social Network Analysis: Future and Current Trends
▸ Collective Intelligence, Overview
▸ Community Detection, Current and Future Research Trends
▸ Crowdsourcing and Human Computation, Introduction
▸ Crowdsourcing and Social Networks
▸ Dark Sides of Social Networking
▸ Data Mining
▸ East Asian Social Networks
▸ Futures of Social Networks: Where Are Trends Heading?

▸ Online Communities
▸ Privacy Issues for SNS and Mobile SNS
▸ Social Computing
▸ Social Media, Definition and History
▸ Social Networking Sites
▸ Structural Holes
▸ Topic Modeling in Online Social Media, User Features, and Social Networks for
▸ Web Communities Versus Physical Communities

References

Barabási A-L (2013) Network science. Philos Trans R Soc A 371: 20120375 http://www.barabasilab.com/pubs/CCNR-ALB_Publications/201302-18_RoyalSoc-NetworkScience/201302-18_RoyalSoc-NetworkScience.pdf

Bayus BL (2013) Crowdsourcing new product ideas over time: an analysis of the Dell IdeaStorm community. Manag Sci 59(1):226–244

Bishop CM, Nasrabadi NM (2006) Pattern recognition and machine learning. Springer, New York

Bonabeau E (2009) Decisions 2.0: the power of collective intelligence. MIT Sloan Manag Rev 50(2):45–52

Bothner MS, Podolny JM, Smith EB (2011) Organizing contests for status: the Matthew effect vs. the Mark effect. Manag Sci 57(3):439–457

Boudreau KJ, Lacetera N, Lakhani KR (2011) Incentives and problem uncertainty in innovation contests: an empirical analysis. Manag Sci 57(5):843–863

Boyd DM, Ellison NB (2007) Social network sites: definition, history, and scholarship. J Comput-Mediat Commun 13(1):210–230

Bozzon A, Brambilla M, Ceri S, Mauri A (2013) Reactive crowdsourcing. In: Proceedings of the 22nd international conference on world wide web, Rio de Janeiro. International World Wide Web Conferences Steering Committee

Bremmer I (2010) Democracy in cyberspace-what information technology can and cannot do. Foreign Aff 89:86

CNNIC (2013) 32nd statistical report on Internet Development in China. China Internet Network Information Center

Difallah DE, Demartini G, Cudré-Mauroux P (2013) Pick-a-crowd: tell me what you like, and i'll tell you what to do. In: Proceedings of the 22nd international conference on world wide web, Rio de Janeiro. International World Wide Web Conferences Steering Committee

Fossato F (2009) Web captives. Index Censorsh 38(3):132–138

Glott R, Schmidt P, Ghosh R (2010) Wikipedia survey–overview of results. United Nations University: Collaborative Creativity Group

Gonzalez R, Cuevas R, Motamedi R, Rejaie R, Cuevas A (2013) Google+ or Google-?: dissecting the evolution of the new OSN in its first year. In: Proceedings of the 22nd international conference on world wide web, Rio de Janeiro. International World Wide Web Conferences Steering Committee

Han J, Kamber M, Pei J (2006) Data mining: concepts and techniques. Morgan Kaufmann, San Francisco

Howe J (2006) The rise of crowdsourcing. Wired Mag 14(6):1–4

ITU (2013) Key ICT indicators for developed and developing countries and the world (totals and penetration rates). International Telecommunications Unions

Java A, Song X, Finin T, Tseng B (2007) Why we Twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on web mining and social network analysis, San Jose. ACM

Jurca R, Faltings B (2009) Mechanisms for making crowds truthful. J Artif Intell Res 34(1):209

Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? In: Proceedings of the 19th international conference on world wide web, Raleigh. ACM

Lazer D, Pentland AS, Adamic L, Aral S, Barabasi AL, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M (2009) Life in the network: the coming age of computational social science. Science (New York, NY) 323(5915):721

Lévy P, Bonomo R (1999) Collective intelligence: mankind's emerging world in cyberspace. Perseus Publishing, Cambridge

Lu J, Qiu Y (2013) Microblogging and social change in China. Asian Perspect 37(3):305–331

Mannes AE (2009) Are we wise about the wisdom of crowds? The use of group judgments in belief revision. Manag Sci 55(8):1267–1279

Manning CD, Schütze H (1999) Foundations of statistical natural language processing. MIT, Cambridge

Moon JY, Sproull LS (2008) The role of feedback in managing the Internet-based volunteer work force. Inf Syst Res 19(4):494–515

Murthy D (2013) Twitter: social communication in the Twitter age. Polity Press, Cambridge

Olivera F, Goodman PS, Tan SS-L (2008) Contribution behaviors in distributed environments. Manag Inf Syst Q 32(1):23

Pickard G, Pan W, Rahwan I, Cebrian M, Crane R, Madan A, Pentland A (2011) Time-critical social mobilization. Science 334(6055):509–512

Schuler D (1994) Social computing. Commun ACM 37(1):28–29

Shadbolt N, Hall W, Hendler JA, Dutton WH (2013) Web science: a new frontier. Philos Trans R Soc A: Math Phys Eng Sci 371 (1987)

Tarrell A, Tahmasbi N, Kocsis D, Tripathi A, Pedersen J, Xiong J, Oh O, de Vreede G-J (2013) Crowdsourcing:

a snapshot of published research. In: Proceedings of the nineteenth Americas conference on information systems, Chicago

Wang F-Y, Carley KM, Zeng D, Mao W (2007) Social computing: from social informatics to social intelligence. IEEE Intell Syst 22(2):79–83

Wang F-Y, Zeng D, Hendler JA, Zhang Q, Feng Z, Gao Y, Wang H, Lai G (2010) A study of the human flesh search engine: crowd-powered expansion of online knowledge. Computer 43(8):45–53

Zhang Q (2012) Analyzing cyber-enabled social movement organizations: a case study with crowd-powered search. Ph.D., The University of Arizona

Zhang J, Ackerman MS, Adamic L (2007) Expertise networks in online communities: structure and algorithms. In: Proceedings of the 16th international conference on world wide web, Banff. ACM

Zhang Q, Feng Z, Wang F-Y, Zeng D (2010) Modeling cyber-enabled crowd-powered search. In: The second Chinese conference on social computing, Beijing

Zhang Q, Wang F-Y, Zeng D, Wang T (2012) Understanding crowd-powered search groups: a social network perspective. PLoS ONE 7(6):e39749

Zheng X-L, Zhong Y-G, Wang F-Y, Zeng D-J, Zhang Q-P, Gui K-N (2011) Social dynamics research based on web information. Complex Syst Complex Sci 8(3): 1–12

Recommended Reading

Adamic LA, Glance N (2005) The political blogosphere and the 2004 US election: divided they blog. In: Proceedings of the 3rd international workshop on link discovery, Chicago. ACM

Albert R, Barabasi A-L (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97

Antal T, Krapivsky P, Redner S (2005) Dynamics of social balance on networks. Phys Rev E 72(3):036121

Bakshy E, Rosenn I, Marlow C, Adamic L (2012) The role of social networks in information diffusion. In: Proceedings of the 21st international conference on world wide web, Lyon. ACM

Barabási A-L (2002) Linked: the new science of networks. Basic Books, New York

Barabási A-L, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Bond RM, Fariss CJ, Jones JJ, Kramer AD, Marlow C, Settle JE, Fowler JH (2012) A 61-million-person experiment in social influence and political mobilization. Nature 489(7415):295–298

Bothner MS, Podolny JM, Smith EB (2011) Organizing contests for status: the Matthew effect vs. the Mark effect. Manag Sci 57(3):439–457

Burt RS (2009) Structural holes: the social structure of competition. Harvard University Press, Cambridge, MA, United States

Cha M, Mislove A, Gummadi KP (2009) A measurement-driven analysis of information propagation in the flickr social network. In: Proceedings of the 18th international conference on world wide web, Madrid. ACM

Chaney AJ-B, Blei DM (2012) Visualizing topic models. In: ICWSM, Dublin

Chang J, Boyd-Graber J, Blei DM (2009) Connections between the lines: augmenting social networks with text. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, Paris. ACM

Cheng X, Dale C, Liu J (2008) Statistics and social network of YouTube videos. In: 16th international workshop on quality of service, 2008 (IWQoS 2008), Enskede. IEEE

Conover M, Ratkiewicz J, Francisco M, Gonçalves B, Menczer F, Flammini A (2011) Political polarization on Twitter. In: ICWSM, Barcelona

Cook J, Kenthapadi K, Mishra N (2013) Group chats on Twitter. In: Proceedings of the 22nd international conference on world wide web, Rio de Janeiro. International World Wide Web Conferences Steering Committee

De Choudhury M, Sundaram H, John A, Seligmann DD (2009) What makes conversations interesting?: themes, participants and consequences of conversations in online social media. In: Proceedings of the 18th international conference on world wide web, Madrid. ACM

Diakopoulos NA, Shamma DA (2010) Characterizing debate performance via aggregated Twitter sentiment. In: Proceedings of the SIGCHI conference on human factors in computing systems, Atlanta. ACM

Dwyer C, Hiltz SR, Passerini K (2007) Trust and privacy concern within social networking sites: a comparison of Facebook and MySpace. In: AMCIS, Keystone

Ellison NB, Steinfield C, Lampe C (2007) The benefits of Facebook "friends:" social capital and college students' use of online social network sites. J Comput-Mediat Commun 12(4):1143–1168

Ghannam J (2011) Social media in the Arab world: leading up to the uprisings of 2011. Center for International Media Assistance/National Endowment for Democracy 3

Goetz M, Leskovec J, McGlohon M, Faloutsos C (2009) Modeling blog dynamics. In: ICWSM, San Jose

Harper FM, Moy D, Konstan JA (2009) Facts or friends?: distinguishing informational and conversational questions in social Q&A sites. In: Proceedings of the 27th international conference on human factors in computing systems, Boston. ACM

Hinds D, Lee RM (2008) Social network structure as a critical success condition for virtual communities. In: Proceedings of the 41st annual Hawaii international conference on system sciences, Hawaii. IEEE

Huberman B, Romero D, Wu F (2008) Social networks that matter: Twitter under the microscope. First Monday 14(1)

Jiang J, Wilson C, Wang X, Huang P, Sha W, Dai Y, Zhao BY (2010) Understanding latent interactions in

online social networks. In: Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, Melbourne. ACM
Kairam S, Brzozowski M, Huffaker D, Chi E (2012) Talking in circles: selective sharing in Google+. In: Proceedings of the SIGCHI conference on human factors in computing systems, Austin. ACM
Khondker HH (2011) Role of the new media in the Arab Spring. Globalizations 8(5):675–679
Kumar R, Novak J, Raghavan P, Tomkins A (2004) Structure and evolution of blogspace. Commun ACM 47(12):35–39
Lancichinetti A, Fortunato S, Radicchi F (2008) Benchmark graphs for testing community detection algorithms. Phys Rev E 78(4):046110
Larsson AO, Moe H (2012) Studying political microblogging: Twitter users in the 2010 Swedish election campaign. New Media Soc 14(5):729–747
Leskovec J, Huttenlocher D, Kleinberg J (2010a) Predicting positive and negative links in online social networks. In: Proceedings of the 19th international conference on world wide web, Raleigh. ACM
Leskovec J, Huttenlocher D, Kleinberg J (2010b) Signed networks in social media. In: Proceedings of the SIGCHI conference on human factors in computing systems, Atlanta. ACM
Lewis K, Kaufman J, Gonzalez M, Wimmer A, Christakis N (2008) Tastes, ties, and time: a new social network dataset using Facebook.com. Soc Netw 30(4):330–342
Lotan G, Graeff E, Ananny M, Gaffney D, Pearce I, Boyd D (2011) The revolutions were tweeted: information flows during the 2011 Tunisian and Egyptian revolutions. Int J Commun 5:1375–1405
Morris MR, Teevan J, Panovich K (2010) What do people ask their social networks, and why?: a survey study of status message Q&A behavior. In: Proceedings of the SIGCHI conference on human factors in computing systems, Atlanta. ACM
Myers SA, Zhu C, Leskovec J (2012) Information diffusion and external influence in networks. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, Beijing. ACM
Newman ME (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256
Newman ME (2006) Modularity and community structure in networks. Proc Natl Acad Sci 103(23):8577–8582
Qu Y, Huang C, Zhang P, Zhang J (2011) Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake. In: Proceedings of the ACM 2011 conference on computer supported cooperative work, Hangzhou. ACM
Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on world wide web, Raleigh. ACM
Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting elections with Twitter: what 140 characters reveal about political sentiment. In: ICWSM 10, Washington, DC, pp 178–185
Wang T, Zhang Q, Liu Z, Liu W, Wen D (2012) On social computing research collaboration patterns: a social network perspective. Front Comput Sci China 6(1):122–130
Watts DJ (1999) Small worlds: the dynamics of networks between order and randomness. Princeton university press, Princeton
Watts D, Strogatz S (1998) Collective dynamics of small-world networks. Nature 393:440–442
Yang Z, Wilson C, Wang X, Gao T, Zhao BY, Dai Y (2011) Uncovering social network sybils in the wild. In: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, Berlin. ACM
Zink M, Suh K, Gu Y, Kurose J (2008) Watch global, cache local: YouTube network traffic at a campus network: measurements and implications. In: Electronic imaging 2008. International Society for Optics and Photonics, San Jose, CA

Social Networking Services

▸ Social Networking on the World Wide Web

Social Networking Sites

▸ Ethics of Social Networks and Mining
▸ Friends Recommendations in Dynamic Social Networks
▸ Online Privacy Paradox and Social Networks
▸ Social Networking on the World Wide Web

Social Networks

▸ Counterterrorism, Social Network Analysis In
▸ Community Identification in Dynamic and Complex Networks
▸ Game-Theoretic Framework for Community Detection
▸ Game Theory and Social Networks
▸ Network Analysis in French Sociology and Anthropology
▸ Networks in Rural Sociology
▸ Sampling Effects in Social Network Analysis

▸ Social Networks in Emergency Response
▸ Social Networks and Politics
▸ Social Capital
▸ Social History of Computing and Online Social Communities
▸ Telecommunications Fraud Detection, Using Social Networks for
▸ Top Management Team Networks

Social Networks and Politics

Elena Pavan
Department of Sociology and Social Research, University of Trento, Trento, Italy

Synonyms

Collective action; Globalization; Governance; ICTs; Politics; Social networks

Glossary

Social Networks Specific type of social organization based on patterns of communication and exchange among actors involved

Politics A wide variety of dynamics aimed at the production of public purpose, from laws and regulations to norms of behavior

Definition

When thinking about the relationship that exists between *social networks* and *politics*, what comes immediately to mind is a whole set of events, ranging from the "Arab Spring" to manifestations of global refusal for the political and financial status quo as those of the *Occupy* movement, passing through the rise of Indignados in Spain and the mobilization in support of Barack Obama during the 2008 US presidential elections. Despite obvious differences in terms of actors involved, size, goals, geographies, identities, as well as in terms of the political dynamics played out, what these events had in common is that political objectives were pursued through the rapid construction of wide (trans)national networks of action which, in turn, were generated, sustained, and stimulated by a heavy use of networked information and communication technologies (ICTs) – Internet and social media such as Facebook, Twitter, and YouTube in the first place.

In fact, all the abovementioned episodes provide good examples of the strict nexus that exists between *politics* as *political participation* and *networks*, as both the *social network of activists* and *the networked communication infrastructure* that sustained political actions – i.e., the Internet. Indeed, there is a long-term reflection on the role of social networks in the fostering, structuring, and renewing of collective action dynamics (e.g., Diani 2003). Thus, the very deployment of dramatic situations like the Arab Spring invites to include systematically ICTs within these reflections and to consider them not only as tool for organizing but, more properly, as the real organizational milieu where contemporary mobilizations and campaigns develop (Bennett and Segerberg 2012).

And yet, the nexus between social networks and politics needs to be addressed from a wider perspective which includes, but is not limited to, widespread ICTs-enforced collective action instances. In fact, politics embraces a wide set of dynamics aimed at public purpose production, which is "an expression of vision, values, plans, policies and regulations that are valid for and directed towards the general public" (Sørensen and Torfing 2006). Moreover, social networks are first and foremost a specific type of social organization based on "reciprocal patterns of communication and exchange" (Powell 1990) which permeates all domains of society, not only politics.

Hence, reviewing the nexus between social networks and politics entails reflecting systematically on the changes of organizational modes in the conducts of politics at all levels, from the local to the global, and for the production of various types of public purpose,

from conventional government and regulatory acts (e.g., Knoke et al. 1996) to norms, i.e., cognitive frameworks that guide the action of institutional and noninstitutional actors (Finnemore and Sikkink 1998). Also, ICTs are to be considered as a crucial intervening element of the "multidimensional context" that shapes the patterns of political actions (Diani 2011) as they foster the construction of communication-based social relations and, in the end, provide "the means of political debate: the arena, the communication links, the agenda" (Bijker 2006).

Introduction

Addressing the nexus between social networks and politics is a complex task which requires, in the first place, an exploration of the background scenario which has led to the consolidation of networks as both metaphors to depict contemporary complex political arrangements and as a true organizational mode for the conducts of politics at all levels, from the local to the global. Also, this task requires some systematization effort aimed at clarifying the main traits of networked politics in its most popular declinations: as forms of collaboration of governments (*government networks*), as multi-actor arrangements for the definition and the implementation of policies (*policy networks*), and as the overall perspective to depict different instances of collective action (*collective action networks*).

Globalization, the Overcoming of Nation-State Politics, and the Role of Networks

One of the main features of the contemporary world, perhaps the most emphasized, is *interconnectedness*. Societies and economies today are linked in complex webs of interactions, influencing each other in non-trivial ways and both enhancing possibilities (let's just think of the worldwide solidarity response to the 2004 Tsunami in the Indian Ocean) as well as augmenting the reach of negative dynamics (as it is in the case of the global financial crisis).

Whether it is considered its cause or its consequence, interconnectedness is often related to the concept of globalization, which can be defined as a set of processes impacting the spatial organization of social relations and transactions "generating transcontinental or inter-regional flows and networks of activity, interaction and the exercise of power" (Held et al. 1999). Globalization processes have taken place in a variety of fields (economy, politics, culture, environmental concerns) and have contributed to the transformation of the world into a "shared social space" (Held et al. 1999), where traditional boundaries (whether these are territorial, thematic, or based on competences) are now blurred.

At the same time, globalization is presenting us with a number of challenging aspects. Looking in particular at the domains of politics, many issues (as the sustainable use of energetic resource, the control of financial markets or even the definition of national labor policies, etc.) are not any longer managed by single and rather homogeneous societies or economies embodied by the nation-state. Rather, they are now spanning a wide range of geographically distant and socioculturally heterogeneous constituencies and represent now global societal challenges characterized by their global scale and by features of *diversity* (of actors and perspectives), *dynamics* (the continuous evolution of issues at stake as well of perspectives upon them), and *complexity* (of the webs of interaction).

Hence, governments and institutions are pressured to intervene in a complex scenario where the distinction between domestic and foreign affairs is blurred and where multiple and diversified knowledge is required to keep under control all the facets of global problems. Shortcomings in facing these challenges have translated into a threefold deficit of *legitimacy*, *knowledge*, and *access* (Hockings 2006) which questions the traditional hierarchical nation-state model as the preferred governance mechanism. Thus, the recent global financial crisis has highlighted the limits of a regulation model based on free market assets.

Furthermore, the increased level of interconnectedness fosters the proliferation of

nontraditional political actors, (e.g., civil society organizations and coalitions, social movements, subcultures, single committed individuals, loose platforms for action) which constitute a plurality of different *publics*, all exerting control on the management of public affairs, possessing the required knowledge for the management of global challenges, proposing alternative solutions to current mechanisms deficiencies, and willing to take part into reformative dynamics and governance experiments to increase the effectiveness and the democratic features of political mechanisms.

Contemporary global settings call then for a "decentralized concept of social organization and governance [for which] society is no longer exclusively controlled by a central intelligence (e.g., the state); rather controlling devices are dispersed and intelligence is distributed amongst a multiplicity of action (or 'processing') units" (Kenis and Schneider 1991).

Here, networks enter as a powerful image to depict the growing complexity, but they also represent a truly new social morphology (Castells 2011): one for which policy outcomes and outputs are "generated within multiple-actors-set in which actions are interrelated in a more or less systematic way" (Kenis and Schneider 1991). Within an overall context of uncertainty, due to the shortcoming of conventional political mechanisms and to the difficulties of reorganizing steering activities so to include all actors and stakeholders (Börzel 1998), networks emerge to *incorporate, supply*, and *challenge* market and hierarchies as governance mechanisms for the production of public purpose (Kahler 2009).

Networks emerge then in response to the lack of a central authority able to set the widely accepted benchmarks for the conduct of public affairs. As a mode of (re)organizing political dynamics, they are based on cooperation (and yet allow for the development and management of conflicts), foster mutual learning and the spread of knowledge, allow a fast translation of knowledge into action, and, hence, are flexible enough to compensate the variability and the overall uncertainty of the future (Powell 1990). For their peculiarities, networks become then the preferred arrangement for sustaining contemporary governance efforts, i.e., for producing rules, norms, and, more broadly, the conditions for ensuring order through new strategies of problem-solving based on relationships between private and public actors that augment governing capacities.

Networked Politics as Forms of Communication Networks

For the strict link that exists between networks of sociopolitical actors and the conducts of politics in a globalized context, the very concept of networks has been applied in several ways. In general, networks in politics have been used to study both the emergence of coalitions within states, with a specific accent on resource mobilization and power redistribution, and the creation of interdependencies between states (Wellman 2002). Over time, labels have multiplied as to depict the variety of situations in which interdependency between political actors is experienced and managed. However, the application of a relational view for studying political transformations has not happened in consistent ways: similar situations have been labelled differently, the same label has been applied to different occurrences, and the underlying assumptions leading to the choice of a specific network concept over the other are seldom made explicit (Börzel 1998). The heterogeneity of uses somehow jeopardizes the heuristic potential of the network idea itself for the study of politics, and despite studies adopting a network point of view have multiplied in this field, an overall *consensus* on what networks mean for politics (a mere metaphor, a method, an analytic tool, or a proper theory) is still missing (Börzel 1998).

However, the heterogeneity of labels and of their uses is not a total impediment to a systematic overview of different conceptualizations of networked politics. In fact, all applications of the (social) network concept in the study of politics share the initial assumption that both the hierarchical nation-state and the market models present major shortcomings that hinder the achievement of satisfactory results. Because they

are not self-sufficient, states need to collaborate with other actors and to internalize the knowledge coming from these collaborations within policy-finding and policy-making processes. This creates an overall situation of interdependency between institutional and noninstitutional actors that is managed first and foremost through the establishment of *communication flows* from one actor to others. In this sense, all applications of the network concept to politics can be summarized through the idea of *communication networks* that join together actors mainly through the exchange of messages across time and space in the attempt to stabilize structures of interaction out of the chaos provided by the globalized context (Monge and Contractor 2003).

On these bases, we can distinguish between different types of political communication networks leaning on the elements that define networks as specific forms of social organizations, i.e., actors and relations. Looking then at which are political actors involved, how heterogeneous they are, and at why they interact, we can then make sense of different applications of the network concept in the study of politics.

Government networks are composed by national governmental and intergovernmental organizations officials with the overall aim of providing traditional political actors with the necessary *global reach* they miss in the contemporary globalized political milieu through their engagement and exploitation of flexible arrangements for collaboration (Slaughter 2004). Examples of such networks are the G-7 or the G-8 and the G-20 as well as the Asia-Pacific Economic Cooperation (APEC) or the Organization for Economic Co-operation and Development (OECD). Actually, these networks are not completely new phenomena, but at the present stage, their scale, scopes, and type of ties are undergoing an unprecedented growth.

Government networks are composed of homogeneous nodes, i.e., they are made of governmental and intergovernmental actors, who can be further differentiated on the bases of interests they carry (Slaughter 2004). Furthermore, government networks can be horizontal (aimed at exchanging information and best practices) or vertical (in which authority is delegated to a higher-level organization, e.g., in the field of justice with international courts). In mobilizing traditional political actors, Slaughter points out how government networks respond to the "governance tri-lemma" for which (a) contemporary political settings see the need for official regulatory activity at global level yet without centralization of power and ensuring accountability across different policy mechanisms, (b) governmental actors can and should interact with a multiplicity of non-governmental organizations that have emerged as important actors but (c) "their role in governance bears distinct and different responsibilities" (Slaughter 2004). In this context, government networks offer "a flexible and relatively fast way to conduct the business of global governance, coordinating and even harmonizing national government action while initiating and monitoring different solutions to global problems. Yet they are decentralized and dispersed, incapable of exercising centralized coercitive authority. Further (...) they can interact with a wide range of NGOs, civic and corporate, but their responsibilities and constituencies are far broader" (Slaughter 2004).

Policy networks is probably the most widely used label to describe a whole set of very different processes revolving around transformations of policy-making processes. In their seminal work, Marin and Mayntz tackle the problematic issue of identification of policy networks which, following their argumentation, "are explicitly defined not only by their structure as interorganizational arrangements, but also by their function – the formulation and the implementation of policy" (Marin and Mayntz 1991). Actors involved in collective decision processes might be of different nature, but their ability to enter the network varies depending on the porosity of the policy domain under discussion (i.e., the more uncertain the domain, the wider the constituency of actors involved).

Policy networks have been studied predominantly on a national scale, sometimes in comparative terms (Knoke et al. 1996) or loosely applied

to represent interdependence between public and private actors at global level. In being the most widespread label for depicting the nexus between social networks and politics, policy networks have been reviewed and classified in several ways (see Börzel 1998; Adam and Kriesi 2007). Overall, existing literature points out the use of this concept to identify, depending on the concrete case studies, structures for interest intermediation among actors; alternative governance structure challenging markets and hierarchies; multi-actor arrangements for policy implementation; or a "formalized, quantitative approach of social network analysis (. . .) that focuses on the relations between actors and not on actors' characteristics" (Adam and Kriesi 2007).

As a specific approach for studying policy-making activities through network analysis techniques, social network analysis of policy networks (e.g., Knoke et al. 1996) is mostly concerned with the redistribution of power along network ties, where the degree of power is proportional to the degree at which interests held by different actors involved are reflected through policy outcomes and not in relation to innate qualities. Concrete operationalizations of this relational view of power have translated into two types of studies: positional, which are primarily concerned with actors' positions within the network, and relational, concerned with characters and effects of relations existing between actors in a system (Lotan et al. 2011).

More recently, the idea of *governance network* has been proposed to expand the reach of the policy network approach also to the production of nonbinding policy outcomes, i.e., of norms (Sørensen and Torfing 2006). In this sense, as a sort of "second generation" of policy network studies, governance networks' studies are not so much focused on the actual existence of networks as distinct and legitimate forms of governance (Sørensen and Torfing 2006). Rather, they start from an explicit recognition of networks' existence and political meaning to model interactions thus keeping into account structural, processual, and cognitive elements. In this sense, governance networks can be defined as "(1) a horizontal articulation of interdependent, but operationally autonomous actors; (2) who interact through negotiations; (3) transpiring within a regulative, normative, cognitive and imaginary framework; (4) that to a certain extent is self-regulating; and (5) which contribute to the production of public purpose within a particular area" (Sørensen and Torfing 2006).

Although they are often studied in the context of policies production and coordination, the potential of governance network as analytical tools goes beyond conventional policy making to include "decision finding rather than decision making processes" (Hemmati 2002). In this sense, governance networks are the preferred label to study those political dynamics that are not necessarily finalized to the formulation of binding provisions but, rather, are aimed at the production of shared norms and knowledge (e.g., the United Nation World Summit on the Information Society or the Internet Governance Forum for the creation of a common vision between governments, private sector, and civil society; (see Pavan 2012)).

Collective action networks is a conceptual perspective based on social networks that has been pushed forward within the study of political participations and contentious politics to differentiate and underline the specificities of the diverse collective action instances: social movements, coalitions, organizational action, and communities/subcultures (Diani and Bison 2004; Diani 2008). According to this specific perspective, and in consistency with the premises of the structural approach to the study of politics, the accent is put on actors' interactions rather than on actors' features (e.g., the level of formalization of organizational assets, the sociodemographic characteristics of citizens who mobilize or participate politically). Thus, this perspective was elaborated in the first place to specify social movements in comparison to other forms of contentious politics or political participation (Diani 2008), but it can be adopted to study of all forms of collective political participation.

It is the combination of three different network characteristics that allows to distinguish between

Social Networks and Politics, Table 1 Typology of collective action networks

	Dense networks	Sparse networks
High collective identity	Social movements	Communities
Low collective identity	Coalitions	Organizational action

Source: Adaptation from Diani (2008)

different realizations of collective action: (i) the presence or absence of conflictual orientations towards clearly identified opponents, (ii) network density (sparse vs. dense networks), and (iii) presence of a strong or weak network collective identity.

While the presence of conflict refers more to specific repertoires of actions adopted within all types of collective actions, levels of density and of identity sharing are the two main axes along which instances of action can be distinguished (see Table 1).

Intensity of network identity divides dense networks into social movements, characterized by a strong identity, and coalitions, where collective identity is weak. Collective identity is important for social movements as it entails the presence of shared visions and values that sustain a long-term involvement over time and, in this sense, is what bonds different individuals and organizations, each of which with its own agenda, modes of behaviors, and perceptions, within the same mobilization effort over time (Melucci 1996). Thus, although social movements can be based on consensual repertoires, they are often coupled by a marked attitude towards conflict, as they rise as explicit expressions of social dissent towards identified opponents (Diani and Bison 2004).

When network identity is lower, dense networks of exchange between actors respond to instrumental and more short-term goals. Instrumentality of action is what characterizes coalitions in general (Gamson 1961), but it is worth noting that this is not tantamount to the lack of values or solidarity within coalitional processes. In fact, although coalitions lack a long-term vision, in their attempt to pursue a specific goal they can repeat over time, as it happens, for example, in the case of the campaign "16 days against VAW (violence against women)" every year from November 25 to December 10. Moreover, especially when coalitions are transnational and the goal they pursue is linked to a reform of societal assets, there is the need to supply instrumentality with shared views and values.

If looser networks are coupled with weak collective identity, the focus shifts to single organizations, while if they are associated to strong identities, they generate communities. Within specific organizations, such as Greenpeace, Sea Shepherd, or Oxfam, action is carried on very much following the agenda and the modus operandi of the single organization, i.e., under an organizational (rather than collective) identity (Diani 2008; Diani and Bison 2004). Thus, participation to action is consequential to the ownership of established membership criteria (e.g., all sorts of eligibility conditions from having paid a fee to possessing some specific skills or competences). Differently from social movements and communities, which join together a plurality of organizations under widely shared frames and beliefs systems, organizational collective action is characterized by a specific entrenchment within the boundaries of the organization itself, which is responsible for determining how the mobilization is carried on. Conversely, social movements, as well as communities, are "multicentric" as none can claim to represent the totality of the network (Diani 2008).

Communities instead carry on collective activity through networks which are sparse and yet are characterized by a shared sense of belonging diffused among members. Here, the idea of community can be detached from that of territoriality (as it is, instead, within classical sociology) and should be rooted in the shared practices and views thus blending the networked structure of mobilization within daily activities, which are conducted following the very values and ideals that jointly define the collective identity (Diani 2008).

Future Directions

For communication is the very backbone of networked politics, developments in the ICTs field have a profound impact on network arrangements in the field. Indeed, in a context of total embeddedness of ICTs in all domains of human action, the social relational infrastructure overlaps with the technical and physical infrastructure generating socio-technical systems (Vespignani 2009), making the distinction between the online and the offline obsolete and the space for social and political action hybrid, nurtured by relations that are built across the boundary between the virtual and the real world.

When it comes to the study of politics, this socio-technical breakthrough implies the difficulty of assessing the role of ICTs, and of social media in particular, in relation to the overall set of political transformations outlined above. The emphasis put by both academics and mainstream media on Facebook or Twitter in commenting events like the Arab Spring or the gatherings of the Occupy movement somehow biased our understanding of these dynamics as if, in the absence of social media, these mobilization episodes could have never happened. In fact, in spite of the overall enthusiasm for the "Revolutions through the Internet," critical approaches to the study of collective action transformations invite us to reflect on the fact that ICTs are not *the cause* of collective action but they remain crucial in determining its shape and forms (Diani 2011). This *caveat* can be easily generalized to the totality of political dynamics beyond the domain of collective action: the nexus between social networks and politics is deepened, even radicalized by the presence and diffusion of ICTs, but it is not *caused* by them. The main issue, then, is how to systematically explore, both conceptually and empirically, the shape, the form, and the consequences of networked politics in the ICTs era.

At present, research activities are growing rapidly and, yet, along two parallel tracks. On the one hand, there are theoretical attempts to properly outline the implications and the very defining features of contemporary forms of political dynamics: from the transformations of collective action into "connective action" (Bennett and Segerberg 2012), to the exploration of genres and repertoires of action that are made possible by an extensive use of the Internet (Lievrouw 2011), to the transformations of supra-national politics towards multi-actor governance arrangements though the construction of offline and online networks of collaboration (Pavan 2012), to the redistribution of power along communication network ties (Castells 2011). On the other hand, there are attempts to empirically investigate network structures sustaining political dynamics, whether these are generated within online forum discussions (González-Bailón et al. 2010), by protest participation through Twitter use (Lotan et al. 2011), by websites pertaining to a certain issue (Pavan 2012), or by mailing list exchanges (Pavan 2012). In these studies, network analysis techniques are often adopted and complex data-retrieval procedures are enacted. Thus, the exploration of political network structures is done in search for mechanisms such as contagion, dyad or triad emergence, and triadic closure effects but also looking at the emergence from network interactions of specific semantics and shared frames.

However, a full integration between the theoretical and the empirical levels appears to be still missing. The exploration of network properties is seldom tied to theoretical considerations on the forms and the effects of networked politics, while sophisticated theoretical models are rarely applied to actual data. In this sense, the current state of research on the nexus between social networks and politics seems to reproduce the fracture between "hard" and "soft" applications of the network idea to the study of politics which has always characterized the field (see above).

In the attempt to recompose this fracture, research activities in the field should be carried on in an integrated manner, joining together considerations on the communicative and relational potential of ICTs, actual exploration of networks' features, and existing knowledge on the transformations of political arrangements in the globalized society. In the first place, this integration requires avoiding to maintain the distinction between online and offline. As we live

in socio-technical systems where the Internet is perhaps the most diffused physical infrastructure upon which we create social relation, not only the social space should be considered hybrid, but no hierarchy or solution of continuity should be imposed between the online and the offline dimensions. In this sense, online relationships *do not substitute* but, rather, *integrate* the relational capital established by actors offline. Hence, in evaluating how the continuous developments of communication technologies affect the creation and the functioning of networked forms of political action, the focus should be set on the *totality* of social relations, whether they are grounded in face-to-face interactions or ICTs mediated.

Second, and in connection to this first point, if we are to make sense of the potential of communication technologies for the creation of politically relevant relations at all levels, monolithic conceptualizations of ICTs (and of the Internet in particular) should be avoided. Not only is there a technical difference between an Internet populated by websites and an Internet crowded by individually generated contents that are then put into global circuits of information transmission and communication. More than this, the passage from a Web 1.0 to a Web 2.0 entailed the passage from a culture of publicity to a culture of participation which is simply germane, almost a precondition, for the overcoming of the nation-state models of politics. However, within the vast realm of Web 2.0, multiple ways of communicating and participating are available: the adoption of one tool in spite of another has consequences on the very structure of the resulting communication networks and, hence, on how networked politics are enacted. In this sense, efforts should be directed towards the identification of the communicative potential that is proper of different social media for the production of different types of public purpose.

Finally, the fruitfulness of future research activities on the nexus between social networks and politics depends to a large extent on interdisciplinarity. There is a relationship of mutual influence between the two infrastructural levels that are present within socio-technical systems, i.e., the social and the technical one. Expertise on social networks as specific forms of social organization, hence on networked forms of politics as alternative modes for organizing and conducting governance practices, should meet technical expertise on communication systems so to balance considerations on the necessities for reforming existing political assets with a systematic knowledge on the very way in which communication technologies are shaped, as their structure and configuration set the overall boundary for the establishment of relations. In this sense, the network approach should not only be the preferred instrument to *conduct research* on how networked politics transform but, more broadly, the very way of reorganizing research practices, fostering a global interconnectedness of disciplinary knowledges across traditional, but now obsolete, boundaries.

Cross-References

- ▸ Community Evolution
- ▸ E-government
- ▸ Mapping Online Social Media Networks
- ▸ Online Communities
- ▸ Policy Networks: History
- ▸ Political Networks
- ▸ Social Capital
- ▸ Social Networking in Political Campaigns
- ▸ Web Communities Versus Physical Communities

References

Adam S, Kriesi H (2007) The network approach. In: Sabatier PA (ed) Theories of the policy process. Westview, Boulder, pp 129–154

Bennett L, Segerberg A (2012) The logic of connective action. Inf Commun Soc 15(5):739–768

Bijker WB (2006) Why and how technology matters. In: Goodin RE, Tilly C (eds) The Oxford handbook of contextual political analysis. Oxford University Press, Oxford, pp 681–706

Börzel T (1998) Organizing Babylon. On the different concepts of social networks. Public Adm 76:253–273

Castells M (2011) A network theory of power. Int J Commun 5:773–787

Diani M (2003) Networks and social movements: a research programme. In: Diani M, McAdam D (eds)

Social movements and networks. Relational approaches to collective action. Oxford University Press, Oxford, pp 299–319
Diani M (2008) Modelli di azione collettiva: quale specificità per i movimenti sociali? Partecipazione e Conflitto 1:43–66
Diani M (2011) Networks and internet into perspective. Swiss Political Sci Rev 17(4):469–474
Diani M, Bison I (2004) Organizations, coalitions and movements. Theor Soc 33:281–309
Finnemore M, Sikkink K (1998) International norm dynamics and political change. Int Organ 52(4):887–917
Gamson WA (1961) A theory of coalition formation. Am Sociol Rev 26(3):373–382
González-Bailón S, Kalterbrunner A, Banchs RE (2010) The structure of political discussion networks: a model for the analysis of online deliberation. J Inf Technol 25(2):230–243
Held D, McGrew A, Goldblatt D, Perraton J (1999) Global transformations: politics, economics, and culture. Polity, Cambridge
Hemmati M (2002) Multi-stakeholder processes for governance and sustainability: beyond deadlock and conflict. Earthscan, London
Hockings B (2006) Multistakeholder diplomacy: forms, functions and frustrations. In: Kurbaljia J, Katrandjiev V (eds) Multistakeholder diplomacy. Challenges and opportunities. Diplo Foundation, La Valletta, pp 13–32
Kahler M (2009) *Introduction-Networked politics, agency, power and governance*. In Kahler M (ed) Networked politics, agency, power and governance. Cornell University Press, Lthaca and London, pp 1–20
Kenis P, Schneider V (1991) Policy network and policy analysis: scrutinizing a new analytical toolbox. In: Marin B, Mayntz R (eds) Policy networks. Empirical evidence and theoretical considerations. Westview, Boulder, pp 25–62
Knoke D, Pappi FU, Broadbent J, Tsujinaka Y (1996) Comparing policy networks. Cambridge University Press, Cambridge
Lievrouw LA (2011) Alternative and activist media. Polity, Cambridge
Lotan G, Graeff E, Ananny M, Gaffney D, Pearce I, Boyd D (2011) The revolutions were tweeted: information flows during the Tunisian and Egyptian revolutions. Int J Commun 5:1375–1405
Marin B, Mayntz R (1991) Introduction: studying policy networks. In: Marin B, Mayntz R (eds) Policy networks. Empirical evidence and theoretical considerations. Westview, Boulder, pp 1–24
Melucci A (1996) Challenging codes: collective action in the information age. Cambridge University Press, Cambridge
Monge P, Contractor N (2003) Theories of communication networks. Oxford University Press, Oxford
Pavan E (2012) Frames and connections in the governance of global communications. A network study of the internet governance forum. Lexington, Lanham
Powell W (1990) Neither market nor hierarchy: network forms of organization. Res Organ Behav 12:295–336
Slaughter A (2004) A new world order. Princeton University Press, Princeton
Sørensen E, Torfing J (2006) Introduction: governance network research: towards a second generation. In: Sørensen E, Torfing J (eds) Democratic network theories. Palgrave Macmillan, London, pp 1–24
Vespignani A (2009) Predicting the behavior of techno-social systems. Science 325:425–430
Wellman B (2002) Structural analysis: from method and metaphor to theory and substance. In: Scott J (ed) Social networks. Critical concepts in sociology, vol. I. Routledge, London/New York, pp 81–122

Social Networks for Quantified Self

Ted Vickey and John Breslin
Digital Enterprise Research Institute, National University of Ireland at Galway, Galway, Ireland

Glossary

Connected Health Health care through the use of technology
mHealth Mobile Health
Mobile Fitness Apps Mobile Fitness Applications used from a smartphone or website

Definition

Over three quarters of US health care spending goes to the care of people with chronic conditions, including heart disease, diabetes, and asthma, while in 2004, nearly half of the Americans were diagnosed with one or more chronic conditions, a number expected to increase dramatically as the baby boomer generation rapidly approaches their retirement age (Accenture 2009). The new reality, dubbed "Connected Health," incorporates a broad range of health and fitness applications that are always on, always active, and always aware(Accenture 2009).

Since many aspects of health promotion professionals involve interdependent actors, social networks are of increasing interest to health

services researchers (O'Malley and Marsden 2008). The creation of a social network map of a person's social network can help visualize and thus better understand the strengths of the social ties of the network (Christakis and Fowler 2009).

Technology Will Transform the Future of Chronic Care

In a 1995 editorial in the American Journal of Public Health, former US Surgeon General C. Everett Koop stated, "Cutting-edge technology, especially in communication and information transfer, will enable the greatest advances yet in public health. Eventually, we will have access to health information 24 hours a day, 7 days a week, encouraging personal wellness and prevention, and leading to better informed decisions about health care" (Koop 1995). Technologies like miniaturized health sensors, broadband networks, and mobile devices are enhancing and creating new health-care capabilities such as remote monitoring and online care (Accenture 2009).

In 2009, management and technology consultant Accenture released a report on how technology will transform the future of chronic care. Cited in the report is the anticipated crisis in care that will be further challenged as the baby boomer generation begins to retire.

> According to the US Census Bureau, the world's population of people age 65 and older is projected to triple by mid-century, from 516 million in 2009 to 1.53 billion in 2050. This growing trend places a tremendous economic burden on governments, private employers and individual consumers alike. It also puts strain on the capacity of skilled care professionals and nursing homes. (Accenture 2009)

In addition to the inexpensive cost of computers and Internet connectivity, the report identifies three technological advancements that are paramount to the future of chronic care:

- Seamless capture and sharing of patient information in real-world settings
- Improvements in ways to combine and interpret data about an individual's health and wellness so that appropriate interventions can be made before an acute situation occurs
- Innovative tools including user modeling, advanced visualization, decision support, and collaboration

Health and Social Networking

One aspect of "Connected Health" is via the power of a person's social network. Research suggests that people interact with their social network with regard to their health. Christakis and Fowler (2009) concluded that "… a person with more friends and social contacts generally has better health than a person with fewer friends, and a person at the center of a network is more susceptible to both the benefits and risks of social connection than those at the periphery of a network." This would suggest that a person is not only affected by their location in a social network but also influenced by the behaviors of those who are "close" to them in the network. Perceived social support and physical activity are directly associated with a person's perceived health status (Almeida 2008).

As technology continues to impact humanity, the understanding of one's social network may be one key to better health. The basic element of a person's social network is simple: a social network starts with a central person (called an ego) and other people (called nodes) that are interconnected by links (called ties). As the numbers of nodes and links increase, the number of possible connections grows exponentially – known as the network effect (Christakis and Fowler 2009).

Christakis and Fowler (2009) suggest that "people are inter-connected and so their health is inter-connected. Inter-personal health effects in social networks provide a new foundation for public health." As online connections between people become ever more interweaved with offline real-world interests, social networking methods are moving towards simulating real-life social interactions, including physical activity, health, and disease management: rather than randomly approaching each other, people meet through things they have in common (Breslin and Decker 2007).

Technology and Health Behavior Modification

By using Mobile Health technology (mHealth), health providers can practice a more "personalized medicine" and potentially reach more individuals with effective health-related advice and information at a very low cost (Strecher 2007). Griffiths et al. (2006) suggest a number of reasons for delivering web-based health, wellness, and fitness interventions including reduced delivery costs, convenience to users, timeliness, reduction of stigma, and reduction of time-based isolation barriers.

Technologies can play three roles with regard to behavior modification: as tools, as media, and as social actors.

- As a tool, interactive technologies can be persuasive by making target behavior easier, leading people through a process, or performing calculations/measurements that motivate.
- As a medium, interactive technologies can be persuasive by allowing people to explore cause-and-effect relationships, providing people with experiences that motivate, or helping people to rehearse a behavior.
- As a social actor, interactive technologies can be persuasive by rewarding people with positive feedback, modeling a target behavior or attitude, and providing a social network of support (Fogg 2002).

Within the health-care field, interactive technologies can be effectively deployed to take on multiple roles at the same time. For example, a simple persuasive tool can measure calories while at the same time giving a reward upon attainment of a personal goal. This type of self-monitoring is a key ingredient in successful behavioral modification. In addition, if several people are connected through the Internet, then social support can be leveraged, which has been shown to impact motivation and behavior change (Chatterjee and Price 2009).

The Quantified Self

The idea of measuring things relative to a business or personal goal is common in today's society. The same measurement tools can be used within the self-tracking of a person's health and fitness. Commonly known as the Quantified Self movement, this is eclectic mix of early adopters, fitness fanatics, technology evangelists, personal development junkies, hackers, and patients suffering from a wide range of health challenges (The Quantified Self – Counting Every Moment 2012). Some measure their hourly mood swings, while others the stages of their nightly sleep habits. Some track every meal, snack, or drink, while others share on Twitter and Facebook their workout routine complete with heart rate, time, distance, calories burned, and musical preferences.

Ongoing research aims to classify and understand why a person shares their workouts within their social network via Twitter and the associated benefits. While there are various personal devices that monitor/track a person's exercise characteristics (e.g., Body Media, Fitbit, MapMyFitness, and Nike+), the effectiveness of online sharing via social networks of one's physical activity is limited in scientific research. Studies have indicated that "lack of motivation" is a key factor in why a person does not exercise.

One factor to address is the relationship between participant and provider (i.e., personal trainer) and/or participant and social network, including their influence. People join gyms not only for health and fitness but also for the social atmosphere. To fully understand the power of combining social networking and exercise adherence, the physical barrier of the four walls of an exercise facility is removed, and technology is used that enables a measurable improvement towards one's fitness goals.

Conclusion

With the move towards making machine-understandable data available for computers, allowing exercise data to become accessible/exchangeable between trusted peers is quite important. However, one's historical exercise records are often locked in to proprietary systems. By publishing selected aspects of these profiles using semantic terms, it will become easier

for people to search for and discover relevant exercise regimes.

Early prevention and healthy lifestyles may be the least expensive and best ways to combat the growing prevalence of avoidable diseases associated with a lack of physical activity including obesity (Almeida 2008). If people who lead sedentary lives would adopt a more active lifestyle, there would be enormous benefit to the public's health and to individual well-being. An active lifestyle does not require a regimented, vigorous exercise program. Instead, small changes that increase daily physical activity will enable individuals to reduce their risk of chronic disease and may contribute to enhanced quality of life (Pate et al. 1995).

Cross-References

▸ Actionable Information in Social Networks, Diffusion of
▸ Data Mining
▸ Twitter Microblog Sentiment Analysis

References

Accenture (2009) Always on, always connected: how technology will transform the future of chronic care. Accenture, New York

Almeida F (2008) The relationship between social networks, social support, physical activity and self-rated health: an exploratory study. University of Denver, Boulder

Breslin J, Decker S (2007) The future of social the need for semantics. IEEE Internet Comput 5:86–90

Chatterjee S, Price A (2009) Healthy living with persuasive technologies: framework, issues, and challenges. J Am Med Inform Assoc (JAMIA) 16(2):171–178 doi:10.1197/jamia.M2859

Christakis NA, Fowler JH (2009) Social network visualization in epidemiology. Health Care 19(1):5–16

Fogg B (2002) Persuasive technology: using computers to change what we think and do. Ubiquity Dec 2002, 5. doi:10.1145/763955.763957

Griffiths F, Lindenmeyer A, Powell J, Lowe P, Thorogood M (2006) Why are health care interventions delivered over the internet? A systematic review of the published literature. J Med Internet Res 8(2):e10. doi:10.2196/jmir.8.2.e10

Koop CE (1995) A personal role in health care reform. Am J Public Health 85(6):759–760. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1615490&tool=pmcentrez&rendertype=abstract. Accessed 13 June 2012

O'Malley A, Marsden P (2008) The analysis of social networks. Health Serv Outcomes Res Methodol 8(4):222–269

Pate RR, Pratt M, Blair SN, Haskell WL, Macera CA, Bouchard C, Buchner D et al (1995) Physical activity and public health: a recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine. JAMA 273(5):402–407. doi:10.1001/jama.1995.03520290054029

Strecher V (2007) Internet methods for delivering behavioral and health-related interventions (eHealth). Annu Rev Clin Psychol 3:53–76. doi:10.1146/annurev.clinpsy.3.022806.091428

The Quantified Self – Counting Every Moment (2012) Economist. Retrieved from http://www.economist.com/node/21548493. Accessed 15 Sept 2012

Social Networks in Emergency Response

Dashun Wang[1, 2], Yu-Ru Lin[3, 4], and James P. Bagrow[5, 6]
[1]Center for Complex Network Research, Northeastern University, Boston, MA, USA
[2]Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, MA, USA
[3]College of Computer and Information Science, Northeastern University, Boston, MA, USA
[4]Institute for Quantitative Social Science, Harvard University, Cambridge, MA, USA
[5]Engineering Sciences and Applied Mathematics, Northwestern University, Evanston, IL, USA
[6]Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL, USA

Synonyms

Collective response; Communication Network; Data mining; Disaster; Emergency; Event detection; Spatiotemporal analysis; Social networks

Glossary

Emergency An unexpected and often dangerous situation, typically affecting multiple individuals and requiring immediate action

Social and Communication Networks Networks of people interacting with each other through web-based (e.g., Twitter) and mobile-based (e.g., mobile phone) technologies

Social Media Web-based tools that enable people to communicate and interact with each other in various media forms including text and multimedia. Examples of these tools include emails, instant messengers (IM), blogs, microblogs (e.g., Twitter), vlogs (e.g., YouTube), podcasts, forum, wikis, social news (e.g., Digg), social bookmarking (e.g., Delicious), and social networks (e.g., Facebook, MySpace, and LinkedIn)

Definition

Modern datasets derived from telecommunication technologies such as online social media and mobile phone systems offer a great potential to understand the behaviors of large populations during emergencies and disasters. This entry reviews recent studies using large-scale, modern data to understand emergency and disaster response, covering work focused on social network activity during earthquakes and disease outbreaks and mobile phone communications following bombing and other emergency events. The key techniques and research trends are also discussed.

Introduction

Large-scale emergencies and disasters are an ever-present threat to human society. With growing populations and looming threat of global climate change, the numbers of people at risk will continue to grow. Thus there is a great need to optimize response efforts from search and rescue to food and resource disbursement. Human dynamics research offers a promising avenue to understand the behaviors of large populations, and modern datasets derived from cutting-edge telecommunications such as online social media and pervasive mobile phone systems bring a wealth of potential new information. Such massive data offers a promising complement to existing research efforts in disaster sociology, which primarily focus on eyewitness interviews, surveys, and other in-depth but small-scale data (Rodríguez et al. 2006).

Yet most current human dynamics research is focused primarily on data collected under normal circumstances, capturing baseline activity patterns. Here we review a number of studies pushing the envelope of modern data into the realm of unexpected deviations in these population behaviors. We discuss research focused on massive datasets from social network activity during earthquakes and disease outbreaks to mobile phone communications following bombings, power outages, and more.

We review a number of recent studies using large-scale, modern data to understand emergency and disaster response. We begin with a review quantifying how expectations of communication in today's world may influence our perception of the severity of an emergency. We then cover works focused on social media and mobile phones. These works use Twitter, a prominent online social media service, to understand more about disease outbreaks and the impact of earthquakes. The mobile phone studies feature a number of emergencies, including earthquakes, bombings, and a plane crash. The results of these studies have the potential to revolutionize disaster response in the future, with the critical goal of saving lives.

Historical Background

Connectivity and information access through global telecommunications have become increasingly pervasive due to modern technologies such as mobile phones and the Internet. People are becoming increasingly reliant on these communication modes and so an important question asked by Sheetz et al. (2010) is as follows: what do people expect about their

access to these communication channels when an emergency occurs? They explored how the expectation of the availability of these communication technologies may influence their perceptions of how they would use these technologies during and after a crisis.

To answer this question, the authors conducted online surveys and follow-up interviews with Virginia Tech students, faculty, and staff (participants). This university suffered a tragic attack on April 16, 2007, and the authors reported that local cellular networks were overwhelmed by traffic. Surveying witnesses and survivors at the university allows the authors to study how the perceptions of information access meshed with the unfortunate events that occurred.

Through these surveys and interviews, they found that participants have a range of expectations for connectedness in normal activities. Most participants did not expect to be able to immediately contact someone. This held even for strong social ties, for example, a student trying to reach his or her parents. Most importantly, the authors discovered that participants who do have high expectations of connectivity (and also tend to be more extroverted individuals) were more likely to report problems with connectivity than users with lesser expectations. These problems can lead these people to form overestimate of the severity of the crisis, compared with individuals who have lower expectations for their communication and are thus less likely to find communication loss a cause for concern. This means that an individual's personal traits may directly influence how he or she estimates the severity of a crisis.

While the authors admitted that they had a small sample size and that their interview methods may not be perfect, this study is an important step towards further understanding the interplay between modern telecommunications and emergency events.

Emergencies and Social Media

Today, social media such as Twitter and Facebook have been popularly used as everyday communication tools. Millions of people use "tweets" or Facebook "statuses" to inform family, friends, colleagues, or any others about information, opinion, and emotions about events just happening, leading to the great potential of using social media for monitoring and rescue purposes. Twitter allows users to send and receive tweets (140-character messages) via text messages and Internet-enabled devices, providing the public with detailed anecdotal information about their surroundings. Given the real-time nature of Twitter and the emerging social networking technologies, social media has the potential to fundamentally alter our discussions of emergencies. We briefly review some of the recent work on detecting disease outbreaks and earthquake response with Twitter.

Twitter and Disease Outbreaks

Various studies have shown the potential of using Twitter data to monitor the current public health status of a population, as people often tweet when they feel ill or recognize disease symptoms. Quincey and Kostkova (2010) collected tweets that contained instances of the keyword "flu" in a week during the swine flu pandemic. Their study suggests that the copresence of other words in tweets can be used by public health authorities to gather information regarding disease activity, early warning, and infectious disease outbreak. For example, in the majority of the collected tweets, the word "swine" was present along with "flu"; the words "have flu" and "has flu" may indicate that the tweet contains information about the users or someone else having flu. The words "confirmed" and "case(s)" perhaps indicate a number of tweets that are publicizing "confirmed cases of swine flu." Culotta (2010) collected over 500,000 influenza-related tweets during 10 weeks and analyzed the correlation between these messages and the Centers for Disease Control and Prevention (CDC) statistics. The paper reported a correlation of 0.78 by leveraging a document classifier. Chew and Eysenbach (2010) collected over 2 million tweets containing the keywords "H1N1," "swine flu," and "swinflu" within 8 months in 2009. Using manual and automated content coding, they found temporal correlation of Twitter activity with major news

stories and H1N1 incidence data. In addition, they found that the majority of these tweets contained resource-related posts (e.g., links to news websites). Gomide et al. (2011) analyzed how the dengue outbreaks in 2009 were mentioned on Twitter. Using a linear regression model, they showed promising results to predict the number of dengue cases by leveraging tweet content and spatiotemporal information. Signorini et al. (2011) tracked time-evolving public sentiments about H1N1 or swine flu and studied the probability of using Twitter stream for real-time estimation of weekly influenza-like illness (ILI) statistics generated by CDC.

There has also been work addressing the technical challenges of collecting tweets that are related to health or disease. Zamite et al. (2011) proposed a system architecture for collecting and integrating epidemiological data based on the principles of interoperability and modularity. Prier et al. (2011) proposed using a Latent Dirichlet Allocation (LDA) model to effectively identify health-related topics in Twitter. Paul and Dredze (2011) collected two billion tweets related to illness, disease symptoms, and treatment from May 2009 to October 2010. They proposed a probabilistic aspect model to separate tweets related to health from unrelated tweets. Aramaki et al. (2011) collected 300 million tweets from 2008 to 2010. They applied the Support Vector Machines (SVMs) to find tweets related to influenza with a correlation of 0.89 % compared with Google Flu Trends (Ginsberg et al. 2008). These tools offer the means to transform the overwhelming flood of big data into more manageable information.

Besides social media, there are also other solutions to estimate a population's health from Internet activity, most notably Google Flu Trends service, which correlates search term frequency with influenza statistics reported by the CDC (Ginsberg et al. 2008).

Twitter and Earthquakes

In recent years, tremendous effort has been made towards leveraging Twitter to study earthquakes, mainly falling into two lines of research: real-time detection (Sakaki et al. 2010; Guy et al. 2010; Earle et al. 2012) and crisis management (Hughes and Palen 2009; Caragea et al. 2011; Li and Rao 2010; Mendoza et al. 2010).

Early earthquake detection and the delivery of timely alerts is an extremely challenging task. Depending on peculiarities of the earthquake, from size to location, alerts may take between 2 and 20 min to publish, owing to the propagation time of seismic energy from the epicenter to seismometers and the latencies in data collection and validation. Therefore, it has been practically impossible for affected populations to know about an earthquake before it arrives. This situation is changing, however, thanks to the pervasive use of Twitter. Users submit their tweets via text messages and Internet-enabled devices, and these messages are available to their followers and the public within seconds, making Twitter an ideal environment for the dissemination of breaking news to large populations. Therefore, by using populations as social sensors, Twitter may be a viable tool for rapid assessment, reporting, and potentially real-time detection of a hazard event. Sakaki et al. (2010) investigated events such as earthquakes and typhoons in Twitter and proposed an algorithm to monitor tweets and to detect earthquakes. They extracted features such as keywords in a tweet by semantic analysis and used Support Vector Machines (SVMs) to classify a tweet into a positive or negative class. By regarding a tweet as a social sensor associated with location information, the authors transformed the earthquake detection problem into an object detection problem in ubiquitous and pervasive computing. They derived a probabilistic model by applying Kalman filtering and particle filtering to estimate the epicenter of an earthquake and the trajectories of a typhoon. They then deployed an earthquake reporting system in Japan, which delivers earthquake notifications to their users faster than the announcements broadcast by Japan Meteorological Agency. Meanwhile, researchers from the US Geological Survey (USGS) reported an earthquake detection system that adopts social network technologies, called Twitter Earthquake Detector (TED) (Guy et al. 2010; Earle et al. 2012). They downloaded tweets that con-

tain the words "earthquake," "gempa," "temblor," "terremoto," or "sismo" from August to the end of November 2009. Based on tweet-frequency time series, they used a short-term-average, long-term-average algorithm to identify earthquakes, finding 48 earthquakes around the globe with only 2 false triggers in 5 months of data. The detections are faster than seismographic detections, with 75 % occurring within 2 min. These results demonstrate the efficiency of using Twitter as a detection tool, potentially achieving better and more accurate results when combined with existing systems.

The rich semantics of tweets and Twitter's broadcasting nature also hint at the potential of using Twitter for rapid emergency response tools to assist in intervention and crisis management. Caragea et al. developed a reusable information technology infrastructure, called Enhanced Messaging for the Emergency Response Sector (EMERSE) Caragea et al. (2011). The system is aimed at classifying tweets and text messages automatically, together with the ability to deliver relevant information to relief workers. EMERSE has four components, including an iPhone application, a Twitter crawler, machine translation, and automatic message classification. The system analyzed the information about the Haiti earthquake relief and provided their output to NGOs, relief workers, and victims and their friends and relatives in Haiti. To use Twitter as an emergency response tool, it is important to assess the information quality of tweets during an emergency situation. Li and Rao (2010) studied Twitter usage following the Sichuan earthquake in China in 2008. They focused on five information quality dimensions: timeliness, accessibility, accuracy, completeness, and collective intelligence, arguing that Twitter is an effective tool for information dissemination in critical moments following earthquake and its broadcasting nature plays an important role in emergency response. Mendoza et al. (2010) studied the dissemination of false rumors and confirmed news following 2010 Chile earthquake, finding that false rumors tend to be questioned much more than confirmed news. Their study indicates the possibility of using Twitter to detect rumors after an earthquake to make the rescue efforts more efficient.

Emergencies and Mobile Phones

In addition to social media websites, the pervasive adoption of mobile phones provides another potentially even more detailed avenue to monitor large populations. Mobile phone records usually include fine-grained longitudinal mobility traces and communication logs. The data allows greater opportunity to study personal social networks through their relationship with physical space, compared to the online social networks (e.g., "friends" and "followers" on Twitter). Mobile phones are well established in many areas, even in third world countries such as Rwanda (Kapoor et al. 2010). Leveraging their presence to assist in emergency response has great potential to save lives. Here we review two recent papers focused on mobile phones and emergencies. The first studied an earthquake that occurred in central Africa (Kapoor et al. 2010). The second analyzed a corpus of events including non-emergency controls such as music festivals occurring in Western Europe (Bagrow et al. 2011).

An Earthquake in Central Africa

To understand how effective mobile phones are at understanding emergency situations, a number of studies have been conducted. Kapoor et al. (2010) studied a 5.9 magnitude earthquake that occurred February 3, 2008, in Lake Kivu region of the Democratic Republic of Congo Kapoor et al. (2010). The dataset is the cellular activity patterns of mobile phone users in Rwanda. They used daily call volume on a per tower basis, and they also had the geographic coordinates of the towers. Their goal was to determine the location of the epicenter algorithmically using only the cellular data and to assess or predict what areas of the country are most in need of aid due to the earthquake.

To study these problems, they assumed that (i) cell tower traffic deviates in a statistically significant manner from normal activity levels

and trends when an event occurs, (ii) areas that are more disturbed by the event will display traffic deviations for longer periods of time, and (iii) disruptions are inversely proportional to the distance from the catastrophe.

To detect an event they assumed the typical daily traffic on a tower obeys a gaussian distribution and they used a negative log-likelihood score to compare the current traffic with this distribution. The higher this score, the more likely there was an anomalous event on that day. They demonstrated that this score spikes on the day of the event, although they did not discuss a specific algorithm to automatically flag scores (e.g., introducing a threshold score such that an event is anomalous when its score exceeds that threshold).

To estimate the location of the event, they assumed the activity levels at a tower during the event follow a normal or gaussian distribution but that the mean of this tower's distribution is now a function of the distance from the epicenter. Specifically they used for tower i a distance-dependent mean $m_i + \alpha D_i(e_x, e_y)^{-1}$, where m_i is the normal mean traffic for i, α is some configurable scaling parameter, and $D_i(e_x, e_y)$ is the geographic distance of tower i from an epicenter located at coordinates (e_x, e_y). They determined this epicenter (e_x, e_y) (and also α) using well-established maximum likelihood estimates, that is, they found the epicenter and scaling parameters that maximize the sum of the log's of all the tower's probabilities.

The other problem they wish to address is to predict what areas are most in need of emergency aid. To do this, they want to predict whether a particular tower will experience a significant increase in traffic some number of days after the event. They accomplished this by building a classifier which allows them to estimate this persistence probability. Since it is reasonable to assume that areas with higher populations are likely to require more aid, they built an "assistance opportunity score" for a location by taking the product of the persistence probability estimate for that location and the population at that location. Such a score allows emergency responders to potentially prioritize aid efforts.

The authors also pointed out an important issue when using mobile phone data to study these problems: the density of towers, and therefore information, is not uniform. Cities have many more towers than rural regions, and this leads to far greater granularity in areas of high population and greater information uncertainty in areas with fewer towers. They exploited this fact to estimate what areas are most valuable to survey manually for information after an event, by prioritizing surveys towards areas with more uncertainty. They did this by devising a simple mechanism to drive down the entropy in the information that may be gained from the system, and they even incorporated geographic distances since it is more expensive in terms of time and effort to survey more remote regions.

All of their methods were validated by comparison with the February 3 earthquake and were shown to work rather well. For future work they discussed a number of interesting advancements such as incorporating richer models of geographic terrain.

Mobile Phones and Disasters

Bagrow et al. (2011) performed a data-driven analysis of a number of emergencies, including bombings, a plane crash, and another earthquake. This work reported a number of empirical discoveries regarding the response of populations in the wake of emergencies (and non-emergency control events such as festivals), as measured from the country-wide data of a single mobile phone provider in Western Europe. The assumptions made by Kapoor et al. (2010) are further justified by their work.

They found that emergencies trigger a sharp spike in call activity (number of outgoing calls and text messages) in the physical proximity of the event, confirming that mobile phones act as sensitive local "sociometers" to external societal perturbations. In Fig. 1a, we plot the relative call volume $\Delta V / \langle V_{\text{normal}} \rangle$ as a function of time, where $\Delta V = V_{\text{event}} - \langle V_{\text{normal}} \rangle$, V_{event} is the number of calls made from nearby towers during the event, and $\langle V_{\text{normal}} \rangle$ is the average call volume during the same time period of the week (Figure adapted from Bagrow et al. (2011)).

The anomalous traffic starts to decay immediately after the emergency occurs, suggesting that the urge to communicate is strongest right at the onset of the event. There was virtually no delay between the onset of the event and the jump in call volume for events that were directly witnessed by the local population, such as the bombing, the earthquake, and the blackout. Brief delay was observed only for the plane crash, which took place in an unpopulated area and thus lacked eyewitnesses. In contrast, non-emergency events, like the festival and the concert, displayed a gradual increase in call activity.

The temporally localized spikes in call activity (Fig. 1a) raise an important question: is information about an event limited to the immediate vicinity of the emergency or do emergencies, often immediately covered by national media, lead to spatially extended changes in call activity (Petrescu-Prahova and Butts 2008)? To investigate this, Bagrow et al. inspected the change in call activity in the vicinity of each event's epicenter, finding that for the bombing, for example, the change in call volume is strongest near the event and drops rapidly with the distance r from the epicenter. To quantify this effect across all emergencies, they integrated the call volume over time in concentric shells of radius r centered on the epicenter. The observed decay in anomalous traffic was approximately exponential, $\Delta V(r) \sim \exp(-r/r_c)$, allowing one to characterize the spatial extent of the reaction with a decay rate r_c (we present their results for the plane crash in Fig. 1b). The observed decay rates ranged from 2 km (bombing) to 10 km (plane crash), indicating that the anomalous call activity is limited to the event's vicinity. An extended spatial range ($r_c \approx 110$ km) was seen only for the earthquake. Meanwhile, non-emergencies are highly localized: they possess decay rates less than 2 km. This systematic split in r_c between the spatially extended emergencies and well-localized non-emergencies persisted for all explored events.

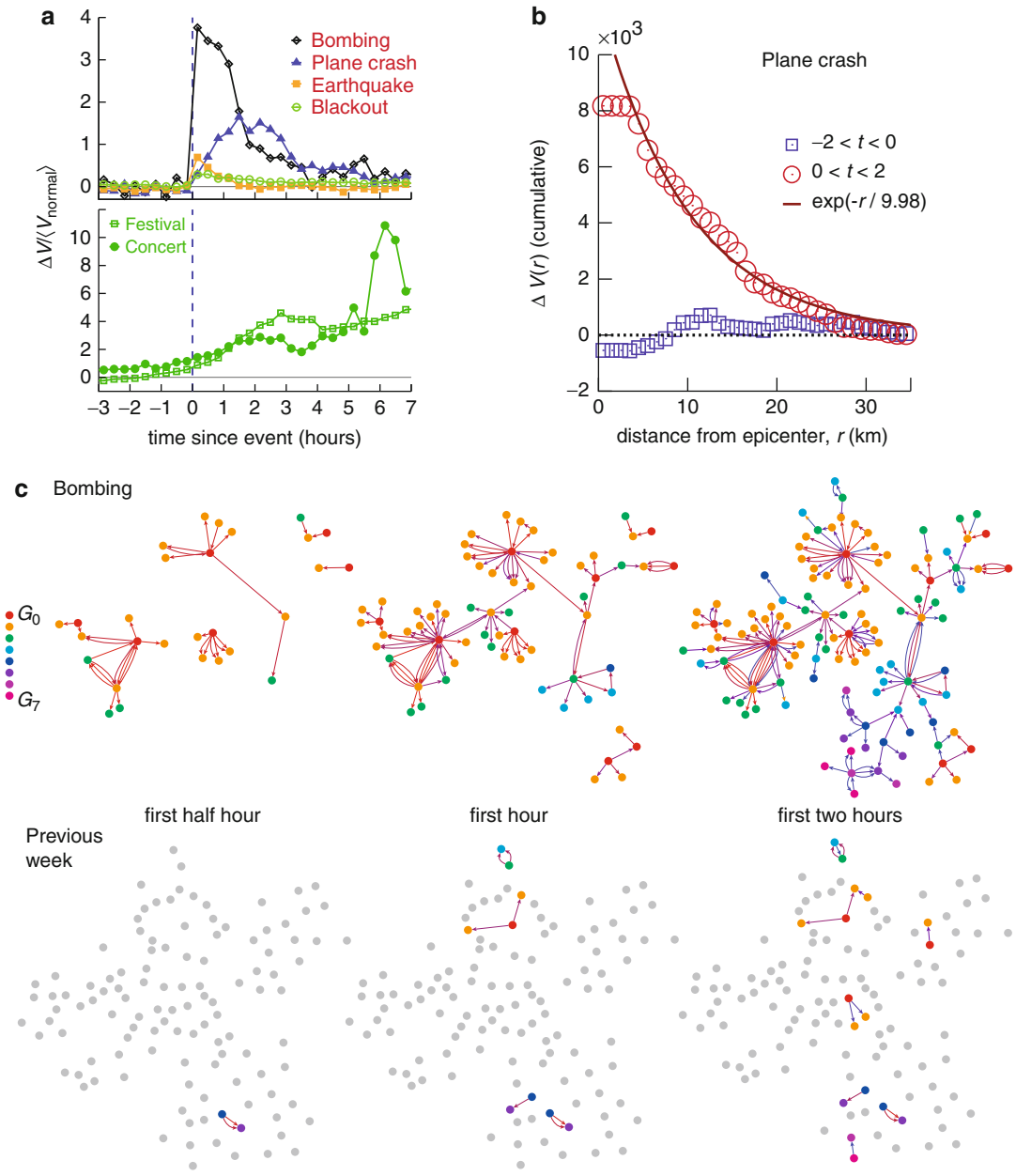
Despite the clear temporal and spatial localization of anomalous call activity during emergencies, one expects some degree of information propagation beyond the eyewitness population. To study how emergency information diffuses through a social network, Bagrow et al. used mobile phone records to identify those individuals located within the event region, forming a population called G_0 as well as a group called G_1 consisting of individuals outside the event region but who receive calls from the G_0 group during the event, a G_2 group that receive calls from G_1, and so on. They reveal that the G_0 individuals typically engage their social network within minutes and that the G_1, G_2, and occasionally even the G_3 group show an anomalous call pattern immediately after an emergency. We present their illustration of a segment of this contact network for the bombing in Fig. 1c. The authors proceeded to further quantify and control for this social propagation and showed that the bombing and plane crash have significant propagation up to the third and second neighbors of G_0, respectively. They found that other emergencies, the earthquake and blackout, displayed relatively little propagation. This seems reasonable given the less severe nature of those events (the earthquake was relatively minor).

Finally, we also presented a breakdown of a number of measurable features for each emergency and non-emergency and showed that these features may be used to distinguish anomalous call activity due to benign events such as music festivals from spikes in call volume that indicate a dangerous event has occurred. Using such factors may allow first responders to more accurately understand rapidly unfolding events and may even allow them to actively solicit information from mobile phone users likely to be near the event.

Key Techniques

We summarize the key techniques that have been used in the above-mentioned studies.

Event Detection. The first challenge in large-scale emergency studies is to determine and collect a subset of data relevant to emergencies under consideration. With Twitter or other social media data where the communication content is



Social Networks in Emergency Response, Fig. 1 Temporal, spatial, and social response during emergencies. **a** The time dependence of call volume $V(t)$ after four emergencies and two non-emergencies. We plot the relative change in call volume $\Delta V / \langle V_{\text{normal}} \rangle$, where $\Delta V = V_{\text{event}} - \langle V_{\text{normal}} \rangle$, V_{event} is the call volume on the day of the event, and $\langle V_{\text{normal}} \rangle$ is the average call volume during the same period of the week. **b** The total change in call volume during 2-h periods before and after the plane crash, as a function of distance r from the epicenter of the crash. Following the event, we see an approximately exponential decay $\Delta V \sim \exp r / r_c$ characterized by decay rate r_c. **c** Part of the contact network formed between mobile phone users in the wake of the bombing. Nodes are colored by group, with G_0 representing phone users calling from the event region, G_1 the recipients of those calls, etc. As time goes by more users are contacted as information propagates. Those same users make little contact during a corresponding time period the week before (Figure adapted from Bagrow et al. (2011))

available in text format, most studies begin with a simple keyword matching, that is, collecting data that contained instances of the relevant keywords such as "flu," "H1N1," and "earthquake." The initial collections could be refined by manual and automated classification process. Classification techniques such as Support Vector Machines (SVMs) have been employed (Aramaki et al. 2011; Sakaki et al. 2010), and topic clustering methods such as Latent Dirichlet Allocation (LDA) can be used to improve the classification (Paul and Dredze 2011; Prier et al. 2011). Validation of this body of work is often conducted based on authority reports such as Centers for Disease Control and Prevention (CDC) statistics (Signorini et al. 2011) (for disease outbreaks) or US Geological Survey (USGS) reports (Guy et al. 2010; Earle et al. 2012) (for earthquakes). While the messages disseminated in social media might be inaccurate, there has been work on determining the quality of information sources (Li and Rao 2010; Mendoza et al. 2010). Further, by applying time-series analysis and spatiotemporal pattern analysis (e.g., Kalman filtering and particle filtering in Sakaki et al. (2010)), researchers have developed powerful earthquake detectors with performance comparative to existing earthquake detection systems.

Event Prediction and Forecasting. The development of event prediction and forecasting is still in its early stage. Gomide et al. (2011) used a linear regression model to predict the number of dengue cases. The earthquake detectors (Sakaki et al. 2010; Guy et al. 2010; Earle et al. 2012) that reported earthquakes faster than the seismographic detection can be used as early warning system. There has been work on developing information infrastructure which has the ability to deliver relevant information to users once events are detected (Caragea et al. 2011).

Spatiotemporal Pattern Recognition of Events. Unlike social media data, the content of communication is often unavailable in mobile phone data, and hence the identification of emergency events in mobile phone data relies on analyses of spatial and temporal anomalies of call logs. The main challenge of this research is to construct reasonable null model in order to recognize anomaly events. Bagrow et al. (2011) proposed using pre-emergency normal activities as well as the activities during non-emergency events to contrast the activities of emergency events. Based on this approach the epicenter of an emergency event can be identified. Kapoor et al. (2010) used a similar methodology to identify event epicenters as well as to predict the locations in need of emergency aid.

Future Trends

Foundational work understanding the sociology of disaster was limited in scale by available data but surveys and interviews can ask a number of in-depth follow-up questions. To understanding population response from, for example, mobile phone call volume alone is potentially more challenging as such data, while perhaps being more objective, is also far shallower. This begs the question: can more depth be found in communications data? The wealth of textual information available within social media such as Twitter can be leveraged to learn more context about how populations respond to emergencies, and advances in data mining and natural language processing techniques offer the promise of even greater information. This may allow researchers to separate relevant information from spurious activity, improving the accuracy and precision of information available to rescuers.

One can reasonably expect a degree of noise from any communication system, as users will be focused on diverse topics. Yet when something of overwhelming importance occurs, such as an emergency, it seems reasonable to expect that event to capture the majority of user attention. This may lead to a communication system that is less noisy and more focused as the severity of the event increases, in the sense that an increasing fraction of the system's communication will be about that event. Given this, it may be worth trying to develop (rigorous) bounds on how much useful information can be successfully extracted from such a system during and immediately following an event. This could allow quantitative

benchmarking of algorithms designed to assist rescuers by comparing, for example, how much emergency information was extracted by an algorithm with the maximum amount possible.

Meanwhile, it will be crucial going forward to develop algorithms that combine and help understand multiple data sources—such as cell phone call volume, twitter messages, and perhaps even security cameras, all from a given geographic locale. This trend towards greater data availability and unification will only continue as more advanced and entirely new forms of telecommunication come into widespread use. Without methods to handle the increased diversity and volume of communication, rescuers may be unable to capitalize on the extra information provided by future telecommunications.

Conclusion

We have reviewed a number of works focused on the use of communications data, from social media to mobile phones, to understand how people react to emergencies and disasters. This problem is of critical importance: in many areas of the world, more people than ever are at risk, as both human populations and threats due to climate change continue to grow. Hopefully tools derived from social media and other communication datasets will help rescuers improve their emergency and disaster response by providing accurate, useful, and timely information in the wake of such events.

Cross-References

▸ Actionable Information in Social Networks, Diffusion of
▸ Counter-Terrorism, Social Network Analysis in
▸ Disaster Response and Relief, VGI Volunteer Motivation in
▸ Extracting Individual and Group Behavior from Mobility Data
▸ Modeling and Analysis of Spatiotemporal Social Networks
▸ Social Network Datasets
▸ Social Networking in the Telecom Industry
▸ Spatiotemporal Proximity and Social Distance
▸ Temporal Networks

References

Aramaki E, Maskawa S, Morita M (2011) Twitter catches the flu: detecting influenza epidemics using twitter. In: Proceedings of the conference on empirical methods in natural language processing, Edinburgh. Association for Computational Linguistics, pp 1568–1576

Bagrow JP, Wang D, Barabśi A-L (2011) Collective response of human populations to large-scale emergencies. PLoS ONE 6(3):e17680

Caragea C, McNeese N, Jaiswal A, Traylor G, Kim H, Mitra P, Wu D, Tapia A, Giles L, Jansen B, et al (2011) Classifying text messages for the Haiti earthquake. In: Proceedings of the 8th international ISCRAM conference, ISCRAM, Harbin, vol 11

Chew C, Eysenbach G (2010) Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. PLoS One 5(11):e14118

Culotta A (2010) Towards detecting influenza epidemics by analyzing twitter messages. In: Proceedings of the first workshop on social media analytics, Washington, DC. ACM, pp 115–122

Earle P, Bowden D, Guy M (2012) Twitter earthquake detection: earthquake monitoring in a social world. Ann Geophys 54(6):708–715

Ginsberg J, Mohebbi M, Patel R, Brammer L, Smolinski M, Brilliant L (2008) Detecting influenza epidemics using search engine query data. Nature 457(7232):1012–1014

Gomide J, Veloso A, Meira W, Almeida V, Benevenuto F, Ferraz F, Teixeira M (2011) Dengue surveillance based on a computational model of spatio-temporal locality of twitter

Guy M, Earle P, Ostrum C, Gruchalla K, Horvath S (2010) Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies. In: Proceedings of the 9th international conference on advances in intelligent data analysis, Tucson, pp 42–53

Hughes A, Palen L (2009) Twitter adoption and use in mass convergence and emergency events. Int J Emerg Manag 6(3):248–260

Kapoor A, Eagle N, Horvitz E (2010) People, quakes, and communications: inferences from call dynamics about a seismic event and its influences on a population. In: Proceedings of AAAI artificial intelligence for development (AI-D'10), Stanford

Li J, Rao H (2010) Twitter as a rapid response news service: an exploration in the context of the 2008 china earthquake. Electron J Inf Syst Dev Ctries 42(0)

Mendoza M, Poblete B, Castillo C (2010) Twitter under crisis: can we trust what we rt? In: Proceedings of the

first workshop on social media analytics, Washington, DC. ACM, pp 71–79
Paul M, Dredze M (2011) You are what you tweet: analyzing twitter for public health. In: Proceedings of the 5th international AAAI conference on weblogs and social media (ICWSM), Barcelona
Petrescu-Prahova M, Butts CT (2008) Emergent coordinators in the World Trade Center disaster. Int J Mass Emerg Disasters 28(3):133–168
Prier K, Smith M, Giraud-Carrier C, Hanson C (2011) Identifying health-related topics on twitter. In: Proceedings of the 4th international conference social computing, behavioral-cultural modeling and prediction, College Park, pp 18–25
Quincey E, Kostkova P (2010) Early warning and outbreak detection using social networking websites: The potential of twitter. Electronic Healthcare. Springer Berlin Heidelberg, 21–24
Rodríguez H, Quarantelli E, Dynes R (2006) Handbook of disaster research. Springer, New York
Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on world wide web, Raleigh. ACM, pp 851–860
Sheetz S, Kavanaugh AL, Quek F, Kim BJ, Lu S-C (2010) The expectation of connectedness and cell phone use in crises. J Emergen Manag 7(2):124–136
Signorini A, Segre A, Polgreen P (2011) The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. PLoS One 6(5):e19467
Zamite J, Silva F, Couto F, Silva M (2011) Medcollector: multisource epidemic data collector. In: Transactions on large-scale data-and knowledge-centered systems IV. Springer, Berlin/Heidelberg, 7(2):124–136

Social Networks in Healthcare, Case Study

Fei Wang
Healthcare Analytics Research Group, IBM T. J. Watson Research Center, Yorktown Heights, NY, USA

Synonyms

Electronic health record; Patient similarity

Glossary

Patient similarity The clinical similarity score between pairwise patients derived from their records

Patient network A network with nodes representing patient entities, edges representing pairwise patient similarities

Definition

Constructing an undirected patient network with patients as nodes and pairwise clinical similarities as edge weights can enable many applications in modern medical informatics such as physician decision support, risk stratification, and comparative effectiveness research, because similar patients have similar clinical characteristics and thus the treatment on one patient might be helpful to his/her similar patients. Therefore constructing such a patient network is very important to data-driven analytics for healthcare, and effective patient similarity evaluation is the key to construct the patient network.

Introduction

Healthcare has undergone a tremendous growth in the use of electronic health records (EHR) systems to capture patient disease and treatment histories. However, these systems store the data in a manner that makes it difficult for clinicians to extract what is necessary to make clinical decisions at the point-of-care. Most of EHR systems are primarily used to record clinical events for bookkeeping and claim purposes as opposed to be used as a decision support tool for better diagnosis and treatment. Constructing a patient network with nodes representing patients and edges connecting clinically similar patients might be very helpful to such a clinical decision support system, as the physician can look at the treatments and disease condition evolutions of the similar patients to come up with a better care plan for the current patient.

Actually besides decision support systems, there are also other areas in medical informatics where such patient network could be very helpful, for example, *comparative effectiveness research* (CER), which is *the direct comparison of existing healthcare interventions to determine which work best for which patients and*

which pose the greatest benefits and harms (http://en.wikipedia.org/wiki/Comparative_effectiveness_research 2013). In such a case, if we can first stratify the patients into different cohorts according to their clinical similarity, then CER can be performed on the patients within the same cohorts. Under a similar setting, patient *risk stratification* aims to stratify the patients according to their disease condition risks. This is a crucial step for effective management of patients, because for patients with different risks, we may have different treatment plans. One step forward, if we can construct an undirected patient network using such patient similarity, we can expect to discover some disease and their evolution patterns, as well as the care/treatment patterns, which would be clinically very useful.

Current Technologies

There have already been quite a few patient similarity evaluation techniques. Before giving an overview of them, we first need to introduce the vector space representation of the patient clinical characteristics, which is an enabling technique to invoke the similarity learning and computations.

Patient Profiling

Patient EHRs contain lots of heterogeneous information, such as demographic information, diagnosis, medication, and lab tests. We call these different information source *features*. To facilitate the process of similarity learning, some researchers proposed to construct a profile for each patient, which is a feature vector with the dimensionality equal to the number of different features. Before constructing such a vector, we first define a time period of interest, within which we will aggregate the features to get the entries in the patient profile (e.g., the average value of a specific lab test or the count of a specific diagnosis code). In this way, after profiling, each patient is represented as a feature vector (Wang et al. 2011a, b, 2012).

Locally Supervised Metric Learning

Locally Supervised Metric Learning (LSML) is a supervised metric learning approach that has been proved to be useful in patient similarity evaluation (Sun et al. 2010b, a; Ebadollahi et al. 2010). This algorithm was initially proposed in Wang et al. (2009) for measuring text similarity. In the following, we use $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ to represent a data matrix from a single specific party, and $\mathbf{y} = [y_1, \cdots, y_n]^T \in \mathbb{R}^n$ is the corresponding label vector with $y_i \in \{1, 2, \cdots, C\}$ denoting the label of $\mathbf{x}_i$, and C is the number of classes. Some examples of the labels here can be diagnosis, for example, the patient has diagnosis or not, or hospitalization, meaning the patient is hospitalized or not, etc.

Our goal is to learn a *Mahalanobis distance* as follows:

$$d_{\Sigma}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^{\top} \boldsymbol{\Sigma} (\mathbf{x}_i - \mathbf{x}_j)} \quad (1)$$

where $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ is a *symmetric positive semi-definite* (*SPSD*) matrix. Following Wang et al. (2009), we define the *homogeneous neighborhood* and *heterogeneous neighborhood* around each data point as

(**Homogeneous neighborhood**). The homogeneous neighborhood of $\mathbf{x}_i$, denoted as $\mathcal{N}_i^o$, is the $|\mathcal{N}_i^o|$-nearest data points of $\mathbf{x}_i$ with the same label.

(**Heterogeneous neighborhood**). The heterogeneous neighborhood of $\mathbf{x}_i$, denoted as $\mathcal{N}_i^e$, is the $|\mathcal{N}_i^e|$-nearest data points of $\mathbf{x}_i$ with different labels.

In the above two definitions, we use $|\cdot|$ to denote set cardinality. In order to define the individual distance metric on this party, we need to first construct the neighborhood $\mathcal{N}_i^o$ and $\mathcal{N}_i^e$. Then we can define the local compactness and scatterness around point $\mathbf{x}_i$ as

$$\mathcal{C}_i = \sum\nolimits_{j:\mathbf{x}_j \in \mathcal{N}_i^o} d_{\Sigma}^2(\mathbf{x}_i, \mathbf{x}_j) \quad (2)$$

$$\mathcal{S}_i = \sum\nolimits_{k:\mathbf{x}_k \in \mathcal{N}_i^e} d_{\Sigma}^2(\mathbf{x}_i, \mathbf{x}_k) \quad (3)$$

Then we can learn an optimal distance metric by minimizing the following *discrimination* criterion

$$\mathcal{J} = \sum\nolimits_{i=1}^{n} (\mathcal{C}_i - \mathcal{S}_i) \quad (4)$$

which makes the data in the same class compact while data in different class diverse. As $\mathbf{\Sigma}$ is SPSD, we can factorize it using incomplete Cholesky decomposition as

$$\mathbf{\Sigma} = \mathbf{W}\mathbf{W}^\top \tag{5}$$

Then $\mathcal{J}$ can be expanded as

$$\mathcal{J} = tr\left(\mathbf{W}^\top \left(\mathbf{\Sigma}_\mathcal{C} - \mathbf{\Sigma}_\mathcal{S}\right)\mathbf{W}\right) \tag{6}$$

where $tr(\cdot)$ is the matrix trace, and

$$\mathbf{\Sigma}_\mathcal{C} = \sum\nolimits_i \sum\nolimits_{j:\mathbf{x}_j \in \mathcal{N}_i^o} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \tag{7}$$

$$\mathbf{\Sigma}_\mathcal{S} = \sum\nolimits_i \sum\nolimits_{k:\mathbf{x}_k \in \mathcal{N}_i^e} (\mathbf{x}_i - \mathbf{x}_k)(\mathbf{x}_i - \mathbf{x}_k)^\top \tag{8}$$

are the local *compactness* and *scatterness* matrices. Hence the distance metric learning problem can be formulated as

$$\min_{\mathbf{W}:\mathbf{W}^\top\mathbf{W}=\mathbf{I}} tr\left(\mathbf{W}^\top \left(\mathbf{\Sigma}_\mathcal{C} - \mathbf{\Sigma}_\mathcal{S}\right)\mathbf{W}\right) \tag{9}$$

Note that the orthogonality constraint $\mathbf{W}^\top\mathbf{W} = \mathbf{I}$ is imposed to reduce the information redundancy among different dimensions of $\mathbf{W}$, as well as control the scale of $\mathbf{W}$ to avoid some arbitrary scaling. The optimal solution of $\mathbf{W}$ can be obtained by doing eigenvalue decomposition to $\mathbf{\Sigma}_\mathcal{C} - \mathbf{\Sigma}_\mathcal{S}$ with the largest eigenvectors.

In summary, the individual distance metric, which is parameterized by a projection matrix $\mathbf{W}$, can finally be learned from local neighborhood information. Next, we will show how to combine these neighborhoods from different base metrics into a single optimal distance metric.

Efficient Metric Updating: Interactive Metric Learning

One issue for applying the above LSML technique in patient similarity evaluation for physician decision support is that the physician may give some feedback after he/she sees the results. Therefore it is important for LSML to be capable of efficiently incorporating those feedbacks. The feedbacks in general can be regarded as in the form of label changes of $\mathbf{y}$, which consequently leads to changes to $\mathbf{\Sigma}_\mathcal{C}$ and $\mathbf{\Sigma}_\mathcal{S}$, and the key is to efficiently updating the eigenvalue and eigenvectors of $\mathbf{\Sigma}_\mathcal{C} - \mathbf{\Sigma}_\mathcal{S}$. In the following we will briefly describe how the authors in Wang et al. (2011b) solve this problem.

Definition and Setup To facilitate the discussion, we define the following matrix:

$$\mathbf{\Sigma} = \mathbf{\Sigma}_\mathcal{C} - \mathbf{\Sigma}_\mathcal{S} \tag{10}$$

Next we introduce an efficient technique based on matrix perturbation (Stewart and Sun 1990) to adjust the learned distance metric according to changes of $\mathbf{\Sigma}$. Suppose that after adjustment, $\mathbf{L}$ becomes

$$\widetilde{\mathbf{\Sigma}} = \mathbf{\Sigma} + \Delta\mathbf{\Sigma} \tag{11}$$

We define $(\lambda_i, \mathbf{w}_i)$ as one eigenvalue-eigenvector pair of matrix $\mathbf{\Sigma}$. Similarly, we define $(\widetilde{\lambda}_i, \widetilde{\mathbf{w}}_i)$ as one eigenvalue-eigenvector pair of $\widetilde{\mathbf{\Sigma}}$.

Then we can rewrite $(\widetilde{\lambda}_i, \widetilde{\mathbf{w}}_i)$ as

$$\widetilde{\lambda}_i = \lambda_i + \Delta\lambda_i \tag{12}$$

$$\widetilde{\mathbf{w}}_i = \mathbf{w}_i + \Delta\mathbf{w}_i \tag{13}$$

Next we can obtain

$$(\mathbf{\Sigma} + \Delta\mathbf{\Sigma})(\mathbf{w}_i + \Delta\mathbf{w}_i) = (\lambda_i + \Delta\lambda_i)(\mathbf{w}_i + \Delta\mathbf{w}_i) \tag{14}$$

Now the key questions are how to compute changes to the eigenvalue $\Delta\lambda_i$ and eigenvector $\Delta\mathbf{w}_i$, respectively.

Eigenvalue Update Expanding Eq. (14) and using the fact that $\mathbf{\Sigma}\mathbf{w}_i = \lambda_i\mathbf{w}_i$, we can obtain the following equation:

$$(\mathbf{\Sigma} + \Delta\mathbf{\Sigma})\mathbf{w}_i = \lambda_i\Delta\mathbf{w}_i + \Delta\lambda_i\mathbf{w}_i \tag{15}$$

Now multiplying both sides of Eq. (15) with $\mathbf{w}_i^\top$ and because of the symmetry of $\mathbf{\Sigma}$, we get

$$\Delta\lambda_i = \mathbf{w}_i^\top\Delta\mathbf{\Sigma}\mathbf{w}_i \tag{16}$$

Eigenvector Update Since the eigenvectors are orthogonal to each other, we assume that the change of the eigenvector $\Delta\mathbf{w}_i$ is in the subspace spanned by those original eigenvectors, i.e.,

$$\Delta \mathbf{w}_i \approx \sum_{j=1}^{d} \alpha_{ij} \mathbf{w}_j \tag{17}$$

where $\{\alpha_{ij}\}$ are small constants to be determined. Bringing Eq. (17) into Eq. (15), we obtain

$$\boldsymbol{\Sigma} \sum_{j=1}^{d} \alpha_{ij} \mathbf{w}_j + \Delta \boldsymbol{\Sigma} \mathbf{w}_i = \lambda_i \sum_{j=1}^{d} \alpha_{ij} \mathbf{w}_j + \Delta \lambda_i \mathbf{w}_i$$

Multiplying $\mathbf{w}_k^\top$ $(k \neq i)$ on both side of the above equation and discarding the high-order term and bringing in Eq. (17), we get

$$\Delta \mathbf{w}_i = -\sum_{j \neq i} \frac{\mathbf{w}_j^\top \Delta \boldsymbol{\Sigma} \mathbf{w}_i}{\lambda_i - \lambda_j} \mathbf{w}_j \tag{18}$$

Collective Intelligence: Composite Distance Integration

Another challenge in patient similarity is that different physicians have different opinions, then how to integrate all of them to come up with an objective patient similarity? The authors in Wang et al. (2011a, 2012) presented an approach on integrating neighborhood information from multiple parties (physicians) when performing LSML. Next we will briefly review this technique.

Objective Function

The goal here is still learning a Mahalanobis distance as in Eq. (1) but integrating the neighborhood information from all parties. Here the q-th party constructs homogeneous neighborhood $\mathcal{N}_i^o(q)$ and heterogeneous neighborhood $\mathcal{N}_i^e(q)$ for the i-th data point in it. Correspondingly, the compactness matrix $\boldsymbol{\Sigma}_{\mathcal{C}}^q$ and the scatterness matrix $\boldsymbol{\Sigma}_{\mathcal{S}}^q$ are computed and shared by the q-th party:

$$\boldsymbol{\Sigma}_{\mathcal{C}}^q = \sum_{i \in \mathcal{X}_q} \sum_{j: \mathbf{x}_j \in \mathcal{N}_i^o(q)} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top$$

$$\boldsymbol{\Sigma}_{\mathcal{S}}^q = \sum_{i \in \mathcal{X}_q} \sum_{k: \mathbf{x}_k \in \mathcal{N}_i^e(q)} (\mathbf{x}_i - \mathbf{x}_k)(\mathbf{x}_i - \mathbf{x}_k)^\top$$

Similar to one party case presented in Eq. (6), we generalize the optimization objective as

$$\begin{aligned} \mathcal{J} &= \sum_{q=1}^{m} \alpha_q \mathcal{J}^q \\ &= \sum_{q=1}^{m} \alpha_q tr\left(\mathbf{W}^\top \left(\boldsymbol{\Sigma}_{\mathcal{C}}^q - \boldsymbol{\Sigma}_{\mathcal{S}}^q\right) \mathbf{W}\right) \end{aligned} \tag{19}$$

where α_q is the importance for the q-th party and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_m)^\top$ is constrained to be in a simplex as $\alpha_q \geq 0$, $\sum_q \alpha_q = 1$, and m is the number of parties. Note that by minimizing Eq. (19), the proposed approach actually leverages the local neighborhoods of all parties to get a more powerful discriminative distance metric. Thus it aims at solving the following optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\alpha}, \mathbf{W}} \; & \sum_{q=1}^{m} \alpha_q tr\left(\mathbf{W}^\top \left(\boldsymbol{\Sigma}_{\mathcal{C}}^q - \boldsymbol{\Sigma}_{\mathcal{S}}^q\right) \mathbf{W}\right) \\ & + \lambda \Omega(\boldsymbol{\alpha}) \\ s.t. \quad & \boldsymbol{\alpha} \geq 0, \; \boldsymbol{\alpha}^\top \mathbf{e} = 1 \\ & \mathbf{W}^\top \mathbf{W} = \mathbf{I} \end{aligned} \tag{20}$$

Here $\Omega(\boldsymbol{\alpha})$ is some regularization term used to avoid trivial solutions, and $\lambda \geq 0$ is the trade-off parameter. In particular, when $\lambda = 0$, i.e., without any regularization, only $\alpha_q = 1$ for the best party, while all the others have zero weight. The best λ can be selected through cross-validation.

Problem (Eq. 20) can be solved by alternating optimization and the procedure is guaranteed to converge to a local optimum.

Future Trends

Although LSML is a powerful methodology and it has been proved to be useful on some real-world clinical data (Ebadollahi et al. 2010; Sun et al. 2010b; Wang et al. 2012), there are still some limitations which include: (1) It is a supervised approach, meaning, for all the training data, we need to have their supervision information (either in terms of labels or pairwise constraints) – this is difficult in medical scenario as the supervision information is expensive and time-consuming to obtain – and (2) it needs to construct different types of neighborhoods; this could be time-consuming when the data set scale

is large. Therefore the future research towards effective patient similarity evaluation should be the following: (1) Use less supervisions and more unsupervised data. Semi-supervised learning techniques (Zhu and Goldberg 2009) could be helpful in this scenario. (2) Improve the scalability of the algorithm and make it fit in the scenario when we have millions of patients.

Conclusion

This chapter reviews the state-of-the-art technology for patient similarity evaluation, which can be used for constructing a patient network. Specifically, we introduced the Locally Supervised Metric Learning (LSML) algorithm as well as its two variants on how to make real-time updates and integrate multiple experts' opinions. We finally point out that the future research directions of this research topic.

Cross-References

- ▸ Data Mining
- ▸ Disease Surveillance, Case Study
- ▸ Distance and Similarity Measures
- ▸ Online Healthcare Management

References

(2013) http://en.wikipedia.org/wiki/Comparative_effectiveness_research

Ebadollahi S, Sun J, Gotz D, Hu J, Sow D, Neti C (2010) Predicting patient's trajectory of physiological data using temporal trends in similar patients: a system for near-term prognostics. In: AMIA annual symposium proceedings, pp 192–196

Stewart G, Sun JG (1990) Matrix perturbation theory. Academic, Boston

Sun J, Sow DM, Hu J, Ebadollahi S (2010a) Localized supervised metric learning on temporal physiological data. In: International conference on pattern recognition, pp 4149–4152

Sun J, Sow DM, Hu J, Ebadollahi S (2010b) A system for mining temporal physiological data streams for advanced prognostic decision support. In: IEEE international conference on data mining, pp 1061–1066

Wang F, Sun J, Li T, Anerousis N (2009) Two heads better than one: Metric+active learning and its applications for it service classification. In: IEEE international conference on data mining, pp 1022–1027

Wang F, Sun J, Ebadollahi S (2011a) Integrating distance metrics learned from multiple experts and its application in inter-patient similarity assessment. In: SIAM data mining conference, pp 59–70

Wang F, Sun J, Hu J, Ebadollahi S (2011b) imet: interactive metric learning in healthcare applications. In: SIAM data mining conference, pp 944–955

Wang F, Sun J, Ebadollahi S (2012) Composite distance metric integration by leveraging multiple experts' inputs and its application in patient similarity assessment. Stat Anal Data Min 5(1):54–69

Zhu X, Goldberg AB (2009) Introduction to semi-supervised learning. Synthesis lectures on artificial intelligence and machine learning. Morgan & Claypool, San Rafael

Social Networks Members

▸ Social Influence Analysis

Social Networks Users

▸ Social Influence Analysis

Social Order in Online Social Networks

Tina Eliassi-Rad
Department of Computer Science, Rutgers University, Piscataway, NJ, USA

Glossary

Tie A relationship between two individuals

Social Network A set of individuals connected by a set of dyadic ties

Online Social Network A social network on the World Wide Web

Definition

Social order, a technical term from social sciences (Frank 1944), is the study of how social creatures (such as human beings) are both

individual and social (Hechter and Horne 2003). As Hechter and Horne (2003) point out, social order occurs when individuals coordinate and cooperate with each other.

Social order in online social networks and the coordination and cooperation that give rise to them appear in many different structural forms. Examples include *homophily*, *communities* (a.k.a. groups), *weak ties*, *structural holes*, and *social capital*.

Homophily The notion of homophily (i.e., "of like attracting like") has been around since the ancient Greeks. It is often quoted that Plato said, "Similarity begets friendship." Previous research (McPherson et al. 2001) has shown that homophily is a major criterion governing the formation of ties in social networks. Many social networks have high levels of homophily (Easley and Kleinberg 2010, pp. 79–81). Coordination and cooperation is often more successful between people who are similar to each other – either in terms of status or value (McPherson et al. 2001).

Communities Generally speaking, communities are defined as groups of individuals that are well connected to each other. The existing literature contains many objective functions and algorithms that formalize the aforementioned definition and produce communities (Leskovec et al. 2010). The one pertinent to social order is where a community has low conductance, i.e., where the ratio of ties crossing the community boundary to ties within the community is low (Leskovec et al. 2010). Members of such communities are often tightly connected. These highly connected structures, in turn, promote trust among their members – an important property for social order.

Leskovec et al. (2008) found that the sizes of communities in large online networks roughly follow the Dunbar number (∼150) (Dunbar 1998) and that large well-defined communities are absent in online networks. These findings make intuitive sense since maintaining relationships besides the trivial ones requires substantial investment in terms of our neocortex processing capabilities (Dunbar 1998).

Moreover, Leskovec et al. (2008) describe large social networks as having a nested core-periphery structure, where the network is composed of layers of large cores and a small number of dense communities loosely connected to the core. This result indicates the presence of a hierarchy or nested social order in online social networks. In other words, the levels of coordination and cooperation vary depending on where in the nested core-periphery structure a person resides.

Weak Ties Granovetter (2003) was the first to distinguish between weak and strong ties in social networks. He informally defined *tie strength* as the "amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie" (Granovetter 2003, p. 1361). Weak ties correspond to "local bridges" (Easley and Kleinberg 2010), where two people have zero common friends. The lack of common friends can make coordination and cooperation difficult and reduce social order.

Structural Holes Burt (2004) defined structural holes as the empty spaces (i.e., no connections) between groups in the social network. People who fill these structural holes bring social order to the network because they control the information flow and are rewarded with power and wealth.

Social Capital Being members of a community has many advantages (Portes 1998). For example, belonging to a community with high *triadic closure* (where friend of a friend is a friend) and *embeddedness* (where two people share many of the same friends) enforces norms and maintains reputational effects. In other words, this "closure" of friends promotes trust. The counterbalance to closure is *brokerage*. People who are "brokers" interact at the boundary of various communities – i.e., they fill the structural holes. As mentioned above, such people have more social capital compared to others in the community.

Social order, in terms of closures and brokerages, is essential in the preservation of social networks. Closures give rise to communities, while brokerages give rise to connections across various communities.

Cross-References

- ▸ Centrality Measures
- ▸ Community Detection, Current and Future Research Trends
- ▸ Human Behavior and Social Networks
- ▸ Role Discovery
- ▸ Role Identification of Social Networkers
- ▸ Social Capital
- ▸ Structural Holes
- ▸ Trust in Social Networks

References

Burt RS (2004) Structural holds and good ideas. Am J Sociol 110(2):349–399

Dunbar R (1998) Grooming, Gossip, and the evolution of language. Harvard University Press, Cambridge, MA

Easley D, Kleinberg J (2010) Networks, crowds, and markets. Cambridge University Press, New York, NY

Frank LK (1944) What is social order? Am J Sociol 49(5):470–477

Granovetter M (2003) The strength of ties. In: Hechter M, Horne C (eds) Theories of social order: A reader. Stanford University Press, Stanford, CA, pp 323–332

Hechter M, Horne C (2003) Theories of social order: A reader. Stanford University Press, Stanford, CA

Leskovec J, Lang K, Dasgupta A, Mahoney M (2008) Statistical properties of community structure in large social and information networks. In: The 17th International Conference on World Wide Web, Beijing, China, pp 695–704

Leskovec J, Lang KJ, Mahoney MW (2010) Empirical comparison of algorithms for network community detection. In: The 19th International Conference on World Wide Web, Raleigh, NC, pp 631–640

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. Ann Rev Sociol 27:415–444

Portes A (1998) Social capital: Its origins and applications in modern sociology. Ann Rev Sociol 24:1–24

Recommended Reading

Blau P, Schwartz J (1997) Crosscutting social circles: Testing a macro-structural theory of intergroup relations. Transaction Publishers, Piscataway, NJ

Burt RS (2005) Brokerage and closure: An introduction to social capital. Oxford University Press, Oxford, UK

Gellner E (2003) Trust, cohesion, and the social order. In: Hechter M, Horne C (eds) Theories of social order: A reader. Stanford University Press, Stanford, CA, pp 300–305

Henderon K, Eliassi-Rad T, Papadimitriou S, Faloutsos C (2010) HCDF: A hybrid community discovery framework. In: The 10th SIAM International Conference on Data Mining, Columbus, OH, pp 754–765

Lazarsfeld P, Merton RK (1954) Friendship as a social process: A substantive and methodological analysis. In: Berger M, Abel T, Page CH (eds) Freedom and control in modern society, Van Nostrand, pp 18–66

Newman MEJ (2006) Modularity and community structure in networks. Proc Natl Acad Sci USA 103: 8577–8582

Simmel G (2003) The web of group-affiliations. In: Hechter M, Horne C (eds) Theories of social order: A reader. Stanford University Press, Stanford, CA, pp 316–322

Watts D (2004) Six degrees: The science of a connected age. W.W. Norton & Company, New York, NY

Social Provenance

Zhuo Feng, Pritam Gundecha, and Huan Liu
Data Mining and Machine Learning Lab, School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA

Glossary

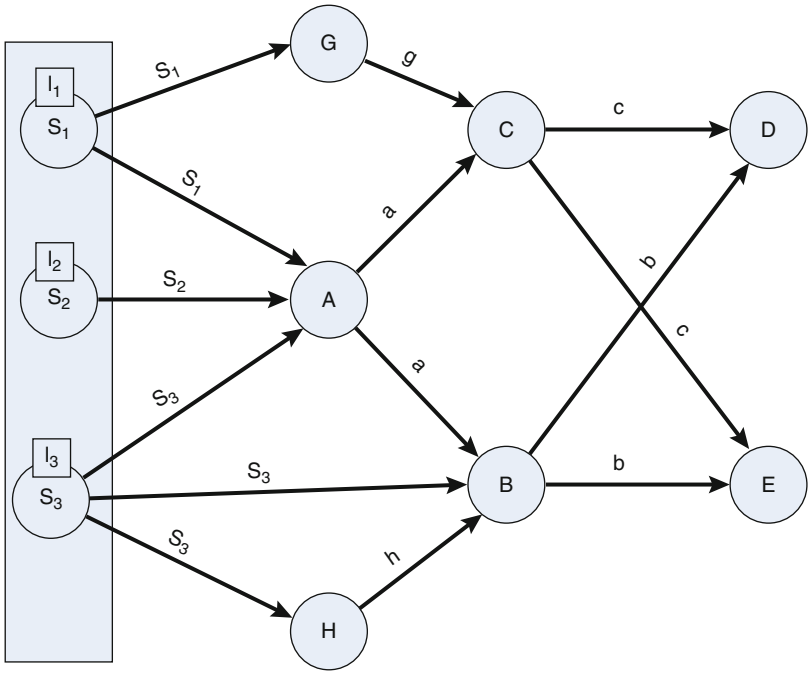
Provenance Sources of a piece of information

Provenance Paths Paths of information propagation from sources to terminals

Definition

Social Provenance

An information propagation network can be represented as a directed graph $G(V, E)$, where V is the node set and E is the edge set. Each node in the graph represents the entity, which publishes a piece of information on social media. Entity may refer to an individual user or a webpage. A directed edge between nodes represents the direction of information flow. For a given piece of information propagating through the social media, the ***social provenance*** informs a user about the



Social Provenance, Fig. 1 Information propagation in social media

sources of a given piece of information. Sources refer to the nodes that first publish the concerned messages.

Figure 1 shows an information propagation graph indicating the flow of information $I = \{I_1; I_2; I_3\}$ which is about the same event. S_1, S_2, and S_3 are the source nodes, or the originators of I_1, I_2, and I_3, respectively. The information is transmitted through different nodes in social media or recipients. These nodes propagate information; some may retransmit it with modifications. Each edge is labeled with the information indicating where it comes from, e.g., "a" on edge "A–C" means that it is from "A." A social provenance problem is to help a recipient (say, node D) to answer what are possible information sources in social media for a given piece of information. A provenance path delineates how information spreads from a source to a recipient, including those responsible for retransmitting the information from the sources through intermediaries. If the provenance paths are known, the sources of information can be determined. More often than not, however, provenance paths of a known piece of information are unknown.

Provenance has been studied in the data management field. In data management, provenance represents the creator of the data and how data has been modified and transferred. Provenance information is used to determine the authenticity and trustworthiness of information. Provenance is the key to solve the data conflict problem (Moreau 2009). Unlike social media, data propagation can be captured in data management systems. Social provenance has received little attention in the literature. Shah and Zaman (2011) proposed a centrality-based method to determine the single information source among all known recipients on an undirected network. It assumes that information spread on a network follows the susceptible infected (SI) model. Since this method requires the knowledge of all recipients, it is not practical for social provenance. Also, the source computed using this method is more biased towards higher-degree nodes. Barbier, in his dissertation (Barbier 2012), proposed a method to collect metadata about the received information. Such metadata is referred as ***provenance attributes***. Provenance attributes can play a vital role in obtaining social provenance.

Social Provenance

Social media can help in solving the problem of social provenance due to its unique features: user-generated content (e.g., tweets, blog posts, news articles), user profiles, user interactions (e.g., links between friends, hyperlinks on the blog, or news articles), and spatial or temporal information. These features can help reconstruct an information propagation network of a given message, and the network is essential for social provenance.

The ***social provenance problem*** answers which nodes are the possible sources of some particular information, say a text message. The ***provenance path problem*** seeks to identify the paths that allow us to trace back possible sources. Solving the social provenance problem entails solving the provenance path problem. We present some key research issues in this burgeoning area below:

(a) What are the characteristics of sources such that we can identify a source when we encounter one? It is a challenging task because source nodes are not necessarily those without incoming links in social media networks.
(b) How can we use different parts of social media data for inferring provenance paths? Content, user profiles, and interaction patterns can play complementary roles in backtracking information propagation. As a popular source can lead to a shallow cascade (Leskovec et al. 2009), the study of node centrality measures can be of help.
(c) How can we infer missing links in reconstructing a provenance path with partial information? By the nature of social media, most information is informal and partial. Links can expand the network (i.e., new nodes can be added), and data associated with a node provides more information, though still partial.
(d) How can we limit the search space in the vast land of social media? It is incumbent to develop a scalable solution for the social provenance problem.
(e) What are effective and objective ways of verifying and comparing different approaches to social provenance and provenance path problems? Lack of ground truth constitutes one of the foremost difficulties.

An Illustrative Example and Impact

One of the important applications of social provenance is to find the rumormongers or disinformation centers in social media. The "Assam Exodus" is a recent example that illustrates the importance of social provenance. Assam is a large state in the northeast of India and a series of riots broke out in July and August 2012. Following the riots, virulent messages along with disinformation were spread in other parts of India via social media. Bulk text messages (short message services, SMS) and social media sites were extensively used to spread information, aiming to incite certain Indian population against the Northeast Indian population. For example, a Wall Street journalist reported that a twitter user used a gory video clip on riots in Indonesia as that of Assam riots (Twitter 2012). Violent messages were also spread on Facebook that incite hatred and vengeance against the Northeast Indian population (Facebook 2012). The disinformation as well as virulent messages resulted in deep fear among Northeast Indian population, which ultimately led to their exodus from some major metropolitan cities across India, which includes Bangalore, Mumbai, Hyderabad, Chennai, and Pune (Wikipedia 2012). In all of these cases, social provenance might be able to help to find the rumormongers or disinformation sources early and to help stop the viral spread of disinformation.

The social provenance problem presents an unprecedented challenge and its research progress can pave way for many equally challenging and important issues such as source trustworthiness, information reliability, and user credibility.

Cross-References

▸ Social Media
▸ Trust in Social Networks
▸ User Behavior in Online Social Networks, Influencing Factors

References

Barbier G (2012) Finding provenance data in social media. Doctoral dissertation, Arizona State University

Facebook (2012) https://www.facebook.com/photo.php?fbid=268506716591158 & set=a.247241168717713 49889.247222755386221&type=3&theater. Accessed 2 Oct 2013

Leskovec J, Backstrom L, Kleinberg J (2009) Meme-tracking and the dynamics of the news cycle. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining, Paris, France. ACM, pp 497–506

Moreau L (2009) The foundations for provenance on the web. Found Trends Web Sci 2:99–241

Shah D, Zaman T (2011) Rumors in a network: who's the culprit? IEEE Trans Inf Theory 57:5163–5181

Twitter (2012) https://twitter.com/dhume01/status/236321660184178688. Accessed 2 Oct 2013

Wikipedia (2012) http://en.wikipedia.org/wiki/2012_Assam_violence#Attacks_o_people_from_North_East.Exodus. Accessed 2 Oct 2013

Social Recommendation in Dynamic Networks

Hao Ma[1], Irwin King[2], and Michael R. Lyu[2]
[1]Microsoft Research, Redmond, WA, USA
[2]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

Synonyms

Collaborative filtering; Matrix factorization; Social network analysis; Social recommender system

Glossary

Recommender System A system that provides recommendations for users

Collaborative Filtering A type of recommendation technique

Social Relations Various social relationships between users, like social trust relationships

Matrix Factorization Factorizing the user-item matrix into user latent matrix and item latent matrix

Definition

The research of social recommendation aims at modeling recommender systems more accurately and realistically. The characteristic of social recommendation that is different from the tradition recommender system is the availability of social network, i.e., relational information among the users. Social recommendation focuses on how to utilize user social information to effectively and efficiently compute recommendation results.

Introduction

As the exponential growth of information generated on the World Wide Web, the *Information Filtering* techniques like *recommender systems* have become more and more important and popular. Recommender systems form a specific type of information filtering technique that attempts to suggest information items (movies, books, music, news, Web pages, images, etc.) that are likely to interest the users. Typically, recommender systems are based on *collaborative filtering*, which is a technique that automatically predicts the interest of an active user by collecting rating information from other similar users or items. The underlying assumption of collaborative filtering is that the active user will prefer those items which other similar users prefer (Ma et al. 2007). Based on this simple but effective intuition, collaborative filtering has been widely employed in some large, well-known commercial systems, including product recommendation at Amazon and movie recommendation at Netflix.

Due to the potential commercial values and the great research challenges, recommendation techniques have drawn much attention in data mining, information retrieval, and machine learning communities. Recommendation algorithms suggesting personalized recommendations greatly increase the likelihood of customers making their purchases online.

Traditional recommender systems assume that users are independent and identically distributed. This assumption ignores the social relationships among the users. But the fact is, offline, social

recommendation is an everyday occurrence. For example, when you ask a trusted friend for a recommendation of a movie to watch or a good restaurant to dine, you are essentially soliciting a verbal social recommendation. In (2001), Sinha and Swearingen have demonstrated that, given a choice between recommendations from trusted friends and those from recommender systems, in terms of quality and usefulness, trusted friends' recommendations are preferred, even though the recommendations given by the recommender systems have a high novelty factor. Trusted friends are seen as more qualified to make good and useful recommendations compared to traditional recommender systems (Bedi et al. 2007). From this point of view, the traditional recommender systems that ignore the social network structure of the users may no longer be suitable.

Thanks to the popularity of the Web 2.0 applications, recommender systems are now associated with various kinds of social information. This kind of information contains abundant additional information about users, hence providing a huge opportunity to improve the recommendation quality. For example, in users' social trust network, users tend to share their similar interests with the friends they trust. In reality, we always turn to friends we trust for movie, music, or book recommendations, and our tastes and characters can be easily affected by the company we keep. Hence, how to incorporate social information into the recommendation algorithms becomes a trend in the research of recommender systems.

Historical Background

As mentioned in Huang et al. (2004), one of the most commonly used and successfully deployed recommendation approaches is collaborative filtering. In the field of collaborative filtering, two types of methods are widely studied: neighborhood-based approaches and model-based approaches.

Neighborhood-based methods mainly focus on finding the similar users (Breese et al. 1998; Jin et al. 2004) or items (Deshpande and Karypis 2004; Linden et al. 2003; Sarwar et al. 2001) for recommendations. User-based approaches predict the ratings of active users based on the ratings of similar users found, while item-based approaches predict the ratings of active users based on the computed information of items similar to those chosen by the active user. User-based and item-based approaches often use Pearson Correlation Coefficient (PCC) algorithm (Resnick et al. 1994) and Vector Space Similarity (VSS) algorithm (Breese et al. 1998) as the similarity computation methods. PCC method can generally achieve higher performance than VSS approach, since the former considers the differences of user rating style.

In contrast to the neighborhood-based approaches, the model-based approaches to collaborative filtering use the observed user-item ratings to train a compact model that explains the given data, so that ratings could be predicted via the model instead of directly manipulating the original rating database as the neighborhood-based approaches do (Liu and Yang 2008). Algorithms in this category include the clustering model (Kohrs and Merialdo 1999), the aspect models (Hofmann 2003, 2004; Si and Jin 2003), the latent factor model (Canny 2002), the Bayesian hierarchical model (Zhang and Koren 2007), and the ranking model (Liu and Yang 2008). Kohrs and Merialdo (1999) presented an algorithm for collaborative filtering based on hierarchical clustering, which tried to balance both robustness and accuracy of predictions, especially when few data were available. Hofmann (2003) proposed an algorithm based on a generalization of probabilistic latent semantic analysis to continuous-valued response variables.

Recently, due to the efficiency in dealing with large datasets, several low-dimensional matrix approximation methods (Rennie and Srebro 2005; Salakhutdinov and Mnih 2008a, b; Srebro and Jaakkola 2003) have been proposed for collaborative filtering. These methods all focus on fitting the user-item rating matrix using low-rank approximations and employ the matrix to make further predictions. The Low-rank matrix factorization methods are very

efficient in training since they assume that in the user-item rating matrix, only a small number of factors influence preferences and that a user's preference vector is determined by how each factor applies to that user. Low-rank matrix approximations based on minimizing the sum-squared errors can be easily solved using Singular Value Decomposition (SVD), and a simple and efficient Expectation Maximization (EM) algorithm for solving weighted low-rank approximation is proposed in Srebro and Jaakkola (2003). In (2004), Srebro et al. proposed a matrix factorization method to constrain the norms of U and V instead of their dimensionality. Salakhutdinov and Mnih presented a probabilistic linear model with Gaussian observation noise in (2008b). In Salakhutdinov and Mnih (2008a), the Gaussian-Wishart priors are placed on the user and item hyperparameters.

Traditional recommender systems have been well studied and developed both in academia and in industry, but they are all based on the assumption that users are independent and identically distributed, and ignore the relationships among users. Based on this intuition, many researchers have recently started to analyze trust-based recommender systems (Bedi et al. 2007; Massa and Avesani 2004, 2007; O'Donovan and Smyth 2005).

Bedi et al. in (2007) proposed a trust-based recommender system for the Semantic Web; this system runs on a server with the knowledge distributed over the network in the form of ontologies and employs the Web of trust to generate the recommendations. In Massa and Avesani (2004), a trust-aware method for recommender system is proposed. In this work, the collaborative filtering process is informed by the reputation of users, which is computed by propagating trust. Trust values are computed in addition to similarity measures between users. The experiments on a large real dataset show that this work increases the coverage (number of ratings that are predictable) while not reducing the accuracy (the error of predictions). In O'Donovan and Smyth (2005), two trust-aware methods are proposed to improve standard collaborative filtering methods. The experimental analysis shows that these trust information can help increase recommendation accuracy.

Previously proposed trust-aware methods are all neighborhood-based methods which employ only heuristic algorithms to generate recommendations. There are several problems with this approach, however. The relationship between the trust network and the user-item matrix has not been studied systematically. Moreover, these methods are not scalable to very large datasets since they may need to calculate the pairwise user similarities and pairwise user trust scores.

Social Recommendation Using Matrix Factorization

Matrix Factorization

In this subsection, we review one popular matrix factorization method that is widely studied in the literature.

Considering an $m \times n$ matrix R describing m users' ratings on n items, a low-rank matrix factorization approach seeks to approximate the frequency matrix R by a multiplication of d-rank factors $R \approx U^T V$, where $U \in \mathbb{R}^{d \times m}$ and $V \in \mathbb{R}^{d \times n}$ with $d \ll \min(m, n)$. The matrix R in the real world is usually very sparse since most of the users only visited a few Web sites.

Traditionally, the Singular Value Decomposition (SVD) method is employed to estimate a matrix R by minimizing

$$\min_{U,V} \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} I_{ij} (r_{ij} - \mathbf{u}_i^T \mathbf{v}_j)^2, \qquad (1)$$

where $\mathbf{u}_i$ and $\mathbf{v}_j$ are column vectors with d values and I_{ij} is the indicator function that is equal to 1 if user i rated item j and equal to 0 otherwise.

In order to avoid overfitting, two regularization terms are added into (1). Hence we have the following Regularized SVD equation:

$$\min_{U,V} \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} I_{ij} (r_{ij} - \mathbf{u}_i^T \mathbf{v}_j)^2 + \frac{\lambda_1}{2} ||U||_F^2 + \frac{\lambda_2}{2} ||V||_F^2, \qquad (2)$$

where $\lambda_1, \lambda_2 > 0$. The optimization problem in (2) minimizes the sum-of-squared-errors objective function with quadratic regularization terms. Gradient-based approaches can be applied to find a local minimum. It also contains a nice probabilistic interpretation with Gaussian observation noise, which is detailed in Salakhutdinov and Mnih (2008b). In Salakhutdinov and Mnih (2008b), the conditional distribution over the observed data is defined as

$$p(R|U,V,\sigma_R^2) = \prod_{i=1}^{m}\prod_{j=1}^{n} \times \left[\mathcal{N}\left(r_{ij}|\mathbf{u}_i^T\mathbf{v}_j,\sigma_R^2\right)\right]^{I_{ij}}, \tag{3}$$

where $\mathcal{N}(x|\mu,\sigma^2)$ is the probability density function of the Gaussian distribution with mean μ and variance σ^2. The zero-mean spherical Gaussian priors are also placed on user and item feature vectors:

$$p(U|\sigma_U^2) = \prod_{i=1}^{m}\mathcal{N}(\mathbf{u}_i|0,\sigma_U^2\mathbf{I}),$$
$$p(V|\sigma_V^2) = \prod_{j=1}^{n}\mathcal{N}(\mathbf{v}_j|0,\sigma_V^2\mathbf{I}). \tag{4}$$

Through a Bayesian inference, we can easily obtain the objective function in (2).

By adopting a simple stochastic gradient descent technique, for each observed rating r_{ij}, we have the following efficient updating rules to learn latent variables $\mathbf{u}_i, \mathbf{v}_j$:

$$\mathbf{u}_i \leftarrow \mathbf{u}_i + \gamma_1(\Delta_{ij}\mathbf{v}_j - \lambda_1\mathbf{u}_i),$$
$$\mathbf{v}_j \leftarrow \mathbf{v}_j + \gamma_2(\Delta_{ij}\mathbf{u}_i - \lambda_2\mathbf{v}_j), \tag{5}$$

where $\Delta_{ij} = r_{ij} - \mathbf{u}_i^T\mathbf{v}_j$, and γ_1, γ_2 are the learning rates.

The Regularized SVD algorithm introduced in this section is both effective and efficient in solving the collaborative filtering problem, and it is perhaps one of the most popular methods in collaborative filtering.

Social Trust Ensemble

However, the above algorithm does not consider any information from users' social network. In order to better model the recommendation problem, in Ma et al. (2009), Ma et al. proposed a matrix factorization-based Social Trust Ensemble (STE) method upon the following intuitions:

- Users have their own tastes.
- Users can also be easily influenced by the trusted friends they have.
- A user's final rating is composed of the combination of this user's own taste and this user's friends' tastes.

Based on the above interpretations, the objective function can be formulated as

$$L = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}\left(r_{ij} - \left(\alpha\mathbf{u}_i^T\mathbf{v}_j + (1-\alpha)\sum_{k\in\mathcal{T}(i)} w_{ik}\mathbf{u}_k^T\mathbf{v}_j\right)\right)^2 + \frac{\lambda_1}{2}\|U\|_F^2 + \frac{\lambda_2}{2}\|V\|_F^2, \tag{6}$$

where α is a parameter to balance the impact of user's own taste and user's friends' tastes, $\mathcal{T}(i)$ represents a list of user i's trusted friends, and w_{ik} is a normalized weight that equals to $1/|\mathcal{T}(i)|$.

We can see that in this approach, a user's latent factor is smoothly integrated with this user's trusted friends' tastes. This equation also coincides with the real-world observation that we always ask our friends for movies, books, or music recommendations.

For each observed rating r_{ij}, the stochastic gradient decent learning rules for this method are

$$\mathbf{u}_i \leftarrow \mathbf{u}_i + \gamma_1\Big(\Delta_{ij}\Big(\alpha + (1-\alpha)\sum_{p\in\mathcal{B}(i)} w_{pi}\Big)\mathbf{v}_j - \lambda_1\mathbf{u}_i\Big),$$

$$\mathbf{v}_j \leftarrow \mathbf{v}_j + \gamma_2\Big(\Delta_{ij}\Big(\alpha\mathbf{u}_i + (1-\alpha)\sum_{k\in\mathcal{T}(i)} w_{ik}\mathbf{u}_k\Big) - \lambda_2\mathbf{v}_j\Big), \tag{7}$$

where

$$\Delta_{ij} = r_{ij} - \Big(\alpha\mathbf{u}_i^T\mathbf{v}_j + (1-\alpha)\sum_{k\in\mathcal{T}(i)} w_{ik}\mathbf{u}_k^T\mathbf{v}_j\Big), \tag{8}$$

and $\mathcal{B}(i)$ is the set that includes all the users who trust user i.

Social Regularization

The STE method mentioned above is originally designed for trust-aware recommender systems. In trust-aware recommender systems, we can always assume that users have similar tastes with other users they trust. Unlike trust relationships among users, the tastes among social friend relationships are more diverse. User k is a friend of user i does not necessarily indicate that user k has similar taste with user i. Hence, in order to model the social recommendation problems more accurately, another more general social recommendation approach, Social Regularization (SR), is proposed in Ma et al. (2011).

The objective function of this approach is formulated as

$$\begin{aligned} L = {} & \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}(r_{ij} - \mathbf{u}_i^T\mathbf{v}_j)^2 \\ & + \frac{\alpha}{2}\sum_{i=1}^{m}\sum_{f\in\mathcal{F}^+(i)} s_{if}\|\mathbf{u}_i - \mathbf{u}_f\|_F^2 \\ & + \frac{\lambda_1}{2}\|U\|_F^2 + \frac{\lambda_2}{2}\|V\|_F^2, \end{aligned} \tag{9}$$

where s_{if} indicates the similarity between user i and user f and $\mathcal{F}^+(i)$ represents user i's outlink friends.

In this method, the social network information is employed in designing the social regularization term to constrain the matrix factorization objective function. The social regularization term also indirectly models the propagation of tastes. More specifically, if user i has a friend f and user f has a friend user g, this regularization term actually indirectly minimizes the distance between latent vectors $\mathbf{u}_i$ and $\mathbf{u}_g$. The propagation of tastes will reach a harmonic status once the learning is converged.

Similarly, for each observed rating r_{ij}, we have the following stochastic gradient descent updating rules to learn the latent parameters:

$$\begin{aligned} \mathbf{u}_i &\leftarrow \mathbf{u}_i + \gamma_1\Big(\Delta_{ij}\mathbf{v}_j - \alpha\sum_{f\in\mathcal{F}^+(i)} s_{if}(\mathbf{u}_i - \mathbf{u}_f) - \alpha\sum_{g\in\mathcal{F}^-(i)} s_{ig}(\mathbf{u}_i - \mathbf{u}_g) - \lambda_1\mathbf{u}_i\Big), \\ \mathbf{v}_j &\leftarrow \mathbf{v}_j + \gamma_2(\Delta_{ij}\mathbf{u}_i - \lambda_2\mathbf{v}_j), \end{aligned} \tag{10}$$

where $\Delta_{ij} = r_{ij} - \mathbf{u}_i^T\mathbf{v}_j$, and $\mathcal{F}^-(i)$ represents user i's inlink friends.

The experiments conducted in Ma et al. (2009, 2011) suggest that social recommendation algorithms outperform traditional recommendation algorithms, especially when the user-item matrix is sparse. This indicates that using social information is a promising direction in the research of recommender systems.

Future Directions

The methods mentioned above can be solved efficiently by using simple gradient descent or stochastic gradient descent algorithms. However, for statistical machine learning's point of view, the methods themselves are not full Bayesian methods. Hence, learning those methods can easily have the overfitting problem. How to apply full Bayesian method on these models hence becomes worth of studying.

We already demonstrate how to recommend by incorporating users' social trust and friend information. Actually, sometimes there are more data sources available on Web 2.0 sites, such as tags issued by users to items and temporal information. These sources are also valuable information to improve recommender systems.

Acknowledgments

The work described in this article is supported by grants from the Research Grant Council of the Hong Kong Special Administrative Region, China (Project No.: CUHK 413212 and CUHK 415212).

Cross-References

- ▸ Human Behavior and Social Networks
- ▸ Inferring Social Ties
- ▸ Matrix Decomposition
- ▸ Probabilistic Graphical Models
- ▸ Recommender Systems Using Social Network Analysis: Challenges and Future Trends
- ▸ Social Recommender System

References

Bedi P, Kaur H, Marwaha S (2007) Trust based recommender system for semantic web. In: Proceedings of IJCAI'07, Hyderabad, pp 2677–2682

Breese JS, Heckerman D, Kadie C (1998) Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of UAI'98, Madison

Canny J (2002) Collaborative filtering with privacy via factor analysis. In: Proceedings of SIGIR'02, Tampere, pp 238–245

Deshpande M, Karypis G (2004) Item-based top-n recommendation. ACM Trans Inf Syst 22(1):143–177

Hofmann T (2003) Collaborative filtering via gaussian probabilistic latent semantic analysis. In: Proceedings of SIGIR'03, Toronto, pp 259–266

Hofmann T (2004) Latent semantic models for collaborative filtering. ACM Trans Inf Syst 22(1):89–115. doi:http://doi.acm.org/10.1145/963770.963774

Huang Z, Chen H, Zeng D (2004) Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. ACM Trans Inf Syst 22(1): 116–142

Jin R, Chai JY, Si L (2004) An automatic weighting scheme for collaborative filtering. In: Proceedings of SIGIR'04, Sheffield, pp 337–344

Kohrs A, Merialdo B (1999) Clustering for collaborative filtering applications. In: Proceedings of CIMCA, Gold Coast

Linden G, Smith B, York J (2003) Amazon.com recommendations: item-to-item collaborative filtering. IEEE Intern Comput 76–80

Liu NN, Yang Q (2008) Eigenrank: a ranking-oriented approach to collaborative filtering. In: Proceedings of SIGIR'08, Singapore, pp 83–90

Ma H, King I, Lyu MR (2007) Effective missing data prediction for collaborative filtering. In: Proceedings of SIGIR'07, Amsterdam, pp 39–46

Ma H, King I, Lyu MR (2009) Learning to recommend with social trust ensemble. In: Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval, SIGIR'09, Boston, pp 203–210

Ma H, Zhou D, Liu C, Lyu MR, King I (2011) Recommender systems with social regularization. In: Proceedings of the fourth ACM international conference on Web search and data mining, WSDM'11, Hong Kong, pp 287–296

Massa P, Avesani P (2004) Trust-aware collaborative filtering for recommender systems. In: Proceedings of CoopIS/DOA/ODBASE, Irvine, pp 492–508

Massa P, Avesani P (2007) Trust-aware recommender systems. In: Proceedings of RecSys'07, Minneapolis, pp 17–24

O'Donovan J, Smyth B (2005) Trust in recommender systems. In: Proceedings of IUI'05, San Diego, pp 167–174

Rennie JDM, Srebro N (2005) Fast maximum margin matrix factorization for collaborative prediction. In: Proceedings of ICML'05, Bonn

Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl J (1994) Grouplens: an open architecture for collaborative filtering of netnews. In: Proceedings of CSCW'94, Chapel Hill

Salakhutdinov R, Mnih A (2008a) Bayesian probabilistic matrix factorization using markov chain monte carlo. In: Proceedings of ICML'08, Helsinki

Salakhutdinov R, Mnih A (2008b) Probabilistic matrix factorization. In: Proceedings of NIPS'08, vol 20, Vancouver
Sarwar B, Karypis G, Konstan J, Reidl J (2001) Item-based collaborative filtering recommendation algorithms. In: Proceedings of WWW'01, Hong Kong, pp 285–295
Si L, Jin R (2003) Flexible mixture model for collaborative filtering. In: Proceedings of ICML'03, Washington, DC
Sinha RR, Swearingen K (2001) Comparing recommendations made by online systems and friends. In: DELOS workshop: personalisation and recommender systems in digital libraries, Dublin
Srebro N, Jaakkola T (2003) Weighted low-rank approximations. In: Proceedings of ICML'03, Washington, DC, pp 720–727
Srebro N, Rennie JDM, Jaakkola T (2004) Maximum-margin matrix factorization. In: Proceedings of NIPS'04, Vancouver
Zhang Y, Koren J (2007) Efficient bayesian hierarchical user modeling for recommendation system. In: Proceedings of SIGIR'07, Amsterdam, pp 47–54

Social Recommender System

▸ Recommender Systems, Semantic-Based
▸ Social Recommendation in Dynamic Networks

Social Reconnaissance

▸ Reconnaissance and Social Engineering Risks as Effects of Social Networking

Social Relationships

▸ Actionable Information in Social Networks, Diffusion of
▸ Inferring Social Ties

Social Selection

▸ Stability and Evolution of Scientific Networks

Social Spam

▸ Spam Detection, E-mail/Social Network

Social Status, Social Location, Position

▸ Role Discovery

Social Structural Analysis

▸ Origins of Social Network Analysis

Social Tagging

▸ Folksonomies

Social Tags

▸ Social Bookmarking

Social Theory

▸ Web Communities Versus Physical Communities

Social Ties Inferring

▸ Link Prediction: A Primer

Social Trends Discovery

▸ Semantic Social Networks Analysis

Social Trust

▸ Social Interaction Analysis for Team Collaboration

Social Web Search

Maryam Shoaran, Alex Thomo, and Jens Weber
Department of Computer Science, University of Victoria, Victoria, BC, Canada

Synonyms

Recommender systems; Social analysis; Social content search; Social navigation

Glossary

Social Media Online systems with high public participation and interaction rate

User Metadata Data created as the result of user interactions in an information space

Socially Enhanced Search Quality-improved search resulting from employing user metadata

Personalization Adjustment of a system or process to fit user preferences

Blogosphere Collection of interconnected Web logs

Facebook Graph Search Information lookup in the Facebook graph-structured data

Collaborative Filtering Discovery of new knowledge and patterns through filtering data produced by collaboration between different individuals

Definition

Social search is an online search process that employs user-generated data and user-user relationships produced by social systems including bookmarking sites, Web forums, social networks, and blogs to discover the best matching content to user queries in an information space. This is different from the methods used in traditional Web search engines in the sense that search techniques in the latter are mostly based on page-author-generated data such as page content, anchor text, and link connections. User-created data forms a rich source of metadata that expresses single-user or community preferences, ideas, and needs. User tags and queries can be considered as new descriptions of Web page content. Social search utilizes this new and fast expanding source of information to establish a fine-grained and more personalized or community-based online search.

A variety of information systems ranging from the World Wide Web to special purpose social systems, such as social networks, bookmarking sites, document-or media-sharing communities, and e-commerce, benefit from the capabilities of social search. In the literature, social search also refers to the process of the analysis and discovery of new knowledge from social media.

Introduction

The objective of an online search system is to locate the relevant objects (e.g., Web pages) to a user-generated query from the Web-or a community-based collection. Over the decades, Web search engines have improved their quality of search by inventing new techniques to retrieve query-relevant documents and rank them based on their quality. The characteristic of almost all of these techniques is that they are based on the data created by the Web page builders or document authors. Two types of ranking methods are used in search engines: first, query-dependent or similarity measures that use document content, title, and anchor text to find similar documents and second, query-independent or static measures that use page connectivity (link structure) as a quality measure to rank similar documents. The prominent static metrics are PageRank (Page et al. 1999) and HITS (Kleinberg 1999).

Recently, with the ever-increasing activity and popularity of social media, a new type of

information – user-created metadata – is available that can be used to enhance the quality of search.

User-generated content can be categorized as *explicit* or *implicit*. Explicit user data is created by visitors of Web sites in the form of annotations and viewpoints in order to describe, organize, and share their favorite entities (URLs, movies, songs, books, articles, etc.) online. Social systems capture explicit user annotations and viewpoints (feedback) in different forms. For example, book, article, and movie review sites collect user reviews and ratings as text and star points. Social bookmarking sites store user tags and favorite URLs, and social networks capture user comments and their likes. User annotations and viewpoints constitute a precious information source that can be utilized to extract, for example, Web page descriptions (using tags and comments), page or media popularities (using bookmarks and ratings), and user preferences (using ratings and likes).

Monitoring user online behavior builds another valuable source of information. Implicit user data is automatically extracted from system logs containing user search queries, browsing history (clickthrough data), and amount of time spent by users on different pages. This data can help improve the quality of search in different ways. For instance, user queries can be considered as "URL tags" describing the content of pages. User browsing history is an indication of user interest and can be used to resolve the ambiguity that often exists in user queries. The amount of time spent by users reading the content of Websites might be an indication of the importance of sites and can be used to improve the ranking process, especially in community-based search environments.

Social Web Search and Analysis

Online social systems holding a rich public participation have been able to accumulate valuable and heterogeneous collections of user metadata. Social search and analysis is focused on taking advantage of such data sources by new techniques that either help to improve the functionality of already existing systems or devise novel analysis and knowledge discovery schemes. Social search and analysis is active in different areas as follows: (1) socially enhanced Web search, (2) social navigation, (3) social analysis, (4) recommender systems, and (5) social content search.

Socially Enhanced Web Search

Social Web search aims at improving the quality of Web search by combining traditional search methods, e.g., query-document similarity and PageRank, with new techniques that employ social content. For instance, *SocialSimRank* (SSR) and *SocialPageRank* (SPR) (Bao et al. 2007) are two new methods that integrate social annotations available in social bookmarking sites, e.g., del.icio.us, into the page ranking process.

SocialSimRank (SSR) is a similarity ranking algorithm for queries and social annotations. The algorithm is based on the assumption that social annotations provide good summaries of Web pages from various user perspectives. Based on this observation, similarities for every pair of annotations and similarities for every pair of pages are iteratively computed. The similarities are recursively defined as follows. The more similar the pages are, the more similar their corresponding annotations are. Conversely, the more similar annotations are, the more similar their associated pages are. These similarities are integrated into each other's computation. That is, in the equation that calculates the similarity between two annotations, one of the parameters is the similarity between two pages to which these annotations are assigned, and vice versa. After several iterations, this process typically converges, and the system is ready to answer queries. Each query term is considered to be a page annotation. The similarity of a query q to a Web page p is computed as the sum of the similarities of each term in q to each annotation associated with p.

SocialPageRank (SPR) computes the page quality (popularity) with the intuition that the number of annotations assigned to a Web page indicates the quality of the page in some sense. SPR uses an iterative algorithm to compute page popularities based on user and annotation popularities. Integrating SSR and SPR into

a ranking function that also uses traditional documentsimilarity metric and PageRank improves the quality of Web search (Bao et al. 2007).

Other enhanced search methods are also proposed that benefit from different aspects of user annotations. A hybrid search technique is presented in Yanbe et al. (2007) that combines a link-based ranking method with a new metric that is based on user-generated data in social bookmarking sites (e.g., del.icio.us). The new metric utilizes SBRank (Social Bookmarking Rank) which is the number of user bookmarks on a page, the sentiment-based and temporal information extracted from user annotations, as well as general statistics derived from user interactions with Web pages.

Community-based search systems improve the quality of search by incorporating user search behaviors within the community, e.g., user queries and result selections, into the ranking method. The underlying intuition is that among the users of similar mind, e.g., social network or enterprise intranet users, the context of queries is similar, the query repetition is high, and also there rarely exist malicious behaviors that can negatively affect popularity metrics (Freyne et al. 2007). This type of social search is also called *Collaborative Web Search* (CWS) (Morris and Teevan 2009), and I-SPY (Smyth et al. 2004) is an implementation of it. Such systems record the queries and result selection of the community searchers, and upon exposure to a new query, the information of search sessions of a similar pattern is retrieved. The system re-ranks the result returned by the underlying search engines to reflect the implicit preferences of the community. Each item in the result list is also augmented by a set of past related queries that can be used to start new searches.

Social Navigation

The goal of social navigation is to enhance the quality of user browsing by providing various types of navigational assistance based on the visiting behavior of similar-minded users in the past. Social navigation systems benefit from different implicit and explicit user-generated data. They keep track of the browsing behavior of the users by collecting user queries and browsing paths (personal footprints). The time spent reading a page is also taken into account as an indication of user intention. Such systems also benefit from user annotations that can provide useful information about the importance of visited pages. When a user clicks on a source or on a (page) link in the search result list, the system provides a visual guide containing different navigational cues, for example, the source or page visit frequency (browsing popularity), the number of associated annotations (annotation popularity), and a list of queries leading to this source or page (search popularity).

Knowledge Sea II (Brusilovsky et al. 2004) is an example of a social browsing system that was developed to help students in a class to find the most useful sources for a particular course. This system organizes sources in a table with each cell associated with one source. The available navigational clues include the background color of cells indicating visit frequency, a sticky note for the presence of annotations, and a thermometer representing the number of positive annotations.

Another interesting system is the one presented in Freyne et al. (2007) that facilitates community-based access to the Communications of the ACM (CACM) magazine. This system integrates social search and social navigation in both the interface level and its internal mechanisms. When a new search query is initiated, the search component of the system retrieves similar queries and their associated search results. Then, the results are scored based on their relevance to the new query, and finally the top-k results are placed ahead of the other results returned by the ACM search engine. Each result item is appended by complementary information presented as icons. Five icons with different levels of filling indicate, respectively, (1) the relevance of the result to the query (the percentage of times the result has been selected for the query by community users), (2) a list of other queries that have led to the selection of this result by community users, (3) the last time the result was encountered by the

users (a view of the freshness), (4) the browsing popularity of the result (footprints), and (5) the user annotations. When a result is selected, the browsing component also augments the opened pages with social assistance icons.

Social Quality Analysis

The quality of user-generated content in online social systems varies from excellent to spam due to the participation of individuals with different intentions and levels of expertise. This is especially important in knowledge-based social systems such as question/answering portals, online forums, and networks of email exchangers. Social quality analysis aims at identifying knowledge experts and high-quality user-created content in order to improve the quality of information-retrieval tasks (cf. Zhang et al. 2007; Campbell et al. 2003; Agichtein et al. 2008; Yang et al. 2011). Various analysis methods are used ranging from link-based ranking algorithms, e.g., PageRank (Page et al. 1999) and HITS (Kleinberg 1999), to text classification techniques and user clickthrough information.

Since 2006, some interesting systems have been presented that automatically evaluate the quality of questions and answers in question/answering domains (cf. Jeon et al. 2006; Agichtein et al. 2008). The framework presented in Agichtein et al. (2008) first identifies a collection of quality-indicating features of social media and associated interactions. Then, these features are used as input to a classifier (a stochastic gradient boosted tree), in order to extract high-quality content. A wide range of information sources are used to extract features of the following categories: (1) contentbased: textual features of questions and answers, such as word n-grams, punctuation and typos, syntactic and semantic complexity measures, and grammaticality measures; (2) connectivitybased: link-based metrics (authority scores and PageRank) in user-item and user-user relationship graphs, where an item is a query or an answer; (3) usagebased: temporal statistics, number of clicks on items, and time spent on reading.

Recommender Systems

In online shopping, movie, and music Web sites, the goal is to improve the user experience by providing appropriate recommendations about new items that match user interests, ideas, and needs. These Web sites collect different types of user-produced data, ranging from explicit user ratings to implicit purchase history, browsing, and search activities. Recommender systems (Resnick and Varian 1997), using sophisticated algorithms, combine data from independent contributors to discover new knowledge about relations between users and items.

There are two major approaches in recommender systems: *content filtering* and *collaborative filtering* (Koren et al. 2009; Koren and Bell 2011). Content filtering discovers matching users and items based on their individual characteristics. Items (products) are profiled by domain experts and user profiles are created by users' explicit answers to specific questions, e.g., demographic questions. A problem with content-based filtering is the difficulty of gathering relevant information.

Collaborative filtering, on the other hand, is based on user behavior in the past, for example, user transactions and product ratings. By analyzing the relationships among users and among items, collaborative filtering predicts new relations between particular users and items. Suppose that in a movie rental site, user u has not watched and rated movie x yet, and the system would like to know whether it should recommend x to u. In the user-centered collaborative approach, first the similarity between u and all other users who have rated x is computed using some similarity measure, e.g., Euclidean distance or Pearson correlation coefficient. Then, the system predicts how u would rate x by computing a weighted average of the ratings for x by the most similar users to u. If the predicted rating is above a certain threshold, the system recommends x to u. In the item-centered approach the prediction is made based instead on the similarity between items.

Motivated by the Netflix prize contest, significant improvements have been made in the quality of recommender systems. *Latent factor* models

are another approach in collaborative filtering that maps the users and items to a common multidimensional space based on the past rating patterns. Latent factor models are based on *sparse matrix factorization*, and they are among the most popular and the best performing approaches (Koren 2009; Koren et al. 2009; Koren and Bell 2011).

Social Content Search

Despite the similarities between social media sites such as social networks, the blogosphere, and microblogging systems like Twitter, they differ in the type of data predominantly posted and shared by users, as well as in the form of user interconnections they offer. Based on these characteristics, special-purpose search and information discovery efforts are applicable on the content of each site (Facebookgraphsearch; Bansal and Koudas 2007; Mathioudakis and Koudas 2010).

Facebook has recently launched *Graph Search* (Facebookgraphsearch) as a new feature to benefit from its massive storage of data and relationships. Using this tool people can search for real-world objects in Facebook's knowledge graph, which is comprised of objects such as people, places, and things and inter-object connections, for example, Friendship and Likes. An important advantage of Facebook's search is that it has access to the collective knowledge of its vast community of users (more than one billion) to answer questions involving different layers of searching. Appealing examples are as follows: "What to read that is liked by my friends in college," "Where to eat in Toronto that my friends living there like," "Where to go in Asia that my friends and friends of friends of my age found interesting," "What iPhone app to download that my friends use to track their jogging and cycling," etc. People's experience with Facebook Graph Search will highly depend on the level of their connectedness and participation in the system.

Blogging is another online social activity that has received an increasing popularity in recent years. The free context of blogs makes the blogosphere (the collection of connected blogs) a rich source of heterogeneous information including personal experiences and opinions about a variety of subjects. Mining and analysis of blog data can capture public insight in different topics (Gruhl et al. 2005; Bansal and Koudas 2007). For instance, BlogScope (Bansal and Koudas 2007) is one of the systems designed to analyze the textual content of blogs and to provide information such as *when*, *where*, and *why* about interesting topics. When the user selects one of the daily hot keywords provided by the system or poses a query, all the relevant blog posts are retrieved and the result of various analysis on their content is presented. For example, the system can display the following information: (1) a popularity curve for a keyword as a function of time, (2) a list of the most closely related keywords in blog posts, (3) a distribution of the related posts on the map, and (4) a synopsis set which is the maximal set of keywords correlated with query that exhibits a bursty behavior in the associated popularity curve.

Future Directions

Whereas the usefulness of the annotations in small-scaled information communities has been demonstrated by several works, social annotations and bookmarks lack yet the sufficient size and quality to significantly influence the performance of search engines in a large scale (cf. Heymann et al. 2008; Bao et al. 2007). For example, the number of unique URLs in del.icio.us is relatively small in comparison with the indexes of the major search engines that include hundreds of billions of pages.

Augmenting more Web sites with improved tagging systems and also aggregating the data from various bookmarking and tagging Web sites would significantly help to improve the situation. The design of appealing and structured user interfaces that could provide a list of possible tags that do not appear in the page content and title can enhance the quality of tagging. Providing incentives, such as site access privileges or special offers to stimulate user tagging activities, could be another approach to create enriched user metadata.

Cross-References

▸ Microtext Processing
▸ Recommender Systems: Models and Techniques
▸ Social Bookmarking

References

Agichtein E, Castillo C, Donato D, Gionis A, Mishne G (2008) Finding high-quality content in social media. In: WSDM, Stanford, pp 183–194

Bansal N, Koudas N (2007) Blogscope: spatio-temporal analysis of the blogosphere. In: WWW, Banff, pp 1269–1270

Bao S, Xue G-R, Wu X, Yu Y, Fei B, Su Z (2007) Optimizing web search using social annotations. In: WWW, Banff, pp 501–510

Brusilovsky P, Chavan G, Farzan R (2004) Social adaptive navigation support for open corpus electronic textbooks. In: AH, Eindhoven, pp 24–33

Campbell CS, Maglio PP, Cozzi A, Dom B (2003) Expertise identification using email communications. In: CIKM, New Orleans, pp 528–531

del.icio.us. https://delicious.com/

Facebook graph search. https://www.facebook.com/about/graphsearch

Freyne J, Farzan R, Brusilovsky P, Smyth B, Coyle M (2007) Collecting community wisdom: integrating social search & amp; social navigation. In: IUI, Honolulu, pp 52–61

Gruhl D, Guha RV, Kumar R, Novak J, Tomkins A (2005) The predictive power of online chatter. In: KDD, Chicago, pp 78–87

Heymann P, Koutrika G, Garcia-Molina H (2008) Can social bookmarking improve web search? In: WSDM, Stanford, pp 195–206

Jeon J, Croft WB, Lee JH, Park S (2006) A framework to predict the quality of answers with non-textual features. In: SIGIR, Seattle, pp 228–235

Kleinberg JM (1999) Hubs, authorities, and communities. ACM Comput Surv 31(4es):5

Koren Y, Bell RM, Volinsky C (2009) Matrix factorization techniques for recommender systems. IEEE Comput 42(8):30–37

Koren Y, Bell RM (2011) Advances in collaborative filtering. In: Ricci F et al (eds) Recommender systems handbook. Springer, London/New York, pp 145–186

Koren Y, Bell RM, Volinsky C (2009) Matrix factorization techniques for recommender systems. IEEE Comput 42(8):30–37

Mathioudakis M, Koudas N (2010) Twittermonitor: trend detection over the twitter stream. In: SIGMOD Conference, Indianapolis, pp 1155–1158

Morris MR, Teevan J (2009) Collaborative web search: who, what, where, when, and why. Synthesis lectures on information concepts, retrieval, and services. Morgan & Claypool, San Rafael

Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking: bringing order to the web. Technical Report 1999–66, Stanford InfoLab, Nov 1999. Previous number = SIDL-WP-1999-0120

Resnick P, Varian HR (1997) Recommender systems – introduction to the special section. Commun ACM 40(3):56–58

Smyth B, Balfe E, Freyne J, Briggs P, Coyle M, Boydell O (2004) Exploiting query repetition and regularity in an adaptive community-based web search engine. User Model User-Adapt Interact 14(5):383–423

Yanbe Y, Jatowt A, Nakamura S, Tanaka K (2007) Can social bookmarking enhance search in the web? In: JCDL, Vancouver, pp 107–116

Yang L, Bao S, Lin Q, Wu X, Han D, Su Z, Yu Y (2011) Analyzing and predicting not-answered questions in community-based question answering services. In: AAAI, San Francisco

Zhang J, Ackerman MS, Adamic LA (2007) Expertise networks in online communities: structure and algorithms. In: WWW, Banff, pp 221–230

Social Weight

▸ User Behavior in Online Social Networks, Influencing Factors

Social Work

▸ Network Analysis in Helping Professions

Socioeconomic Stratification

▸ Demographic, Ethnic, and Socioeconomic Community Structure in Social Networks

Sociograms

▸ History and Evolution of Social Network Visualization

Socio-Graph Representations, Concepts, Data, and Analysis

Elie Raad[1] and Richard Chbeir[2]
[1]Faculty of Business, Memorial University of Newfoundland, St. John's, Canada
[2]Laboratoire LIUPPA, University of Pau and Adour Countries, Anglet, France

Synonyms

Centrality measures; Graph representation; Online social networks' concepts; Social network data; Social network representation

Glossary

A Network A structure that consists of a set of actors

A Graph Usually used to represent networks and consists of nodes to represent actors and edges to represent relationship

Social Network Data Data available on online social networks

Socio-graph Analysis Graph-based analysis of social networks, their concepts, and their data

Introduction

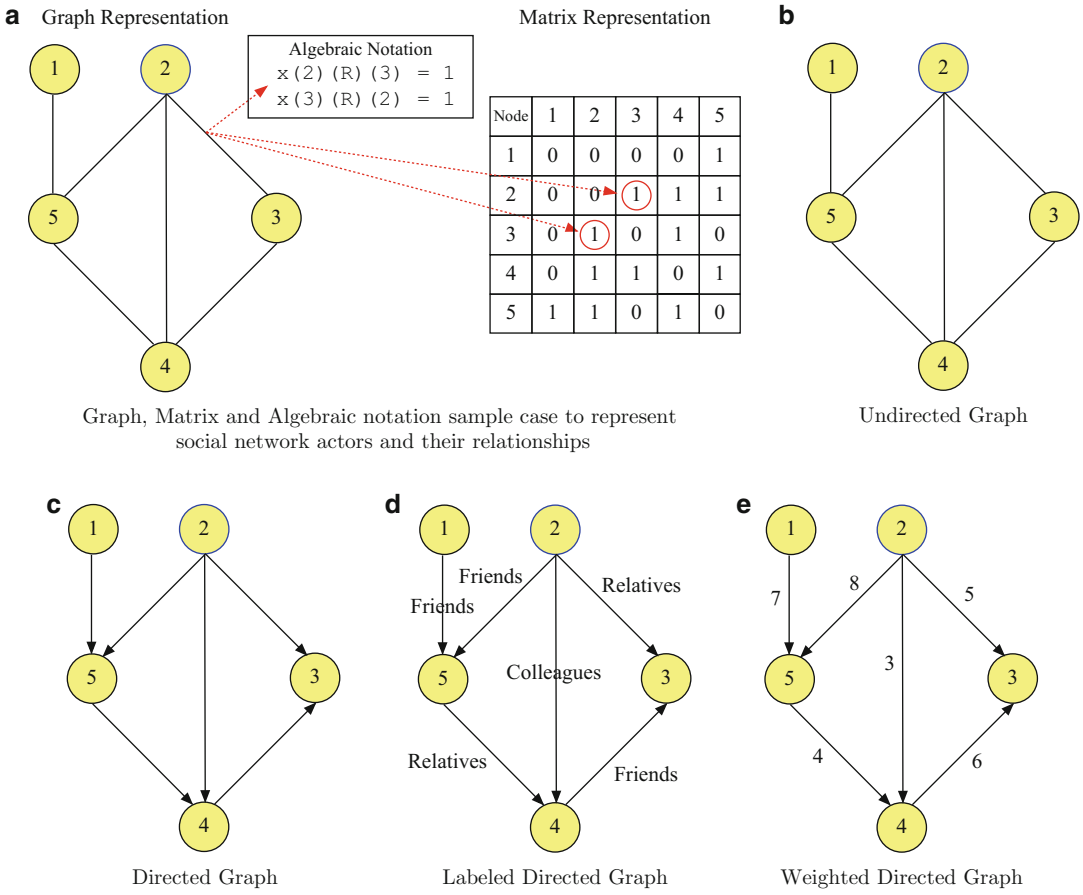
With the proliferation of online social networks, information sharing on these networks is gaining an ever-increasing importance. Obviously, online social networks have found ingenious ways to collect data as users socialize. Not surprisingly, when socializing social network users communicate, interact, and tend to freely reveal personal information in line with their perceptions and preferences. Understanding the characteristics of social networks is of considerable importance. Namely, the structure of the networks, the user-generated content, the level of interaction, as well as other dimensions can be used to analyze users' behaviors and understand their needs. In this work, we detail the most common representations of social networks, define their fundamental concepts, describe their social network data, and provide an overview of their most common analysis measures.

Representation of Social Networks

Finding an appropriate representation that can facilitate efficient and accurate interpretation of network data is an important step in social network studies. Just as graphs are a set of interconnected nodes, social networks are built on the foundation of actors interconnected through relationships. The use of graphs is a powerful visual tool and a formal means to represent social networks as detailed in this section.

Various Notations

There are many notations to represent social networks: algebraic notations, matrices, and graphs. A sample algebraic notation, a matrix representation, and a graph are illustrated in Fig. 1a. Depending on the data to be processed, the notation whose representation best fits the social network to describe is typically selected. But, there are well-known limits to the extent to which social networks can be formalized using matrices or algebraic notations to be recalled here. First, social networks hold valued relations and user-related attributes that algebraic notations cannot handle. Second, matrices are mostly efficient for small networks. Consequently, due to the large size of social networks, matrices are not the most appropriate way to represent these networks. Note that to represent a social network using matrices, a two-way matrix, also called sociomatrix, can be used. A sociomatrix consists of rows and columns that denote social actors and numbers or symbols in cells that denote existing relationships. Thus, graph-based representations are by far the most common form for modeling social networks (Wasserman and Faust 1994; Newman 2003; Boccaletti et al. 2006). Graphically representing social networks facilitate the understanding, labeling, and modeling of many properties of these networks (e.g., friendship networks with



Socio-Graph Representations, Concepts, Data, and Analysis, Fig. 1 A social network representation using a graph, its related matrix, and a sample algebraic notation (**a**), an undirected graph (**b**), a directed graph (**c**), a labeled graph (**d**), and a weighted graph (**e**) with $n = 5$ nodes and $m = 6$ links

labeled actors and relationships). Hence, graphs can represent various social data properties and their attributes while handling large real-world networks. Besides an adequate vocabulary to denote structural properties, graph-based representations have shown their mathematical reliability as well as their capacity to prove theorems for different social structural properties (Wasserman and Faust 1994). More details about the advantages and drawbacks of each representation are provided in Table 1.

Graph Representation

Graphs are usually used to represent networks in different fields such as biology, sociology, and computer science (Fortunato 2010). Graphs consist of nodes to represent actors and edges to represent relationships. The terms *nodes* and *objects* are usually used to denote *actors*. Likewise, *edges* may also be called *links* or *relationships*. Nodes with multiple edges are used to represent *ties* related to pairs of actors with more than one relationship.

More formally, a graph, $G = (V, E)$, consists of a set of nodes, V, and a set of edges, E. The number of elements in V and E is, respectively, denoted as $n = \|V\|$, the number of nodes, and $m = \|E\|$, the number of edges. The ith node, v_i, is usually referred to by its order i in the set V. Note that E consists of a finite set of relationships that is built from all relationships R_i, R_{i+1}, …, R_k, where k is the total number of relationships

Socio-Graph Representations, Concepts, Data, and Analysis, Table 1 Social network representations: advantages and drawbacks

Representation	Advantages	Drawbacks
Algebraic notations	– Useful for multi-relational networks as they can easily denote the combination of relations	– Cannot handle valued relations and user-related attributes
Matrices	– Efficient for small networks – Easy to denotes ties between a set of actors (a matrix for each relationship)	– Not a best choice for large social networks – Difficult to use when network data contain information on attributes
Graphs	– Handle large social networks – Provide a rich vocabulary to easily model social networks (labels, values, weights, etc.) – Provide mathematical operations that can be used to quantify structural properties and prove graph-based theorems	– Scalable visualization techniques are needed – Signed and valued graphs have to be used to represent valued relations

linking the pairs of actors. A subgraph $G' = (V', E')$ of $G = (V, E)$ is a graph such that $V' \subseteq V$ and $E' \subseteq E$. To represent different forms of data and to model the structural properties of social networks, graphs can have their edges and nodes labeled or unlabeled, directed or undirected, and weighted or unweighted as explained in what follows.

Directed and Undirected Graphs

In an undirected graph, the order of the connected vertices of an edge is not important. We refer to each link by a couple of nodes i and j such as $e(i, j)$ or e_{ij}, i and j are the end nodes of the link. A directed graph is defined by a set of nodes and a set of directed edges. The order of the two nodes is important: e_{ij} denotes the link from i to j, and $e_{ij} \neq e_{ji}$. To graphically indicate the direction of the links, directed edges are depicted by arrows. Depending on the nature of the relationship (asymmetric or symmetric), social network graphs can be undirected or directed. In fact, social networks can be modeled as undirected graphs when relationships between actors are mutual (e.g., symmetric relationships on Facebook (http://www.facebook.com) where e_{ij} and e_{ji} both denote a *friendship* link between user i and user j). Social networks can also be modeled as directed graphs when relationships are not bidirectional (e.g., asymmetric relationships on Twitter (http://www.twitter.com) where e_{ij} stands for user i is *following* user j). Figure 1b, c show, respectively, a representation of an undirected and a directed graph, both with $n = 5$ and $m = 6$. Directed links are important to evaluate the role of actors in a social network. They are key factors in measuring the centrality of actors in a social network. An interesting research work conducted by Brams et al. (2006) described how to transform undirected graphs to directed ones in order to explore additional information about the networks' structure. This transformation is an important step in understanding the flow of influence in the context of terrorist networks. In another study, Morselli et al. (2007) investigated and compared the structure of criminal and terrorist networks. The authors used links to compute a number of measures such as degree, betweenness, and centrality measures. These measures are used in order to discover the organizational hierarchy and to identify central and powerful criminal and terrorist actors.

Labeled and Unlabeled Graphs

Labels are important since they can identify the type of relationships between social network actors. When graphs are labeled, this means that a label is used to indicate the type of link that characterizes the relationship between the connected labeled nodes. Note that labeled graphs are considered to be signed graphs whenever their edges are labeled with either a $+$ or a $-$. For example, a signed graph can be used to model the inferred trust or distrust relationships in online social networks (Bachi et al. 2012). Figure 1d shows a labeled graph where the relationship type between linked actors is indicated. On social networks, relationship can be used to organize contacts based on their relationship types. This is useful in different situations such as improving face clustering and annotation of personal photo collections (Zhang et al. 2011), organizing friends into social circles (Raad et al. 2013; McAuley and Leskovec 2012), and enforcing access control (Carminati et al. 2009). Relationship-based access control is highly interesting in order to enable users to manage and fine-tune their privacy settings.

Weighted and Unweighted Graphs

Weights represent the strength of relationships between social network actors. When graphs are weighted, this means that their edges are assigned with a numerical weight, w, that can provide various indications such as link capacity, link strength, level of interaction, or similarity between the connected nodes (e.g., the number of messages that actors have exchanged, the number of common friends). Figure 1e shows a weighted graph (on a scale of 0–10) where the numeric values are assigned to the links and indicate the level of interaction between social network's actors. One way to characterize relationships is by computing their strength. On social networks, link strength is highly correlated with the level of interaction between users. Link strength can be used to model different levels of friendship where high weights represent "close friends" and low weights represent "acquaintances." Xiang et al. (2010) estimated the link strength from interaction activities (e.g., communication, tagging) and user similarities. More recently, another research explored a more specific aspect related to the predictive capacity of link strength to generalize from one social network to another (Gilbert 2012). Typically, link strength is primarily used to build intelligent systems that can favor interactions with strong ties without missing interesting activities derived by weak ties. Specifically, this interesting study showed that the link strength model captured in one social network can be generalized to another network, one in which it did not train.

Social Networks' Concepts

Networks have been used to model many systems of interest such as the World Wide Web, computer networks, biochemical networks, diffusion networks, and social networks. Each of these networks is a structure that consists of a set of actors representing, for instance, web pages on the World Wide Web or persons in a social network, connected together by relations, representing links between web pages or friendships between persons. Besides these structural properties (actors and relations), Wasserman and Faust (1994) identified a number of fundamental concepts like ties, dyads, triads, subgroups, and groups that characterize networks. In the following, we detail the concepts of actors, relations, and ties, the building blocks of social networks, before illustrating their use in online social networks.

General Concepts

The following defined concepts (actors, relations, and ties) are particularly important to understand and to study social networks.

Actor

An actor is a social entity that interacts with other entities not only to maintain existing relations but also to establish new ones. On social networks, the concept of actors can refer to various types of entities such as persons, groups, and organizations. Actors interact with each other through a variety of meaningful relations that denote different patterns of communication. Relations

like friendship, collaboration, and alliance can vary across time, applications, or in terms of the involved actors (Wilson et al. 2012). Consequently, there are two main categories of networks that can be identified based on the type of actors, one-mode networks and two-mode networks. While one-mode networks have a single type of actors, two-mode networks, also called bipartite, are networks with two types of actors. For instance, social networks modeling friendship between actors are an example of one-mode networks, whereas those concerned with group memberships or attendance at events are two-mode networks.

Relation

A relation represents a connection from one actor to another one. A relation, also called relationship, plays an important role when studying the structure of social networks and the interactions among their actors. A relationship is characterized by various features such as its content, direction, and strength. The relationship types have been addressed in several studies. Borgatti et al. (2009) distinguished between four basic types of relationships: similarities, social relations, interactions, and flows. For instance, these relationships can express memberships (e.g., same club), kinships (e.g., mother of), affections (e.g., likes), interactions (e.g., talked to), and flows (e.g., flow of information), among others. Relationships on social networks can be directed or undirected. Depending on their content, relationships may (or may not) have a specific direction. While relationships such as "marriage" and "friendship" are undirected, other relationships such as "parent of" or "fan of" are directed. Social network relationships can also differ in strength. Usually, the strength can be estimated in a variety of ways using information about the actors, their interaction activities, or the correlation between them as the most common indicators (Wilson et al. 2012; Gilbert 2012).

Tie

A tie is the set of all relationships that exist between two actors. It is tightly connected to the concept of relationship as it aggregates the different types of relationships that exist between two actors. Just like relationships, ties also vary in terms of their content, direction, and strength. Actors can be connected either with one relationship exclusively (e.g., employees of the same company) or with many relationships (e.g., employees of the same company and members of a sport club at the same time). Consequently, pairs of actors who maintain more than a single relationship are said to have a tie (Haythornthwaite 1996; Musial and Kazienko 2013). While each individual relationship within a tie carries its own content and direction, the strength of a tie depends on many factors such as the number of relationships that actors maintain, the reciprocity of these relationships, and their duration. Granovetter (1973) distinguished between strong and weak ties on the basis of the time actors spend together, their intimacy, and the emotional intensity of the existing relationships. Generally, weak ties are infrequently maintained with little interactions among actors (e.g., between distant acquaintances). Strong ties link similar actors, such as close friends, whose social circles tightly overlap with each other. Often, actors with strong ties that maintain many kinds of relations tend to communicate frequently with each other and use different channels of communication.

Online Social Networks' Concepts

Social networks and content-sharing sites with social networking functionalities have become an important part of the online activities on the web and one of the most influencing media. Facebook, Twitter, LinkedIn (http://www.linkedin.com), Google+ (http://plus.google.com), MySpace (http://www.myspace.com), Flickr (http://www.flickr.com), and YouTube (http://www.youtube.com) are among the most popular online social networks. These networks are attracting an ever-increasing number of users, many of whom are interested in establishing new connections, maintaining existing relations, and using the various social networks' services. The impact of social-based technologies on users, and particularly the influence of online social networks, is becoming the major source of contemporary fascination and controversy

(Musial and Kazienko 2013; Heidemann et al. 2012). A number of studies shed the light on different research directions like the implications of online social networks on individual connectivity (Hua and Wellman 2010), the capacity of technology to override cognitive limits in order to socialize with larger groups (Dunbar 2012), and the challenge to maintain a balance between security, privacy, usability, and sociability on online social networks (Zhang et al. 2010; Zheleva et al. 2012).

Social Network User

While many definitions exist for the term social network user (Adamic and Adar 2005; Boyd and Ellison 2007; Schneider et al. 2009), all of them are centered around social network users. First, these users create a personal profile which usually contains identifying information (e.g., name, age, photos) and captures users' interests (e.g., joining groups, liking brands). Afterwards, users start to socialize by interacting with other network members using a wide variety of communication tools offered by different social networks. In reality, each social network offers particular services and functionalities to target a well-defined community in the real world. Many of these available services are designed to help foster information sharing, bridge online and off-line connections to enforce interactions, provide instant information help, and enable users to derive a variety of uses and gratifications from these sites. To make use of the provided functionalities and to stay tuned with their related members, users create several accounts on various social networks where they disclose personal information with varying degrees of sensitivity (Raad et al. 2010). Personal information available on these networks commonly describes users and their interactions, along with their published data.

User Profile

Information about each social network user is maintained in a user profile which contains a number of attributes related to the demographics of users, their personal and professional addresses, their interests and preferences, as well as different types of user-generated contents (e.g., posts, photos, videos) (Thelwall 2008). Prior studies have noted the importance of user profiles to shape users' personalities, identities, and behaviors on social networks (Ryan and Xenos 2011; Gentile et al. 2012). These studies showed that among the disclosed attributes such as personal information and user-generated contents, photos and status updates have higher preferences for users.

Social Relationship

While myriad social networks' services assist users to find new contacts and establish new connections (e.g., friend suggestion systems through locations Cranshaw et al. 2010, based on interactions Wilson et al. 2012), users get connected to different types of contacts such as friends, relatives, colleagues, and strangers. Nevertheless, social relationship types between users and their contacts are rarely identified neither by the users nor by the existing social network sites (Raad et al. 2013; Tang and Liu 2009; McAuley and Leskovec 2012). This diversity, yet the different levels of social closeness between users and their contacts, entails an increasing need to analyze social interactions for better relationship (and consequently privacy) management. Currently, users are often provided with an exclusive and default relationship type connecting them to each of their contacts within a single social network site. However, it is common that social network users initiate connections with other contacts without any prior off-line connection (Ellison et al. 2011). On Facebook, for instance, these contacts are known as *friends* even though social network users do not particularly know or trust them. Consequently, many privacy-related concerns are raised in terms of identity disclosure, information sharing, access control, etc. (Zhang et al. 2010). The default social relationship(s) among the users of a number of famous social networks, along with other information, can be found in Table 2.

Social Network Data

Besides the fact that social networks are made of several components and can have various

Socio-Graph Representations, Concepts, Data, and Analysis, Table 2 Famous social networks with their main focus, default relationship(s), and the relationship's direction

Social network	Focus	Default relationship(s)	Relationship direction
Facebook	General use	Friendship	Symmetrical
Flickr	Photo sharing	Contact and optionally friend or family	Symmetrical
Google+	General use	Friends, family, acquaintances, and following	Symmetrical
LinkedIn	Professional	Business	Symmetrical
MySpace	General use	Friendship	Symmetrical
Twitter	Microblogging	Follower followee	Asymmetrical
YouTube	Video sharing	Subscribed to	Asymmetrical

representations, online social networks can also hold different types of data as detailed in the following. There are many types of social network data that can be collected from various sources on the web (i.e., different social network sites) and extracted from the daily activities and interactions between users. In this context, Schneier (2010) proposed a taxonomy of social data that we further develop into two main categories:

1. **Explicit data** is the set of explicit information that is provided by social network users or the data that is embedded in the provided information, i.e., metadata embedded in photos. Explicit information may include different forms of data such as text messages, photos, or videos. In this category, social network users actively participate in the creation of information.
   (a) **Service data** is the set of data that a user provides to the social network to create her account such as the user's name, date of birth, and country.
   (b) **Disclosed data** is what the user posts on her social network profile. This might include comments, posted photos, posted entries, captions, and shared links.
   (c) **Entrusted data** is what the user posts on other users' profiles. This might include comments, captions, and shared links.
   (d) **Incidental data** is what other social network users post about the user. It might include posted photos, comments, and notes.
2. **Implicit data** is the set of information that is not explicitly provided by social network users. However, social networks or third parties can use the set of explicit data to infer more information about the user. Inferring implicit data is founded on the analysis of the users' behaviors or derived from one or more user-provided information. For instance, it is possible to predict the characteristics of relationships between a number of users by examining the different aspects related to the patterns of communication between users (e.g., text messages, published photos, number of common friends) (Diesner et al. 2005; Raad et al. 2013). Consequently, in this category social network users are considered to be passive since the inferred information is extracted from prior activities or previously posted data.
   (a) **Behavioral data** is the data inferred from the user's behaviors. Social networks can collect information about the user's habits by tracking the patterns of activities of the user and consequently analyzing the user's behavior. Inferred behavioral data can reveal various information such as what the user usually do on the social networks, with whom the user usually interacts, and in what news topics the user is interested. Social networks collect such information by analyzing the articles that the user reads, the posts that the user publishes, the game that the user plays on social networks, etc.
   (b) **Derived data** is the data about the user that can be inferred from all other data. It is not related to the habit of the user. For example, the IP address can be used to infer the users' actual location. The derived data can also be inferred from the combination of two (or more) information. For example, if a significant number of

contacts live in one city, one can say that the social network user might live there as well. In this case, social networks or third parties must have access to two information in order to infer the derived data (the contacts of a user as the first information and their corresponding hometown as the second information).

Socio-graph Analysis

Concerned with the structural analysis of social interactions, research in social network analysis developed new models to study the fundamental properties of diverse theoretical and real-world networks (Luke and Harris 2007). Social network analysis has been used in different application domains such as e-mail communication networks, learning networks, epidemiology networks, terrorist networks, and online social networks. These works tried to answer a handful of questions such as how highly an actor is connected within a network? Who are the most influential actors in a network? How central is an actor within a network?

To capture the importance of actors within a network, a number of measures have been proposed in the literature (Koschützki et al. 2005). A commonly accepted measure is the centrality measure. Centrality consists of giving an importance order to the actors of a graph by using their connectivity within the network. Several structure-based metrics have been proposed to compute the centrality of an actor within a network, such as degree, closeness, and betweenness centrality (Freeman 1978). In what follows, we explain each of these metrics in details. Table 3 summarizes the characteristics of these structure-based centrality measures. As shown in Fig. 2, different central actor(s) in a network can be identified using each of these structural measures (degree, closeness, and betweenness).

Socio-Graph Representations, Concepts, Data, and Analysis, Table 3 Main centrality measures and their characteristics

Centrality measure	Characteristic
Degree	Measures how much an actor is highly connected to other actors within a network
Closeness	Computes the length of paths from an actor to other actors in the network
Betweenness	Measures the extent to which an actor lies on the paths between other actors

Degree Centrality

Degree centrality measures how much an actor is highly connected to other actors within a network. Degree centrality is a local measure since its value is computed by considering the number of links of an actor to other actors directly adjacent to it. A high degree centrality denotes the importance of an actor and gives an indication about potentially influential actors in the network. With a high degree of centrality, actors in social networks serve as hubs and as major channels of information in a network. Degree centrality, C_D, of an actor, v_i, can be computed as follows (Freeman 1978):

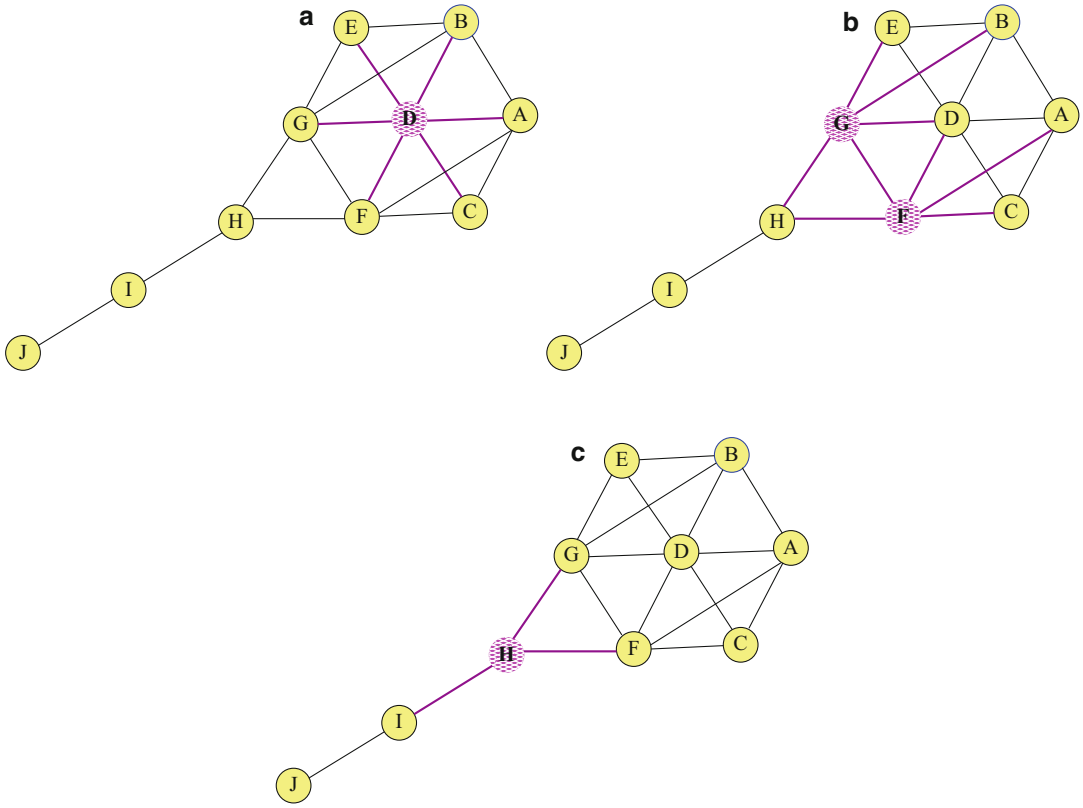
$$C_D(v_i) = \sum_{i=1}^{n} a(v_i, v_j) \tag{1}$$

where n is the total number of actors in the social network, $a(v_i, v_j) = 1$ if and only if v_i, and an actor, v_j, are connected by an edge; otherwise $a(v_i, v_j) = 0$.

Closeness Centrality

Closeness centrality computes the length of paths from an actor to other actors in the network. By measuring how close an actor is to all other actors, closeness centrality is also known as the median problem or the service facility location problem. Actors with small length path are considered more important in the network than those with high length path. Closeness centrality, C_C, of an actor, v_i, can be computed as follows (Freeman 1978):

$$C_C(v_i) = \frac{n-1}{\sum_{i=1}^{n} d(v_i, v_j)} \tag{2}$$



Socio-Graph Representations, Concepts, Data, and Analysis, Fig. 2 A network shaped as a kite graph where each centrality measure yields a different central actor: degree centrality (D), closeness centrality (F and G), and betweenness centrality (H). (**a**) Centrality degree. (**b**) Closeness degree. (**c**) Betweenness degree

where n is the total number of actors in the social network and $d(v_i, v_j)$ is the geodesic distance from actor v_i to another actor v_j.

Betweenness Centrality

Betweenness centrality measures the extent to which an actor lies on the paths between other actors. It denotes the number of times an actor needs to pass via a given actor to reach another one and thus represents the probability that an actor is involved into any communication between two other actors. Actors with high betweenness centrality facilitate the flow of information as they form critical bridges between other actors or groups of actors. Such central actors control the spread of information between groups of nonadjacent actors. Betweenness centrality, C_B, of an actor, v_i, can be computed as follows (Freeman 1978):

$$C_B(v_i) = \sum_{j} \sum_{<k} \frac{g_{jk}(n_i)}{g_{jk}} \qquad i \neq j \neq k \tag{3}$$

where n is the total number of actors in the social network, $C_B(v_i)$ is the betweenness centrality for actor v_i, and g_{jk} is the number of geodesics linking actors v_j and v_k that also pass through actor v_i.

To sum up, structural characteristics of a graph are a key aspect for social networks as they can be used to analyze the activity and to understand the behaviors of social network users. In most cases, networks of interconnected users are mainly represented by graphs, while graphs resulting from users' activity are usually referred to as the activity graphs. The activity captured within social networks is between users (the nodes) sharing various social data, connected with directed or

undirected relationships (the links), and having different levels of interactions (strong and weak ties). In this regard, these characteristics can be used to identify well-connected, central, and influential users. This would give more visibility and understanding for the network analyzer, but at the same time this can possibly reveal additional and sensitive information about the users, thus raising privacy concerns.

Cross-References

- ▸ Centrality Measures
- ▸ Classical Algorithms for Social Network Analysis: Future and Current Trends
- ▸ Graph Classification in Heterogeneous Networks
- ▸ Network Data Collected via the Web
- ▸ Social Networking Sites

References

Adamic L, Adar E (2005) How to search a social network. Soc Netw 27(3):187–203

Bachi G, Coscia M, Monreale A, Giannotti F (2012) Classifying trust/distrust relationships in online social networks. In: International conference on privacy, security, risk and trust (PASSAT), and social computing (SocialCom), Amsterdam, pp 552–557

Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U (2006) Complex networks: structure and dynamics. Phys Rep 424(4–5):175–308

Borgatti S, Mehra A, Brass D, Labianca G (2009) Network analysis in the social sciences. Science 323(5916):892–895

Boyd D, Ellison N (2007) Social network sites: definition, history, and scholarship. J Comput Mediat Commun 13(1):210–230

Brams S, Mutlu H, Ramirez S (2006) Influence in terrorist networks: from undirected to directed graphs. Stud Confl Terror 29(7):703–718

Carminati B, Ferrari E, Perego A (2009) Enforcing access control in web-based social networks. ACM Trans Inf Syst Secur 6(38):1–6

Cranshaw J, Toch E, Hong J, Kittur A, Sadeh N (2010) Bridging the gap between physical location and online social networks. In: UbiComp'10 – proceedings of the 2010 ACM conference on ubiquitous computing, Copenhagen, pp 119–128

Diesner J, Frantz T, Carley K (2005) Communication networks from the enron e-mail corpus "it's always about the people. enron is no different." Comput Math Organ Theory 11(3):201–228

Dunbar RIM (2012) Social cognition on the internet: testing constraints on social network size. Philos Trans R Soc B Biolog Sci 367(1599):2192–2201

Ellison N, Steinfield C, Lampe C (2011) Connection strategies: social capital implications of Facebook-enabled communication practices. New Media Soci 13(6):873–892

Fortunato S (2010) Community detection in graphs. Phys Rep 486(3–5):75–174

Freeman L (1978) Centrality in social networks conceptual clarification. Soc Netw 1(3):215–239

Gentile B, Twenge J, Freeman E, Campbell W (2012) The effect of social networking websites on positive self-views: an experimental investigation. Comput Hum Behav 28(5):1929–1933

Gilbert E (2012) Predicting tie strength in a new medium. In: Proceedings of the ACM 2012 conference on computer supported cooperative work, CSCW'12, Seattle. ACM, pp 1047–1056

Granovetter MS (1973) The strength of weak ties. Am J Sociol 78(6):1360–1380

Haythornthwaite C (1996) Social network analysis: an approach and technique for the study of information exchange. Libr Inf Sci Res 18(4):323–342

Heidemann J, Klier M, Probst F (2012) Online social networks: a survey of a global phenomenon. Comput Netw 56(18):3866–3878

Hua W, Wellman B (2010) Social connectivity in america: changes in adult friendship network size from 2002 to 2007. Am Behav Sci 53(8):1148–1169

Koschützki D, Lehmann K, Peeters L, Richter S, Tenfelde-Podehl D, Zlotowski O (2005) Centrality indices. In: Brandes U, Erlebach T (eds) Network analysis. Volume 3418 of lecture notes in computer science. Springer, Berlin/Heidelberg, pp 16–61

Luke D, Harris J (2007) Network analysis in public health: history, methods, and applications. Ann Rev Public Health 28:69–93

McAuley J, Leskovec J (2012) Learning to discover social circles in ego networks. Adv Neural Inf Process Syst 25:548–556

Morselli C, Giguère C, Petit K (2007) The efficiency/security trade-off in criminal networks. Soc Netw 29(1):143–153

Musial K, Kazienko P (2013) Social networks on the internet. World Wide Web 16(1):31–72

Newman M (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256

Raad E, Chbeir R, Dipanda A (2010) User profile matching in social networks. In: Proceedings of the 13th international conference on network-based information systems, NBiS 2010, Takayama, pp 297–304

Raad E, Chbeir R, Dipanda A (2013) Discovering relationship types between users using profiles and shared photos in a social network. Multimed Tools Appl 64(1):141–170

Ryan T, Xenos S (2011) Who uses Facebook? An investigation into the relationship between the big five, shyness, narcissism, loneliness, and Facebook usage. Comput Hum Behav 27(5):1658–1664
Schneier B (2010) A taxonomy of social networking data. IEEE Secur Priv 8(4):88
Schneider F, Feldmann A, Krishnamurthy B, Willinger W (2009) Understanding online social network usage from a network perspective. In: Proceedings of the 9th ACM SIGCOMM conference on internet measurement conference, IMC'09, Chicago. ACM, pp 35–48
Tang L, Liu H (2009) Scalable learning of collective behavior based on sparse social dimensions. In: Proceedings of the 18th ACM conference on information and knowledge management, CIKM'09, Hong Kong. ACM, pp 1107–1116
Thelwall M (2008) Social networks, gender, and friending: an analysis of mySpace member profiles. J Am Soc Inf Sci Technol 59(8):1321–1330
Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge
Wilson C, Sala A, Puttaswamy KPN, Zhao BY (2012) Beyond social graphs: user interactions in online social networks and their implications. ACM Trans Web 6(4):17:1–17:31
Xiang R, Neville J, Rogati M (2010) Modeling relationship strength in online social networks. In: Proceedings of the 19th international conference on World Wide Web, WWW'10, Raleigh. ACM, pp 981–990
Zhang C, Sun J, Zhu X, Fang Y (2010) Privacy and security for online social networks: challenges and opportunities. IEEE Netw 24(4):13–18
Zhang T, Chao H, Tretter D (2011) Dynamic estimation of family relations from photos. In: Lee K-T, Tsai W-H, Liao H-Y, Chen T, Hsieh J-W, Tseng C-C (eds) Advances in multimedia modeling. Volume 6524 of lecture notes in computer science. Springer, Berlin/Heidelberg, pp 65–76
Zheleva E, Terzi E, Getoor L (2012) Privacy in social networks. Synth Lect Data Min Knowl Discov 3(1):1–85

Sociology of the Web

▸ Web Science

Sociomatrix

▸ History and Evolution of Social Network Visualization

Sociometric Diagram

▸ History and Evolution of Social Network Visualization

Sociopsychological Theories

▸ Friends Recommendations in Dynamic Social Networks

Socio-Technical Systems

▸ Futures of Social Networks: Where Are Trends Heading?

Software

▸ NetMiner

Sources of Network Data

Monika Cerinšek[1] and Vladimir Batagelj[2]
[1]Hruška d.o.o., Ljubljana, Slovenia
[2]Department of Mathematics, University of Ljubljana, FMF, Ljubljana, Slovenia

Synonyms

Almost network data; Archives; Boundary problem; Copyrights; Databases; Ego-centered networks; Ethics; Networks; Observation; Random networks; Semantic web; Surveys

Glossary

Network Analysis A study of networks as a representation of relations between discrete objects

Social Network A social structure based on a set of actors (individuals or organizations) and the ties between these actors

Genealogy A study of families and tracing of their lineages

Web Crawler An Internet bot that automatically browses the World Wide Web

Computer-Assisted Text Analysis – CATA Techniques that model and structure the information content of textual sources

Cloud Technology A use of hardware and software that are delivered as a service over a network (usually the Internet)

Introduction

We can find the network data almost everywhere in our lives:

- The cities are linked with roads.
- People in a group are linked by exchange of messages (mail, phone).
- Works from a field of research are linked with citations.
- Researchers are linked with their collaborations.
- Atoms in molecules are linked with their chemical bonds.
- Words are linked according to their coappearance in sentences of some text.
- In genealogies people are linked by marriage and parent-child ties.

A ***graph*** $\mathcal{G}$ is an ordered pair of sets $(\mathcal{V}, \mathcal{L})$ with the set of ***nodes*** $\mathcal{V}$ and the set of ***links*** $\mathcal{L}$. Every link has two end-nodes. It is either directed, an ***arc***, or undirected, an ***edge***. A ***network*** $\mathcal{N} = (\mathcal{V}, \mathcal{L}, \mathcal{P}, \mathcal{W})$ consists of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{L})$, describing the structure of network, and additional data: ***properties*** $\mathcal{P}$ of nodes and ***weights*** $\mathcal{W}$ on links. There are different types of networks beside ordinary networks.

A ***two-mode*** network is a network $\mathcal{N} = ((\mathcal{I}, \mathcal{J}), \mathcal{L}, \mathcal{P}, \mathcal{W})$, where the set of nodes $\mathcal{V} = \mathcal{I} \cup \mathcal{J}$ is split into two disjoint sets of nodes $\mathcal{I}$ and $\mathcal{J}$ and each link from $\mathcal{L}$ has one end-node in $\mathcal{I}$ and the other end-node in $\mathcal{J}$.

A ***multirelational*** network $\mathcal{N} = (\mathcal{V}, \mathcal{L}, \mathcal{P}, \mathcal{W})$ allows multiple relations to exist in the network $\mathcal{L} = (\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_r)$.

In a ***temporal*** network $\mathcal{N} = (\mathcal{V}, \mathcal{L}, \mathcal{P}, \mathcal{W}, T)$ the time T is attached to a network. For all nodes and links we have to specify the time intervals in which the element is active (present) in the network. Also properties and weights can change through time.

When constructing a network we must first specify what are the nodes and which relation is linking them – the ***network boundary*** problem (Wasserman and Faust 1994; Marsden 2011). According to the plan of network analyses, we need to bound the set of nodes to those that we need. Along with nodes and links, we select also their properties. We have to decide whether the network is one-mode or two-mode and which node properties are important for our intended analyses. About the links we have to answer to several questions: Are the links directed? Are there different types of links (relations)? Can a pair of nodes be linked with multiple links? What are the weights on the links? Is the network static, or is it changing through time?

Sometimes the list of nodes is known in advance (e.g., students in the class). But often the set of nodes is constructed during the network data collection process. In this case we have to specify the membership criteria determining for each potential node whether it belongs to the network or not.

For collecting the network data, the ***snowball*** procedure is often used. We first choose a (small) set of nodes as initial candidates. Then we collect the data about each candidate and determine its neighboring nodes. The new ones among them we add to the list of candidates. The inclusion of the new nodes can be ruled also by some other criteria, for example, by the distance from the closest initial node. We end this process when the list of candidates is exhausted or the limit to the number of inspected nodes is reached.

Another problem that often occurs when defining the set of nodes is the ***identification*** of nodes. The unit corresponding to a node can have different names (***synonymy***), or

the same name can denote different units (***homonymy*** or ***ambiguity***). For example, in a bibliography on mathematics from Zentralblatt MATH, the names Borštnik, N. S. Mankoč; Mankoč Borštnik, N.; Mankoč-Borštnik, Norma; Mankoč Borštnik, Norma Susana; Mankoc-Borstnik, N.S.; and Mankoč Borštnik, N.S. belong to the same author. On the other hand, in Zentralblatt MATH at least two different Smith, John W. are recorded, because publications of the author(s) with this name spanned from 1868 to 2007. There are at least 57 different mathematicians with the name Wang, Li in the MathSciNet Database. Its editors are trying hard, from the year 1985, to resolve the authors identification problem (Martin et al. 2013) during the data entry phase. In the future the problem could be eliminated by general adoption of initiatives such as ResearcherID or ORCID.

The identification problem appears also when the units are extracted from the plain text, for example, "the President of the USA" and "Barack Obama." To resolve it we have to provide lists of equivalent terms. Another source of identification problems is the grammar rules of the language used in text. For example, the action "go" can appear in the text in different forms "go," "goes," "gone," "going," and "went." To resolve these problems we apply the stemming or lemmatization procedures from natural language processing toolkits such as NLTK or MontyLingua.

A special approach of collecting data for a network analysis is by forming ***ego-centered*** networks (Lozar Manfreda et al. 2004). This approach is used when the population of our interest is too large. From the population we select a sample of units (***egos***) and collect the data about them and their neighbors (***alters***) and links among them. An example is the friendship networks of selected persons from Facebook.

Collecting the network data we have to respect legal (copyright) and ethical constraints (Borgatti and Molina 2003; Eynon et al. 2008; Charlesworth 2008; Breiger 2005).

The network data can be obtained in many ways:

- By observation
- With surveys or interviews
- From archives and databases
- From data organized in a network form
- Derived from the data
- From semantic web
- With generating random networks

Each of the above methods for gathering the network data is described in more details in the following subsections. For details additional references are provided.

Observation

To form a network we must first obtain the data. The ways of obtaining the data have been changing through history following the development of the technology. A basic approach is the observation (Mitchell 1969). The observation is a human activity consisting of receiving information about the outside world through the senses, or the recording of data using scientific instruments and includes also any data collected during this activity. Scientific instruments were developed to amplify human powers of observation, such as weighing scales, clocks, telescopes, microscopes, thermometers, cameras, and tape recorders and also to translate into perceptible form events that are unobservable by human senses, such as voltmeters, spectrometers, infrared cameras, oscilloscopes, interferometers, Geiger counters, x-ray machines, and radio receivers (Shipman et al. 2009).

Making direct measurements is the most accurate method for many variables but can be limited by the technology available. The main alternative to direct observation is to require others to report their activities.

An example of the observational network data collection is described in the PhD thesis of Sampson (1968). He did an ethnographic study of community structure in a New England monastery – he divided 18 novices into 4 groups at 5 time points based on his observations and analyses. Another example is the detection of molecular structure of organic molecules.

Surveys

Survey as a method is a data-gathering method that actively includes the observeds (Marsden 1990). They allow us to study attitudes, beliefs, behaviors, and other characteristics. With carefully prepared questionnaires one can collect vast amounts of quality data. A ***questionnaire*** is a list of questions. Answers can be ***closed*** – selected from a given list. They are easier to analyze. But the ***open*** answers, that are not given in advance, allow the analysts to get a wider amount of information. A survey can take different forms: face-to-face, paper and pencil, telephone, e-mail, or online. Nowadays questionnaires are mostly digital (online surveys) that allows them to be adaptable, immediate checking of the entered data, and also collecting some contextual (observational) data.

The use of direct observation in combination with surveys can provide additional information. It can confirm or negate information gained from surveys. As observation itself also the observation in combination with surveys must be prepared. The observant might use appropriate scales, checklists, and other observation materials that are chosen in accordance with the questions and possible closed answers in the survey.

An interesting network obtained by interviewing is the Edinburgh Associative Thesaurus.

Surveys are the most commonly used methods to gather social network data. They are also used to study interorganizational relations (Mizruchi and Galaskiewicz 1993). For details on surveys and questionnaires, see the essay "▸ Questionnaires for Measuring Social Network Contacts."

Archives and Databases

An archive is a collection of historical data, or the physical place where they are located (Schmidt 2011). Archives have a historical, cultural, and evidentiary value. Archives exist everywhere, where data has been stored. Every organization has an archive of past activities; universities have archives of past students' achievements and research; backup on the personal computer is an archive of past usage of the computer, etc. With the transition of office work to computers and the spread of Internet, many archives became digital. A database is an organized collection of data, mostly in digital form. Database is organized in records and for each record it has stored some properties (Ullman and Widom 2008). Because data is organized, it is very easy to transform it in a collection of (often two-mode) networks which are then used in the network analysis. Smaller amounts of data can be presented in a tabular form as spreadsheets.

For example, there exist many bibliographic databases (Web of Science, Scopus, Zentralblatt MATH, etc.) that are keeping data about published papers and books. Even the World Wide Web is being partially collected and preserved as an archive for future researchers, historians, and the public.

As a source of data, the archives of various kinds are inexpensive and advantageous for studying especially social networks in the past (Marsden 1990). The network data can be derived from archived data. For example, relations between corporations can be studied based on information about persons in the boards of directors of the corporations.

Historical archives help researchers to gain knowledge about the development of some field – economics, scholar, military, etc. For example, with data from World War II one can study the military movement through the war, the transfer of refugees or prisoners, the transfer of weapons, etc. Another example is the analysis of alliances between the most powerful countries over a selected time period.

Archived data about the inhabitants of a city or an area can be used for genealogical analysis. In genealogy we can search for typical marriage patterns and their irregularities. For example, marriages among relatives to keep the family's wealth, or on the other hand, marriages outside the family to increase its influence. The genealogical data are often available in the GEDCOM format. Large collection of family genealogies is available at the Genealogy Forum. For "scientific" genealogies

used in anthropological research, see the site KinSource.

Activities on the Internet, such as e-mail, chat, and forums, leave their traces that can be used as sources for network data. A notorious example is the Enron e-mail data.

Especially interesting for network analysis is the World Wide Web as an archive. The web crawlers visit the page with URL from the list of URLs, identify all hyperlinks in it, and add the URLs of these hyperlinks to the list. The largest web archiving organization based on crawling approach is the Internet Archive, but also national libraries, national archives, and other organizations are involved in archiving mostly culturally important web content.

Enormous archives are being formed by different social networking services such as Facebook, Twitter, LinkedIn, and Google+. These organizations are collecting the data about users, their posts, or tweets. Data about users are not publicly available. The user can download only the data about his/her past activity and the data that other users declared visible for him.

A large amount of data is stored in Internet Movie Database (IMDb) and services such as Amazon or lastFM. Converting data into multiple two-mode networks and combining them in network analysis allows us to obtain information about collaboration between actors, producers and composers, similarity of the movies according to different measures, etc.

With the development of technology, different types of databases occurred, where the type of the database is defined with the way the data is stored in a database. With growth of available data the data warehouses were developed. A data warehouses archive data directly from the source. It is a central source of data for use by managers for creating statistical dashboards and reports about it. The other very popular type of database is cloud database that relies on the cloud technology (Voorsluys et al. 2011).

A graph database (Angles and Gutierrez 2008) is also useful in the network analysis and it is interesting because of the way the data is stored in it. It uses graph structure to represent and store information. Specialized graph database uses a network model, which is conceived as a flexible way of representing objects and their relationships.

Everyday large amounts of data are being collected. So big data (White 2012) is considered to be a collection of large data sets. These data sets are so large and complex that it is very difficult to be processed using traditional data processing applications. Also suitable technologies are required such as cluster analysis, machine learning, neural networks, pattern recognition, and anomaly detection.

Many repositories of networks and datasets of other types are available: Repositories of Datasets, KDnuggets Datasets for Data Mining, Data Surfing on the World Wide Web, Public Data Sets on Amazon Web Services, TunedIT, the Internet2 Observatory Data Collections, Infochimps, CAIDA (the Cooperative Association for Internet Data Analysis) Data, and Network Data Sources on Pajek's web page.

Different activities are traced by their logs. Mobile network operators record the usage of the phones by their users, the data from weather stations is collected, online social network providers collect the data about their users (Abdesslem et al. 2012), different sensor networks are being established, peer-to-peer (P2P) networks are more and more interesting, using the radio-frequency identification (RFID) tags we can follow the movement of their owners, etc. Such data can be used for prediction or just for the behavioral analysis of the users.

Almost Network Data

Some data is already organized in a network form. A transportation network is a network of roads, pipes, streets, or any other similar structure that allows transportation of some kind. They are represented as links, and crossings are presented as nodes. Another area that deals with a lot of data in a network form is chemistry. The structure of every molecule is a network with atoms as nodes and covalent chemical bonds as links between them. The most interesting for network analysis are organic molecules as proteins, lipids,

hydrocarbons, and DNA. A lot of molecular data is available at Protein Data Bank.

To analyze such data using the selected network analysis tool, we usually have to transform them into the corresponding input network data format. These issues are elaborated in details in the essay "▸ Network Data File Formats."

Sometimes special programming solutions should be developed to perform the required transformation. For example, the transformation of the ESRI shape file describing the map of borders between the country's administrative units (states, counties) into the neighborhood relation of the administrative units can be done with a short program in R using the function `poly2nb` from the package `spdep`.

Networks Derived from Data

Some data sources require more sophisticated procedures to transform them into corresponding networks.

Very interesting data sources are also the daily news archives of the news agencies (Agence France-Presse, Reuters, United Press International, American Press Agency, Xinhua, ITAR-TASS, etc.). A single news is essentially a (tagged) plain text that can be analyzed with ***computer-assisted text analysis*** (Popping 2000). One of the main approaches to this type of text analysis is the semantic text analysis. The units of the text are encoded according to the Chomsky's ***subject-verb-object*** model which can be directly transformed into temporal multirelational networks with subjects and objects as nodes and verbs as relations. Examples of applications of this approach are the Kansas Event Data System, Paul Hensel's International Relations Data Site, or Correlates of War. An elaboration of this approach is given in the Franzosi's book "From Words to Numbers" (Franzosi 2004). See also the Centering Resonance Analysis approach proposed by Steve Corman.

Another example are the neighbors networks. Let $\mathcal{V}$ be a set of (multivariate) units and $d(u, v)$ a ***dissimilarity*** on it. They determine two types of networks: the ***k-nearest neighbors*** network:

$$\mathcal{N}(k) = (\mathcal{V}, \mathcal{AA}, w)$$

$$(u, v) \in \mathcal{AA} \Leftrightarrow v \text{ is among } k \text{ nearest neighbors of } u, w(u, v) = d(u, v)$$

and the ***r-neighbors*** network: $\mathcal{N}(r) = (\mathcal{V}, \mathcal{E}, w)$

$$(u : v) \in \mathcal{E} \Leftrightarrow d(u, v) \leq r,$$
$$w(u, v) = w(v, u) = d(u, v)$$

These networks provide a link between (multivariate) data analysis and network analysis. For larger sets of units a problem of an efficient algorithm for determining the nearest neighbors arise. David M. Mount wrote the Approximate Nearest Neighbor Library with fast algorithms for the (approximate) nearest neighbor search. In R these algorithms are available through the function `ann` in package `yaImpute`.

Semantic Web

Semantic web (Berners-Lee et al. 2001) is an upgrade and an extension of the ordinary web. It provides a data layer in the World Wide Web to be used by web services. The basis for semantic web is the semantic description of the web content with the use of metadata and ontologies. The aim is to convert web of unstructured documents into a web of data. This would make also easier to analyze this data, because it would be already in a network form.

Semantic web is based on Uniform Resource Identifier (URI), Resource Description Framework (RDF), and Web Ontology Language (OWL). The URI is a string used to identify a name or a resource and enables interaction with representations of the resource over a network using specific protocols. RDF is a W3C standard for encoding knowledge. It is used for conceptual description or modeling of information from web resources and by computers to seek the knowledge. RDF is actually a foundation for processing metadata; it provides interoperability between applications that exchange machine-understandable information on the Web.

The OWL is a family of knowledge representation languages for authoring ontologies.

A piece of knowledge in RDF is represented as a triple subject-predicate-object. A subject denotes the resource; the predicate denotes aspects of the resource and expresses a relationship between the subject and the object. The resources are always named by URIs plus optional anchor IDs (URL and URN are its subsets). The triples form a multirelational network with subjects and objects represented as nodes and predicates determining types of ties – relations. There are large collections of RDF triples: Linked Data – Connect Distributed Data across the Web, Freebase, and DBpedia.

Different syntax formats exist and are quite varying by their complexity: N3, N-Triples, TRiG, TRiX, Turtle, RDF/XML, RDFa, and JSON-LD. The purpose of RDF is to provide an encoding and interpretation mechanism so that resources can be described in a way that a compatible software can understand it. Some formats are not human friendly but more machine friendly. See also SPARQL – an RDF query language.

Generating Random Networks

Generation of random networks (Batagelj and Brandes 2005; van der Hofstad 2011) has become important for studies of complex systems such as electrical power grid, social relations, the World Wide Web and Internet, and collaboration and citation networks of scientists. Random networks are used for modeling classes of graphs.

Paul Erdős and Alfréd Rényi proposed in Erdős and Rényi (1959) an approach to formalize the notion of a random graph. The ***Erdős-Rényi*** model, denoted by $\mathcal{G}(n,m)$, where n is the number of nodes and m is the number of edges, generates a random graph on n nodes and m edges (uniformly) randomly selected among the $\frac{n(n-1)}{2}$ potential edges.

Another, closely related to Erdős-Rényi model, is the ***Gilbert's*** model $\mathcal{G}(n,p)$ (Gilbert 1959), where n is the number of nodes and p is the probability that an edge is included in the random graph. In this model the $\frac{n(n-1)}{2}$ potential edges of a simple undirected graph $G(n,p) \in \mathcal{G}(n,p)$ are included independently with the probability p.

A model called ***small worlds*** was introduced by Watts and Strogatz (1998). This class of random graphs depends on two structural features. The clustering coefficient is high and the average distance between pairs of nodes is short. Networks such as social networks, the Internet, and gene networks all exhibit small world network characteristics.

The degree distribution of random graph from Erdős-Rényi's or Gilbert's model is sharply concentrated around its average degree. In most real-life networks, it roughly follows the powerlaw. Such networks are called ***scale-free***. Barabási and Albert (1999) described a process of ***preferential attachment*** that generates graphs with this property. The preferential attachment process creates one node at a time and each newly created node is attached to a fixed number of already existing nodes. The probability of selecting a specific neighbor is proportional to its current degree.

Different classes of random graphs can be described also as ***probabilistic inductive classes*** of graphs (Kejžar et al. 2008).

Acknowledgments

This work was supported in part by the ARRS, Slovenia, and grant P1-0294, as well as by grant N1-0011 within the EUROCORES Programme EUROGIGA (project GReGAS) of the European Science Foundation.

The first author was financed in part by the European Union, European Social Fund.

Cross-References

▸ Collection and Analysis of Relational Data in Organizational and Market Settings
▸ Ethics of Social Networks and Mining

▸ Ethical Issues Surrounding Data Collection in Online Social Networks
▸ Network Data Collected via the Web
▸ Network Data File Formats
▸ Quality of Social Network Data
▸ Questionnaires for Measuring Social Network Contacts

References

Abdesslem FB, Parris I, Henderson T (2012) Reliable online social network data collection. Computational social networks. Springer, London
Angles R, Gutierrez C (2008) Survey of graph database models. ACM Comput Surv 40(1):1–39
Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512
Batagelj V, Brandes U (2005) Efficient generation of large random networks. Phys Rev E 71(3):036113
Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. Sci Am 284(5):28–37
Borgatti SP, Molina JL (1990) Ethical and strategic issues in organizational network analysis. J Appl Behav Sci 39(3):337–350
Breiger RL (2005) Ethical dilemmas in social network research: introduction to special issue. Soc Netw 27(2):88–93
Charlesworth A (2008) Understanding and managing legal issues in internet research. In: Fielding NG, Lee RM, Blank G (eds) The SAGE handbook of online research methods. SAGE, London
Eynon R, Fry J, Schroeder R (2008) The ethics of internet research. In: Fielding NG, Lee RM, Blank G (eds) The SAGE handbook of online research methods. SAGE, London
Erdős P, Rényi A (1959) On random graphs. Publ Math Debr 6:290–297
Franzosi R (2004) From words to numbers: narrative, data, and social science. Cambridge University Press, Cambridge
Gilbert EN (1959) Random graphs. Ann Math Stat 30:1141–1144
Kejžar N, Nikoloski Z, Batagelj V (2008) Probabilistic inductive classes of graphs. J Math Sociol 32(2): 85–109
Lozar Manfreda K, Vehovar V, Hlebec V (2004) Collecting ego-centred network data via the web. Metodološki zvezki 1(2):295–321
Marsden PV (1990) Network data and measurement. Ann Rev Sociol 16:435–463
Marsden PV (2011) Survey methods for network data. In: Scott J, Carrington PJ (eds) The SAGE handbook of social network analysis. SAGE, London
Martin T, Ball B, Karrer B, Newman MEJ (2013) Coauthorship and citation in scientific publishing. Arxiv: http://arxiv.org/abs/1304.0473v1. Accessed 2013-04-12
Mitchell JC (1969) The concept and use of social networks. In: Mitchell JC (ed) Social networks in urban situations. Manchester University Press, Manchester
Mizruchi MS, Galaskiewicz J (1993) Networks of interorganizational relations. Soc Methods Res 22(1):46–70
Popping R (2000) Computer-assisted text analysis. SAGE, London
Sampson SF (1968) A novitiate in a period of change. An experimental and case study of social relationships. PhD thesis, Cornell University
Schmidt L (2011) Using archives. A guide to effective research. Society of American Archivists, Wheaton
Shipman J, Wilson JD, Todd A (2009) Introduction to physical science, 12th edn. Cengage Learning, Boston
Ullman J, Widom J (2008) First course in database systems, 3rd edn. Prentice-Hall, Upper Saddle River
van der Hofstad R (2011) Random graphs and complex networks. http://www.win.tue.nl/~rhofstad/NotesRGCN.pdf. Accessed 2013-02-27
Voorsluys W, Broberg J, Buyya R (2011) Introduction to cloud computing. In: Buyya R, Broberg J, Goscinski A (eds) Cloud computing: principles and paradigms. Wiley, New York
Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge
Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–442
White T (2012) Hadoop: the definite guide, 3rd edn. O'Reilly Media, Sebastopol

Web References

Approximate Nearest Neighbor Library. http://www.cs.umd.edu/~mount/ANN
CAIDA (The Cooperative Association for Internet Data Analysis) Data. http://www.caida.org/data/
Centering Resonance Analysis approach proposed by Steve Corman. http://www.crawdadtech.com/
Correlates of War. http://www.correlatesofwar.org/
Data Surfing on the World Wide Web. http://it.stlawu.edu/~rlock/datasurf.html
DBpedia. http://en.wikipedia.org/wiki/DBpedia
Edinburgh Associative Thesaurus. http://www.eat.rl.ac.uk/
Enron E-mail Data. http://www.isi.edu/~adibi/Enron/Enron.htm
Freebase. http://www.freebase.com/
Genealogy Forum. http://www.genealogyforum.com/gedcom/
Infochimps. http://infochimps.com/
Internet Archive. http://archive.org/index.php
Internet Movie Database. http://www.imdb.com/
KDnuggets Datasets for Data Mining. http://www.kdnuggets.com/datasets/index.html

KinSource. http://kinsource.net/csac/wiki/kinsrc/KinSources/
Linked Data – Connect Distributed Data across the Web. http://linkeddata.org/
Network Data Sources on Pajek's web page. http://pajek.imfm.si/doku.php?id=data:urls:index
Paul Hensel's International Relations Data Site. http://www.paulhensel.org/data.html
Protein Data Bank. http://www.rcsb.org/pdb/home/home.do
Public Data Sets on Amazon Web Services. http://aws.amazon.com/publicdatasets/
Repositories of Datasets. http://www.trustlet.org/wiki/Repositories_of_datasets
The Internet2 Observatory Data Collections. http://www.internet2.edu/observatory/archive/data-collections.html
The Kansas Event Data System. http://web.ku.edu/keds/
TunedIT. http://tunedit.org/repo
Web Archiving Service. http://webarchives.cdlib.org/

Space-Embedded Networks

▸ Spatial Networks

Spam Detection, E-mail/Social Network

Cailing Dong and Bin Zhou
Department of Information Systems, University of Maryland, Baltimore County, Baltimore, MD, USA

Synonyms

Junk e-mail; Social spam; Unsolicited bulk e-mail

Glossary

Spam Unsolicited, unwanted message intended to be delivered to an indiscriminate target, directly or indirectly, notwithstanding measures to prevent its delivery
Spammer Originator of spam message
Spam Filter An automated tool that is built to detect spam message with the purpose of preventing its delivery
Whitelist A list of contacts whose e-mails should be delivered
Blacklist A list of contacts whose e-mails are deemed to be spam
Classifier A model that identifies which of a set of categories an object belongs to

Definition

Spam generally refers to "unsolicited, unwanted message intended to be delivered to an indiscriminate target, directly or indirectly, notwithstanding measures to prevent its delivery" (Cormack 2008). While e-mail spam is the mostly widely recognized form of spam, spam actually pervades many existing information systems and social media, including instant messaging (Paulson 2004), blogs (Abu-Nimeh and Chen 2010), newsgroups and forums (Shin et al. 2011), and online social media (Jin et al. 2011). Spam also exists in web search, where search engines are used as a delivery mechanism for web spam (Gyöngyi and Garcia-Molina 2005).

The overwhelming of spam messages in existing information systems and social media severely deteriorates the quality of communication. The objective of spam detection is to develop effective and efficient anti-spam techniques with the purpose of preventing the delivery of spam messages.

Introduction

Regardless of various forms of spam in reality, sending spam messages is essentially profit-driven activity. Spammers, the originators of spam message, intend to deliver the information to a large volume of recipients. Spam messages often contain advertising for commercial products, URL links to promoted websites which may serve as means of adult content dissemination and phishing attacks, or even computer malwares that

are specifically designed to hijack the recipient's computers. Although some forms of spam can be identified whenever the message is delivered and viewed by the receipts, spammers can still be profitable even only a very small fraction of recipients take responses to spam messages. Due to the fact that the operating cost of sending spam messages is substantially cheap and the barrier to entry is quite low, the volume of spam has been consistently increasing in the past several years. According to the 2009 report by Ferris Research (Jennings 2009), the worldwide financial losses caused by spam were estimated to be $130 billion in 2009, a 30 % increase over the 2007 estimates. Comparing to the estimated figures in 2005, the total losses were increased by 160 % in 2009.

Spam dramatically deteriorates the quality of communication. From the users' point of view, users are victims affected directly by spam. Not only users' material wealth but also their personal information could be under risk to spammers. From the system providers' point of view, they are forced to waste a significant amount of computational and storage resources for spam messages. Moreover, if a user receives many spam messages, his/her trust in the system can be drastically weakened, which inevitably makes the user switch from the current system provider to another competitor. Therefore, both users and system providers have strong incentive to wipe out spam thoroughly.

Historical Background

The term "spam" is named for Spam luncheon meat by way of a *Monty Python sketch* where Spam is depicted as ubiquitous and unavoidable. Among various forms of spam in the literature, e-mail spam is the most common one. Along with the vigorous development of the Internet since the mid-1990s, e-mail becomes a popular communication and information exchanging method. E-mail spam started to be a serious problem since then, and it grew exponentially over the following years. Nowadays, spam comprises the vast majority of e-mail messages sent daily. It is reported that 78 % of the e-mails are spam (Fletcher 2009). Due to such large impacts, e-mail spam becomes not only a technical challenge but also a legal crisis and political issue. Myriad technological and legal-based attempts have been developed to combat e-mail spam.

Meanwhile, the user-centered design feature of Web 2.0 further involves people in a rapid information sharing and propagation era, which brings the prosperity of many online social websites. Social websites are designed to support and foster various social interactions, which on the other hand heavily rely on users for content contribution and distribution. However, such interactive and dynamic features also provide fertile soil for spam. Spammers in social websites disguise their spam messages as links, content, video, audio, and executable files. Unlike traditional e-mail spam which usually comes from strangers, this new form spam, also named as "social" spam, often appears to be from a "friend" in social websites. Spammers find social websites alluring because they can broadcast spam messages through a chain of trusted sources and target at a large number of users. According to Fowler et al. (2012), 4 % of the content shared on Facebook is spam. Making matters even worse, a statistical report from Facebook (Fowler et al. 2012) indicates that the volume of social spam is growing much faster than its user base.

Foundations

The majority of existing efforts to combat spam are based on filtering spam messages using spam filters. In this section, we first describe the general framework of spam detection using spam filters and then discuss some representative spam detection solutions for e-mail spam and social spam, respectively.

A General Framework

The success of spam detection relies on spam filters, which are automated tools that are built to detect spam with the purpose of preventing its delivery. For every effective spam filter, the core

component is a spam classifier which categorizes whether a specific message is spam or not. The decision of spam classification using a spam filter is often made based on different pieces of information. For example, in order to construct an effective e-mail spam filter, the content of the e-mail messages and the characteristics of the e-mail sender and the e-mail receiver are usually considered. In addition, collaborative knowledge such as the feedback of other receipts receiving similar e-mails is also valuable for building the spam filter. In many cases, some external pieces of information (e.g., spam repositories) provided by some third parties are also considered.

Once a spam classifier is constructed, a specific message can be categorized as spam or not. A simple way to obtain the spam categorization results is using a binary classifier. In a binary classifier, a message is either labeled as spam or non-spam. Although this solution is simple, it suffers from the lack of adaptability. Users have little control in the spam detection process. A more common way is to use probabilistic classifiers which can provide more informative indications on how likely the spam classifier considers the message to be spam. For example, some spam classifiers can calculate a spamicity score for each message. The spamicity score is in the range [0,1]. The larger the spamicity score, the more likely the spam classifier considers the message to be spam. In practice, a spamicity threshold value is often configured to filter spam messages. Different from the binary spam classifier, users are able to adjust the threshold value so as to capture different spam detection scenarios.

Spam filters automatically filter those messages that are labeled as spam. For example, spam e-mails are automatically placed into the junk folder in each user's e-mail account; spam web pages are automatically removed from the web search results; spam information in online social media is automatically flagged and is prohibited to be propagated through social networks.

E-mail Spam Detection

The methods for e-mail spam detection have been evolved continuously in the past years. In the early stage, spam filter for e-mail spam detection is mainly based on receiver's judgements. For example, users have the opportunity to handcraft several logical rules and guidelines to filter spam e-mails. A common practice is to maintain a whitelist and a blacklist in each user's e-mail account. A whitelist refers to a list of contacts whose e-mails should be delivered. Oppositely, a blacklist refers to a list of contacts whose e-mails are deemed to be spam. This solution is acceptable when the volume of contacts and e-mails is low. However, this solution becomes problematic when the volume gets larger and larger. In addition, the success of manually handcrafted spam filter relies on user's judgements. The assumption that users are savvy enough to construct robust spam filtering rules is questionable. To make matters worse, when spam e-mails change over time, it becomes an even time-consuming and error-prone process for users to constantly turn and refine those spam filtering rules.

To address the problems with the manual construction of spam filtering rules, latest spam filters for e-mail spam detection are able to automatically adapt to the changing characteristics of spam e-mails over time. Machine learning algorithms have been widely adopted to build robust and adaptive spam filters. Since those spam filters are built directly from user's e-mail repository, they are able to be personalized to meet particular characteristics of each user. In other words, these spam filters are tailored specifically to meet each individual's requirements on spam judgements.

Among many machine learning-based spam filters, the one based on Bayesian classifier is most popular and widely adopted (Sahami et al. 1998). A Bayesian classifier is a probabilistic classifier. It uses Bayesian inference to calculate a probability which indicates how likely an e-mail message is spam. The motivation of using Bayesian classifier for e-mail spam detection is that particular information (e.g., words in the e-mail) has particular probabilities to occur in either spam or legitimate e-mails. If all such probabilities are calculated, they can be used to compute the overall probability that a specific e-mail message with a particular set of

information in it belongs to either spam category or non-spam category.

In practice, the Vector Space model in which each dimension corresponds to a given word in the entire corpus is often adopted. Thus, each individual message can be represented as a binary vector denoting which words are present or absent. Some other pieces of information, for example, domain-specific properties, can also be incorporated into the representation. Specifically, an e-mail message is represented as a vector of n features $x = (x_1, x_2, \ldots, x_n)$. Each feature X_i $(1 \leq i \leq n)$ represents a specific piece of information, and x_i $(1 \leq i \leq n)$ is the value pertaining to feature X_i.

Assume that there are m different classes to categorize e-mail messages, each is denoted as c_j $(1 \leq j \leq m)$. Consider a specific e-mail message x, the Bayesian classifier can calculate the probability $P(C = c_j | \mathbf{X} = \mathbf{x})$ for each possible class c_j $(1 \leq j \leq m)$. The calculation is achieved according to the well-known Bayes' theorem, that is,

$$P(C = c_j | \mathbf{X} = \mathbf{x}) = \frac{P(\mathbf{X} = \mathbf{x} | C = c_j) P(C = c_j)}{P(\mathbf{X} = \mathbf{x})}. \quad (1)$$

In Equation 1, the calculation $P(\mathbf{X} = \mathbf{x} | C = c_j)$ is often impractical without imposing some independence assumptions. Thus, the Naive Bayesian classifier is often adopted for the calculation. Given the class variable C, the Naive Bayesian classifier assumes that each feature X_i is conditionally independent of every other feature. As a result, we have

$$P(\mathbf{X} = \mathbf{x} | C = c_j) = \prod_i P(X_i = x_i | C = c_j). \quad (2)$$

Spam filter using Bayesian classifier has been shown a very powerful technique for dealing with e-mail spam. It can tailor itself to the specific needs of individual users and provides low false positive spam detection rates. However, due to the intrinsic problem of Bayesian classifier, the constructed spam filter may also be susceptible to Bayesian poisoning, a technique used by spammers in an attempt to degrade the effectiveness of spam filters.

In the recent decade, many efforts have been devoted to constructing effective and efficient machine learning-based spam filters. Many existing studies mainly focus on two critical issues in building a spam filter: one is feature selection, that is, selecting a subset of relevant features for building robust spam filters; the other is classifier construction, that is, using different machine learning methods to "learn" robust spam classifiers from training data.

As many pieces of information can be extracted from e-mail messages but not all of them are useful for building spam filters, the feature selection problem needs to find the best subset of the available features. The concept of "best" may rely on different factors including the number of selected features, the effectiveness of the trained spam classifier, and the tractability of the algorithm to perform the selection process. In general, given n features of e-mail messages, a straightforward but prohibitively inefficient solution is to examine all 2^n subsets of n features and choose the one which achieves the highest spam detection result. In practice, some greedy heuristics are adopted for ranking features in decreasing order based on some characterizations of their usefulness for spam detection. For example, Sebastiani (2002) analyzed several heuristics relying on statistics such as term frequency and information gain for ranking and selecting features. Regardless of particular statistics of different features, selecting the optimal feature set is always a challenge.

The process of feature selection should not be considered separately from the process of classifier construction. Different machine learning algorithms have been considered for constructing spam filters. For example, the k-Nearest Neighbors Classifier (Firte et al. 2010) builds on top of the k-NN algorithm and classifies a message according to the classes of its nearest neighbors in the training data. Artificial neural networks are also applicable for constructing spam filters. In particular, the algorithms such as perceptron and multilayer perceptron (Tran et al. 2008) have been shown

quite successful for filtering e-mail spams. Several recent studies also considered the application of SVM for spam detection. The motivation of the SVM classification (Zhang et al. 2004) is to find a separation boundary which can correctly classify training samples. Different from the perceptron algorithm, the SVM-based approach tries to find a special maximal margin separating hyperplane such that the distance to the closest training sample is maximal.

In practice, there exist some types of e-mail messages that cannot be clearly categorized as either spam or non-spam. Such examples include newsletters and legitimate advertisements. These types of e-mail messages are usually regarded as gray mail (Chang et al. 2008). Detecting gray mail introduces more challenges. Even an optimal spam filter could inevitably perform unsatisfactorily on gray mail. Chang et al. (2008) systematically studied the problem of gray mail detection and concluded that user preferences are needed to be considered. The experimental results in Chang et al. (2008) indicate that e-mail messages which are labeled differently in the training data are the most reliable source for learning a gray mail detector.

Social Networks Spam Detection

Online social websites have different characteristics comparing to e-mail systems. Some spam filtering techniques for e-mail spam detection may be useful to detect spam in social networks as well; however, some particular requirements of social spam detection need to be considered. Following are the four most important features of online social websites, which to some extent differentiate the characteristics of social spam detection compared to traditional e-mail spam detection:

- Existence of one managing entity. In online social websites, there exists an entity who manages and maintains the system, defines the system policy, and determines the privileges of participated users.
- Well-defined social interactions. Users have very close interactions with the social websites to contribute social contents. Meanwhile, social websites also provide some functionalities to share and distribute users' contents. However, the available interactions of participated users are constrained in the system.
- Unique identifier. In online social websites, each user has to maintain a unique identifier or a personal profile. This unique identifier is associated with each user's interaction in the websites.
- Multiple views of information access. Users in online social websites have multiple views to get access to the available contents.

Users are a key component in social spam detection. Social spam detection has several unique challenges. First, unlike many e-mail systems, the managing entity and restricted interactions in online social websites provide the opportunity to prevent spam effectively even before its emergence. For example, by defining appropriate terms of service and adjusting the trade off between users' privileges and the information flow rate, social websites are able to keep spam in the prevention stage. Second, due to the unique identifier in social websites, the origins of social spam can be controlled since users' interactions are tied to a specific identifier. Third, multiple views of information access lead to different snapshots of available contents in the websites. Therefore, social spam detection should consider all the possible spam tricks and various relations among them. Last but not the least, social websites contain large population and their social interactions, which makes information propagation much faster. This results in increasing and dynamic evolution of social spam. Inevitably, scalability and timely detection requirements become key issues in social spam detection.

Several popular anti-spam strategies for online social websites, named as detection, demotion, and prevention, are analyzed thoroughly in Heymann et al. (2007). Detection is made based on the predefined discriminative features extracted from given spam and non-spam instances. This is similar to e-mail spam detection. The features used for building spam filters are mainly extracted from contents and topological structure of social networks. In addition, the analysis of users' social behavior

and domain-specific features (e.g., features extracted from figures or videos) are often largely considered. Different from e-mail spam detection, demotion and prevention are also considered in social spam detection. The demotion strategy adopts rank-based methods to downgrade the prominence of contents that are deemed to be spam. In some situations, due to the fast propagation and evolution capabilities of social websites, decreasing the rankings of spam messages might not be enough.

Many current techniques of social spam detection largely depend on the set of features extracted from user behaviors and social interactions. Although such features are useful for social spam detection, there is always a considerable time delay until the spam is successfully identified. The fast information exchanging rate in social websites requires a real-time framework to combat spam. To achieve this goal, prevention-based strategy to identify social spam becomes quite useful (Irani et al. 2010). Once user profiles are created in social websites, some features are directly extracted from the static profile contents. The motivation is to identify the potential spammers in the early stage, even before the creation and propagation of spam messages in social networks. A popular solution is to treat social spam detection as an adversarial classification problem (Dalvi et al. 2004). However, this prevention-based solution may be vulnerable. In practice, this technique is often used as a filter even before many sophisticated spam detection techniques are employed. For example, user profiles that are deemed to be spammers are treated as gray profiles. These user profiles need particular attentions for further analysis to support spam detection.

Some recent studies (Boykin and Roychowdhury 2005) proposed an integrated framework of social spam detection and e-mail spam detection and applied social network analysis for e-mail spam detection. The algorithm proposed in Boykin and Roychowdhury (2005) analyzes "From," "To," Cc" and "Bcc" fields of the e-mail headers so as to construct a network representing social relations of different users. The foundation is based on the fact that the underlying e-mail social networks are useful for judging the trustworthiness of users. For example, the trust can be measured based not only on how well a user knows a specific person but also on how well the other users in the e-mail network know that person. Once the social network of e-mail communications is built, an automated anti-spam tool can exploit the properties of social networks to distinguish spam messages from non-spam ones.

Conclusion

There is an adversarial relationship between spam and anti-spam techniques. In recent years, machine learning-based spam detection approaches, the probabilistic classifier-based spam filter in particular, have been widely applied to detect various forms of spam. However, the performance of anti-spam techniques is still far from perfect. The creativity and efforts of spammers who manage to violate laws and social norms to deliver spam messages will provide a continuing challenge for developing anti-spam techniques. Developing robust and adaptive anti-spam techniques is a long-term strategy to combat spam.

Cross-References

- Dark Sides of Social Networking
- Ethics of Social Networks and Mining
- Online Social Network Phishing Attack
- Social Engineering/Phishing
- Trust in Social Networks

References

Abu-Nimeh S, Chen T (2010) Proliferation and detection of blog spam. IEEE Secur Priv 8(5):42–47

Boykin PO, Roychowdhury VP (2005) Leveraging social networks to fight spam. Computer 38(4):61–68. doi:http://dx.doi.org/10.1109/MC.2005.132

Chang M, Yih W, McCann R (2008) Personalized spam filtering for gray mail. In: Proceedings of the fifth conference on email and anti-spam (CEAS'08). Mountain View, CA

Cormack GV (2008) Email spam filtering: a systematic review. Found Trends Inf Retr 1(4):335–455

Dalvi N, Domingos P, Mausam, Sanghai S, Verma D (2004) Adversarial classification. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, KDD '04. ACM, New York, pp 99–108. doi:10.1145/1014052.1014066. http://doi.acm.org/10.1145/1014052.1014066

Firte L, Lemnaru C, Potolea R (2010) Spam detection filter using KNN algorithm and resampling. In: Proceedings of the 6th international conference on intelligent computer communication and processing (ICCP '10). Cluj-Napoca, Romania

Fletcher D (2009) A brief history of spam. Time. http://www.time.com/time/business/article/0,8599,1933796,00.html

Fowler GA, Raice S, Efrati A (2012) Spam finds new target: Facebook and Twitter build up their defenses as hackers attack social networks. The Wall Street Journal

Gyöngyi Z, Garcia-Molina H (2005) Web spam taxonomy. In: First international workshop on adversarial information retrieval on the Web (AIRWeb'05). Chiba, Japan

Heymann P, Koutrika G, Garcia-Molina H (2007) Fighting spam on social web sites: a survey of approaches and future challenges. IEEE Internet Comput 11(6):36–45

Irani D, Webb S, Pu C (2010) Study of static classification of social spam profiles in MySpace. In: Proceedings of the fourth international conference on weblogs and social media. Washington, DC

Jennings R (2009) Cost of spam is flattening: our 2009 predictions. Ferris research. http://email-museum.com/2009/01/28/cost-of-spam-is-flattening-our-2009-predictions/

Jin X, Lin CX, Luo J, Han J (2011) Socialspamguard: a data mining-based spam detection system for social media networks. PVLDB 4(12):1458–1461

Paulson LD (2004) Spam hits instant messaging. In: Computer, vol 37, issue 4. IEEE Computer Society Press, Los Alamitos

Phuoc TT, Po-Hsiang T, Tony J (2008) An adjustable combination of linear regression and modified probabilistic neural network for anti-spam filtering. In: Proceedings of the19th International Conference on Pattern Recognition. Florida, USA

Sahami M, Dumais S, Heckerman D, Horvitz E (1998) A bayesian approach to filtering junk E-mail. In: Learning for text categorization: papers from the 1998 workshop, AAAI Technical Report WS-98-05, Madison, Wisconsin. citeseer.ist.psu.edu/sahami98bayesian.html

Sebastiani F (2002) Machine learning in automated text categorization. ACM Comput Surv 34(1):1–47. doi10.1145/505282.505283. http://doi.acm.org/10.1145/505282.505283

Shin Y, Gupta M, Myers SA (2011) Prevalence and mitigation of forum spamming. In: Proceedings of the 30th IEEE international conference on computer communications. Shanghai, China

Zhang L, Zhu J, Yao T (2004) An evaluation of statistical spam filtering techniques. ACM Trans Asian Lang Inf Process 3(4)

SPARQL

Axel Polleres
Siemens AG Österreich, Vienna, Austria

Synonyms

SPARQL 1.1; W3C Standard RDF Query Language

Glossary

RDF Resource Description Framework
RDFS RDF Schema, a lightweight ontology language on top of RDF
Triple An atomic statement of the form (subject predicate object) in RDF
RDF Graph A set of RDF triples
W3C World Wide Web Consortium
W3C Recommendation Standards published by the W3C
SPARQL initially "Simple Protocol and RDF Query Language," or nowadays more often referred to by the recursive acronym "SPARQL Protocol and RDF Query Language"; in its 1.1 version, SPARQL comprises not only a query language and a protocol but also a data manipulation language and other features
SPARQL Protocol defines how to invoke SPARQL queries and updates via a SPARQL endpoint and how results should be returned via HTTP

SPARQL Service Any implementation conforming to the SPARQL Protocol

SPARQL Endpoint The URI at which a SPARQL service listens for requests from clients

OWL Web Ontology Language, a schema language on top of RDF, rooted in Description Logics

RIF Rule Interchange Format, a standard to encode and exchange rules

BGP Basic Graph Pattern, a set of RDF triple "templates" where variables are allowed in either subject predicate or object position, which can be read as a conjunctive query

HTTP Hypertext Transfer Protocol

URI Universal Resource Identifiers, a generalization of URLs (cf. IETF RFC1630)

Definition

SPARQL, the "Simple Protocol and RDF Query Language," is the W3C's standard query language for RDF (the Resource Description Framework, an emerging data format on the growing Web of Data). However, the SPARQL standard does not only comprise a query language but a family of W3C standards to access and manipulate RDF data; in its current version SPARQL 1.1, the standard comprises:

- A query language (SPARQL 1.1 Query Language)
- A data manipulation language (SPARQL 1.1 Update)
- A mechanism to describe and discover SPARQL endpoints (SPARQL 1.1 Service Description)
- An extension to delegate parts of a query to a remote SPARQL endpoint (SPARQL 1.1 Federated Query)
- Various result formats (SPARQL 1.1 Query Results JSON Format, SPARQL 1.1 Query Results CSV and TSV Formats, SPARQL Query Results XML Format)
- A normative way to return additional results entailed by schema and rules languages such as RDFS, OWL, and RIF (SPARQL 1.1 Entailment Regimes)
- A protocol to invoke SPARQL queries and updates via HTTP (SPARQL 1.1 Protocol)
- An extension to the SPARQL Protocol, to perform certain operations to manage collections of graphs directly via HTTP (SPARQL 1.1 Graph Store HTTP Protocol)

Introduction

The Semantic Web is in principle a family of standards to enable a Web of Data, with the final goal of enabling nothing less than the vision of the Web as a database (Berners-Lee 1999). The architecture of these standards comprises of (i) a simple graph-based data model, RDF; (ii) schema languages, RDFS and OWL; (iii) rules languages, RIF; and last but not least, (iv) a query language, sparql. The existence of such a standard query language has significantly contributed to the increasing uptake of RDF as a basic data format on the Web over the past years. After SPARQL's first edition has become a W3C recommendation in 2008, the community and implementers have requested a variety of additional features that the SPARQL 1.1 working group took as a starting point in 2009 for re-shaping the next version of the standard. In March 2013, the group concluded its work by publishing 11 specification documents (listed above) as a W3C recommendation.

Methodology

In this section, we introduce various parts of the SPARQL specification by a short example.

We will illustrate the use of SPARQL's languages, protocols, and related specifications with a small example RDF graph published on the Web at the URL "http://example.org/alice" which contains personal information about Alice and her social contacts. We use Turtle (Beckett et al. 2013) syntax here for illustration:

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

<http://example.org/alice#me> a foaf:Person .
<http://example.org/alice#me> foaf:name "Alice" .
<http://example.org/alice#me> foaf:mbox
<mailto:alice@example.org> .
<http://example.org/alice#me> foaf:knows
<http://example.org/bob#me> .
<http://example.org/bob#me> foaf:knows
<http://example.org/alice#me> .
<http://example.org/bob#me> foaf:name "Bob" .
<http://example.org/alice#me> foaf:knows
<http://example.org/charlie#me> .
<http://example.org/charlie#me> foaf:knows
<http://example.org/alice#me> .
<http://example.org/charlie#me> foaf:name "Charlie" .
<http://example.org/alice#me> foaf:knows
<http://example.org/snoopy> .
<http://example.org/snoopy> foaf:name "Snoopy"@en .
```

With SPARQL 1.1, one can query such graphs, load them into RDF stores, and manipulate them in various ways.

Firstly, the *SPARQL 1.1 Query Language* (Harris and Seaborne 2013) can be used to formulate queries against RDF ranging from simple graph pattern matching to complex queries. For instance, one can ask using a SPARQL SELECT query for names of persons and the number of their friends:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name (COUNT(?friend) AS ?count)
WHERE{
 ?person foaf:name ?name .
 ?person foaf:knows ?friend .
 }GROUP BY ?person ?name
```

Complex queries may include union, optional query parts, and filters; new features like value aggregation, path expressions, and nested queries have been added in SPARQL 1.1. Apart from SELECT queries – which return variable bindings – SPARQL supports ASK queries, i.e., Boolean "yes/no" queries, and CONSTRUCT queries, by which new RDF graphs can be constructed from a query result; all the new query language features of SPARQL 1.1 are likewise usable in ASK and CONSTRUCT queries.

Results of SELECT queries in SPARQL comprise bags of mappings from variables to RDF terms, often conveniently represented in tabular form. For instance, the query from Section 2 has the following results:

In order to exchange these results in machine-readable form, SPARQL supports four standard

?name	?count
"Alice"	3
"Bob"	1
"Charlie"	1

formats to exchange results, namely, the Extensible Markup Language (XML) (Hawke 2013), the JavaScript Object Notation (JSON) (Seaborne 2013a), as well as the Comma-Separated Values (CSV) and Tab-Separated Values (TSV) (Seaborne 2013b).

The *SPARQL 1.1 Federated Query* (Prud'hommeaux and Buil-Aranda 2013) extension allows to explicitly delegate certain subqueries to different SPARQL endpoints. For instance, in our example, one may want to know whether there is anyone among Alice's friends with the same name as the resource identified by the IRI <http://dbpedia.org/resource/Snoopy> at DBpedia. This can be done by combining a query for the names of friends with a remote call to the SPARQL endpoint at http://dbpedia.org/sparql finding out the name of <http://dbpedia.org/resource/Snoopy> using the SERVICE keyword as follows:

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
SELECT ?name WHERE {
<http://example.org/alice#me> foaf:knows [ foaf:name ?name] .
   SERVICE <http://dbpedia.org/sparql>
   { <http://dbpedia.org/resource/Snoopy> foaf:name ?name } }
```

Here, the first part of the pattern in the WHERE part is still matched against the local SPARQL service, whereas the evaluation of the pattern following the SERVICE keyword is delegated to the respective remote SPARQL service.

SPARQL can be used together with *entailment regimes* (Glimm and Ogbuji 2013), that is, exploiting ontological information in the form of, for example, RDF Schema (RDFS) or OWL axioms. For instance, let us assume that – apart from the data about Alice – some ontological information in the form of RDFS (Brickley and Guha 2004) and OWL (2012) constructs defining the FOAF vocabulary is loaded into our example SPARQL service.

The FOAF ontology (cf. http://xmlns.com/foaf/spec/,retrievedApril2013), of which we only give a small excerpt here, contains, for instance, the following RDFS axiom:

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
...
foaf:name rdfs:subPropertyOf rdfs:label .
...
```

The following query asks for labels of persons:

```
SELECT ?label
WHERE {?person rdfs:label ?label}
```

A SPARQL engine that does not consider any special entailment regimes (on top of standard simple entailment) would not return any results for this query, whereas an RDF Schema aware query engine will return

Since foaf:name is a sub-property of rdfs:label.

?label
"Alice"
"Bob"
"Charlie"
"Snoopy"@en

The *SPARQL 1.1 Update* (Gearon et al. 2013) specification defines the syntax and semantics of SPARQL 1.1 Update requests. Update operations can consist of several sequential requests and are performed on a collection of graphs in a Graph Store. Operations are provided to update, create, and remove RDF graphs in a Graph Store. For instance, the following request inserts a new friend of Alice named Dorothy into the default graph of our example SPARQL service and thereafter deletes all names of Alice's friends with an English language tag.

```
PREFIX foaf: <http://xmlns.com/foaf/0.1/> .

INSERT DATA { <http://www.example.org/alice#me> foaf:knows
[foaf:name "Dorothy" ]. } ;
DELETE { ?person foaf:name ?mbox }
WHERE {
   <http://www.example.org/alice#me> foaf:knows ?person .
   ?person foaf:name ?name . FILTER ( lang(?name) = "EN" ) .}
```

As the second operation shows, insertions and deletions can be dependent on the results of queries to the Graph Store; the respective syntax used in the WHERE part is derived from the SPARQL 1.1 Query Language.

The *SPARQL 1.1 Protocol for RDF* (Feigenbaum et al. 2013) defines how to transfer SPARQL 1.1 queries and update requests to a SPARQL service via HTTP. It also defines how to map requests to HTTP GET and POST operations and what respective HTTP responses to such requests should look like. Additionally, the *SPARQL 1.1 Service Description* (Williams 2013) document describes a method for discovering and an RDF vocabulary for describing SPARQL services made available via the SPARQL 1.1 Protocol. According to this specification, a service endpoint, when accessed via an HTTP GET operation without further (query or update request) parameters, should return an RDF description of the service provided.

For many applications and services that deal with RDF data, the full SPARQL 1.1 Update language might not be required. To this end, the *SPARQL 1.1 Graph Store HTTP Protocol* (Ogbuji 2013) provides means to perform certain operations to manage collections of graphs directly via HTTP operations.

For instance, the first part of the update request in above is a simple insertion of triples into an RDF graph. On a service supporting this protocol, such insertion can – instead of via a SPARQL 1.1 Update request – directly be performed via an HTTP POST operation taking the RDF triples to be inserted as payload.

Implementations

A list of SPARQL 1.0 implementations is available at http://www.w3.org/wiki/SparqlImplementations (retrieved April 2013), whereas a list of implementations of the new features of SPARQL 1.1 along with reports on test coverage is available at http://www.w3.org/2009/sparql/implementations/ (retrieved April 2013). As for performance evaluations, a list of benchmarks is available at http://www.w3.org/wiki/RdfStoreBenchmarking; The Europeana report (Haslhofer et al. 2011) compares and describes various current SPARQL

implementations, not yet mentioning SPARQL 1.1 implementations, though.

SPARQL in Academia

The formal semantics of the SPARQL Query Language in its original recommendation in 2008 has been very much inspired by academic results, such as by the papers of Pérez et al. (2006, 2009). Angles and Gutierrez (2008) later showed that SPARQL – as defined in those papers – has exactly the expressive power of non-recursive safe Datalog with negation. Another translation from SPARQL to Datalog has been presented in Polleres (2007).

Extensions that were now standardized in SPARQL 1.1 such as subqueries (Angles and Gutierrez 2011), path expressions (Alkhateeb et al. 2009; Pérez et al. 2010), or aggregates (Polleres et al. 2007) have also been discussed or proposed in some variants in the academic literature. Details about the differences of the semantics as defined in the official W3C specification and in most of these academic papers are discussed in Polleres (2012). Query optimization and particularly equivalence of SPARQL queries have been discussed to some extent already in Pérez et al. (2009). These results were refined and extended in Schmidt et al. (2008), Letelier et al. (2012), and Chekol et al. (2012). The semantics of SPARQL entailment regimes has been discussed in Kollia et al. (2011); foundational aspects of federated queries are discussed in Buil-Aranda et al. (2013). The semantics of path expressions in SPARQL 1.1 has been discussed in Arenas et al. (2012) and Losemann and Martens (2012), and it should be noted that these papers to some extent influenced the definition of the semantics of path expressions in the final specification. More practical proposals for query optimizations are discussed in Stocker et al. (2008) and Vidal et al. (2010). Overall, SPARQL is a source of ongoing research and inspired various academic works on its foundations, optimization, and extensions, a full account of which would be beyond the scope of this article.

Future Directions

Various additional features requested to the query language could not yet be taken into account for SPARQL 1.1, and the working group has collected a list of open work items on its wiki page (http://www.w3.org/2009/sparql/wiki/Future_Work_Items, retrieved April 2013) which comprises features that had to be left out either for reasons of priorities or missing implementation experience to be standardized already. It may be expected that – just like in the transition from SPARQL 1.0 to SPARQL 1.1 – upon implementation experience and community feedback from implementers, a new working group by W3C will be formed in the future to add additional features. As already mentioned in the previous section, academic research can potentially impact these future directions; for instance, extensions of regular path queries (a small subset of which is now incorporated into SPARQL 1.1) which are currently being investigated in academia (e.g., Barceló et al. 2010) might be viewed as very valuable additions to query graph data in RDF.

Acknowledgments

The author would like to thank all members of the W3C SPARQL working group as well as various people who helped to improve the standard by their comments to public-rdf-dawg-comments@w3.org. A more detailed version of the examples provided herein can be found in the SPARQL 1.1 Overview document (SPARQL 1.1 Overview 2013).

Cross-References

▸ Web Ontology Language (OWL)
▸ RDF
▸ Reasoning
▸ RIF: The Rule Interchange Format
▸ Xpath/XQuery

References

Alkhateeb F, Baget J-F, Euzenat J (2009) Extending SPARQL with regular expression patterns (for querying RDF). J Web Semant 7(2):57–73

Angles R, Gutierrez C (2008) The expressive power of SPARQL. In: International semantic web conference, Karlsruhe, pp 114–129

Angles R, Gutierrez C (2011) Subqueries in SPARQL. In: Alberto Mendelzon international workshop on foundations of data management, Santiago

Arenas M, Conca S, Pérez J (2012) Counting beyond a Yottabyte, or how SPARQL 1.1 property paths will prevent adoption of the standard. In: WWW 2012, Lyon, pp 629–638

Barceló P, Hurtado CA, Libkin L, Wood PT (2010) Expressive languages for path queries over graph-structured data. In: PODS 2010, Indianapolis, pp 3–14

Beckett D, Berners-Lee T, Prud'hommeaux E, Carothers G (eds) (2013) Terse RDF Triple Language. W3C Candidate Recommendation, 19 Feb 2013

Berners-Lee T (1999) Weaving the web. Harper, San Francisco

Brickley D, Guha RV (eds) (2004) RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation, 10 Feb 2004

Buil-Aranda C, Arenas M, Corcho O, Polleres A (2013) Federating queries in SPARQL 1.1: syntax, semantics and evaluation. J Web Semant 18(1):1–17

Chekol MW, Euzenat J, Genevès P, Layaïda N (2012) SPARQL query containment under SHI axioms. In: AAAI 2012, Toronto

Feigenbaum L, Williams GT, Clark KG, Torres E (eds) (2013) SPARQL 1.1 Protocol. W3C Recommendation, 21 Mar 2013

Gearon P, Passant A, Polleres A (eds) (2013) SPARQL 1.1 Update. W3C Recommendation, 21 Mar 2013

Glimm B, Ogbuji C (eds) (2013) SPARQL 1.1 Entailment Regimes. W3C Recommendation, 21 Mar 2013

Harris S, Seaborne A (eds) (2013) SPARQL1.1 Query Language. W3C Recommendation, 21 Mar 2013

Haslhofer B, Roochi EM, Schandl B, Zander S (2011) Europeana RDF store report. Technical report, University of Vienna, Vienna

Hawke S (ed) SPARQL Query Results XML Format, 2nd edn. W3C Recommendation, 21 Mar 2013

Kollia I, Glimm B, Horrocks I (2011) SPARQL query answering over OWL ontologies. In: ESWC 2011 (1), Heraklion, pp 382–396

Letelier A, Pérez J, Pichler R, Skritek S (2012) Static analysis and optimization of semantic web queries. In: PODS 2012, Scottsdale, pp 89–100

Losemann K, Martens W (2012) The complexity of evaluating path expressions in SPARQL. In: PODS 2012, Scottsdale, pp 101–112

Ogbuji C (ed) (2013) SPARQL 1.1 Graph Store HTTP Protocol. W3C Recommendation, 21 Mar 2013

OWL (2012) OWL 2 Web Ontology Language Document Overview (2nd edn). W3C Recommendation, 11 Dec 2012

Pérez J, Arenas M, Gutierrez C (2006) Semantics and complexity of SPARQL. In: International semantic web conference, Athens, pp 30–43

Pérez J, Arenas M, Gutierrez C (2009) Semantics and complexity of SPARQL. ACM Trans Database Syst 34(3):6

Pérez J, Arenas M, Gutierrez C (2010) nSPARQL: a navigational language for RDF. J Web Semant 8(4):255–270

Polleres A (2007) From SPARQL to rules (and back). In: WWW, Banff, pp 787–796

Polleres A (2012) How (well) do datalog, SPARQL and RIF interplay? In: Datalog 2012 workshop, Vienna, pp 27–30

Polleres A, Scharffe F, Schindlauer R (2007) SPARQL++ for mapping between RDF vocabularies. In: OTM conferences (1), Vilamoura, pp 878–896

Prud'hommeaux E, Buil-Aranda C (eds) (2013) SPARQL 1.1 Federated Query. W3C Recommendation, 21 Mar 2013

Seaborne A (ed) (2013a) SPARQL 1.1 Query Results JSON Format. W3C Recommendation, 21 Mar 2013

Seaborne A (ed) (2013b) SPARQL 1.1 Query Results CSV and TSV Formats. W3C Recommendation, 21 Mar 2013

Schmidt M, Meier M, Lausen G (2008) Foundations of SPARQL query optimization. CoRR abs/0812.3788

SPARQL 1.1 Overview (2013) SPARQL 1.1 Overview. W3C Recommendation, 21 Mar 2013

Stocker M, Seaborne A, Bernstein A, Kiefer C, Reynolds D (2008) SPARQL basic graph pattern optimization using selectivity estimation. In: WWW 2008, Beijing, pp 595–604

Vidal ME, Ruckhaus E, Lampo T, Martínez A, Sierra J, Polleres A (2010) Efficiently joining group patterns in SPARQL queries. In: ESWC 2010 (1), Heraklion, pp 228–242

Williams GT (ed) (2013) SPARQL 1.1 Service Description. W3C Recommendation, 21 March 2013

SPARQL 1.1

▸ SPARQL

Spatial Analysis

▸ Spatial Statistics

Spatial Interaction

▸ Spatiotemporal Footprints in Social Networks

Spatial Networks

Marc Barthelemy
Institut de Physique Théorique, CEA,
CNRS-URA 2306, Gif-sur-Yvette, France

Synonyms

Network geography; Space-embedded networks; Transportation systems; Urban networks

Glossary

Graph (or Network) A set of vertices connected by edges

Adjacency Matrix A matrix A which represents the structure of a graph. The element A_{ij} is either 0 if i and j are not connected or $A_{ij} = 1$ if there is an edge from i to j. For a spatial network, the position of the nodes $\{x_i\}$ is needed in order to completely characterize the network

Betweenness Centrality The betweenness centrality of a vertex (or an edge) x is defined as $BC(x) = \sum_{s,t \in V} \frac{\sigma_{st}(x)}{\sigma_{st}}$, where $\sigma_{st}(x)$ is the number of shortest paths between s and t using x and σ_{st} is the number of all shortest paths between s and t

Betweenness Centrality Impact Measures how a new link affects the average betweenness centrality of a graph. This quantity can help in characterizing the different types of new links during the evolution of a (spatial) network

Cell Also called *face* for planar network is a region bounded by edges. The Euler formula relates the number of nodes, edges, and cells (faces)

Diameter The diameter of a graph is defined as the maximum value of all $\ell(i, j)$, is the distance between i and j, and is used to measure the "size" of it. For most real-world spatial network, the diameter scales as the number of nodes to the power $1/d$ where d is the dimension of the embedding space

Planar Graph A planar graph can be drawn in 2-D such that none of its edges are crossing

Organic Ratio Measures the proportion of degree 1 ("dead ends") and degree 3 nodes ("T-shaped intersections"). If the organic ratio is small, the corresponding spatial network is very close to a regular rectangular lattice

Alpha Index Also called the meshedness, it measures the ratio of observed circuits to the maximum number of elementary circuits which can exist in the network

Gamma Index Ratio of the number of edges to the maximum number possible for a planar graph with the same number of nodes

Shape Factor Ratio of the area of a cell to the area of the circumscribed circle

Route Distance Distance between two nodes measured by the length of the shortest path connecting them

Detour Index Ratio of the route distance between two nodes and the euclidean distance between them

Network Cost Ratio of the total length of the network to the total length of the minimum spanning tree constructed on the same set of nodes

Network Performance Ratio of the average shortest path of the network to the average shortest path of the minimum spanning tree constructed on the same set of nodes

Definition

More generally, the term "spatial network" has come to be used to describe any network in which the nodes are located in a space equipped with a metric (Barthelemy 2011). For most practical applications, the space is the two-dimensional space and the metric is the usual euclidean distance. For these networks we thus need both the topological information about the graph (given by the adjacency matrix) and the spatial information

about the nodes (given by the position of the nodes).

Transportation and mobility networks, Internet, mobile phone networks, power grids, social and contact networks, and neural networks are all examples where space is relevant and where topology alone does not contain all the information.

Characterizing and understanding the structure and the evolution of spatial networks is crucial for many different fields ranging from urbanism to epidemiology. An important consequence of space on networks is that there is usually a cost associated to the length of edges which in turn has dramatic effects on the topological structure of these networks. Indeed, a long link will be very costly and can exist if this cost if balanced with another good reason (economical or connection to a hub, ...). For most real-world spatial networks, we indeed observe that the probability of finding a link between two nodes will decrease with the distance. Spatial constraints affect not only the structure and properties of these networks but also processes which take place on these networks such as phase transitions, random walks, synchronization, navigation, resilience, and disease spread.

All planar graphs can be embedded in a two-dimensional space and can be represented as spatial networks, but the converse is not necessarily true: there are some spatial and nonplanar graphs. In general, however, most spatial networks are, to a good approximation, planar graphs (Clark and Holton 1991), such as road or railway networks, but there are some important exceptions such as the airline network (Barrat et al. 2004): in this case the nodes are airports and there is a link connecting two nodes if there is at least one direct connection. For many infrastructure networks, however, planarity is unavoidable. Power grids, roads, rail, and other transportation networks are to a very good accuracy planar networks. For many applications, planar spatial networks are the most important and most studies have focused on these examples.

Also, the above definition does not imply that the links are necessarily embedded in space. Indeed, in social networks, individuals are connected through a friendship relation which is a virtual network of relations. There is however a strong spatial component in these networks as the probability that individuals located in space are friends generally decreases with the distance between them (Liben-Nowell et al. 2005).

Introduction

For many critical infrastructures, communication or biological networks, space is relevant: most of the people have their friends and relatives in their neighborhood, power grids and transportation networks depend obviously on distance, many communication network devices have short radio range, the length of axons in a brain has a cost, and the spread of contagious diseases is not uniform across territories. In particular, in the important case of the brain, regions that are spatially closer have a larger probability of being connected than remote regions as longer axons are more costly in terms of material and energy (Bullmore and Sporns 2009). Wiring costs depending on distance are thus certainly an important aspect of brain networks, and we can probably expect spatial networks to be very relevant in this rapidly evolving topic. Another particularly important example of such a spatial network is the Internet which is defined as the set of routers linked by physical cables with different lengths and latency times. More generally, the distance could be another parameter such as a social distance measured by salary, socio-professional category differences, or any quantity which measures the cost associated with the formation of a link.

All these examples show that these networks have nodes and edges which are constrained by some geometry and are usually embedded in a two- or three-dimensional space, and this has important effects on their topological properties and consequently on processes which take place on them. If there is a cost associated to the edge length, longer links must be compensated by some advantage, for example, being connected to a well-connected node – that is, a hub. The topological aspects of the network are then correlated

to spatial aspects such as the location of the nodes and the length of edges.

Tools for Characterizing Spatial Networks

Graphs are usually characterized by the adjacency matrix A where the elements are $A_{ij} = 1$ if nodes i and j are connected (see, e.g., a graph textbook Clark and Holton 1991). This matrix completely characterizes the topology of the graph and is enough for most applications. This is however not the case for spatial networks where the spatial information is contained in the location of the nodes x_i. Two topologically identical graphs can then have completely different spatial properties, and this is at the heart of the richness and complexity of spatial networks.

In this section we will discuss some tools which can be helpful to characterize some aspects of spatial networks.

Degree Distribution, Clustering, and Average Shortest Path Length

Degree Distribution

In complex networks, the degree distribution, the clustering spectrum, and the average shortest distance are of utmost importance (Albert and Barabasi 2002). Their knowledge already gives a useful picture of the graph under study. In contrast, in spatial networks, physical constraints impose some of the properties. In particular, there is usually a sharp cutoff on the degree distribution $P(k)$ which is therefore not broad. This is true for most spatial and planar networks such as power grids or transportation networks, for example. For a spatial, nonplanar network such as the airline network, the cutoff can be large enough and the degree distribution could be characterized as broad.

Clustering

The clustering coefficient of a node counts how its neighbors are connected with each other. For spatial networks, the dominant mechanism is usually to minimize cost associated with length, and nodes have a tendency to connect to their nearest neighbors, independently from their degree. This in general implies that the clustering spectrum $C(k)$ is relatively flat for spatial networks. The same argument can be used to show that the assortativity "spectrum" defined as the function $k_{nn}(k)$ is also approximately constant in general when spatial constraints are very strong (see Barthelemy 2011 for more details).

Average Shortest Distance

Usually, there are many paths between two nodes in a connected network, and the shortest one defines a distance on the network:

$$\ell(i,j) = \min_{\text{paths}(i\to j)} |\text{path}| \tag{1}$$

where the length $|\text{path}|$ of the path is defined as its number of edges. This quantity is infinity when there are no paths between the nodes and is equal to one for the complete graph (for which $\ell(i,j) = 1$). For weighted graphs, we assign to each link e a weight w_e and the length of a path is given by $|\text{path}| = \sum_{e\in\text{Path}} w_e$.

In most complex networks, one observes a small-world behavior (Watts and Strogatz 1998) of the form

$$\langle \ell \rangle \sim \log N \tag{2}$$

In contrast, for a real-world spatial network embedded in a d-dimensional space, we usually observe the very different behavior:

$$\langle \ell \rangle \sim N^{1/d} \tag{3}$$

which also means that to go from one node to another one, one has to cross a path of length of the order of the diameter (which is not the case when shortcuts exist). The measure of the average shortest path length could thus be a first indication whether a network is close to a lattice or if long-range links are important.

Organic Ratio

We note that more recently, other interesting indices were proposed in order to characterize specifically road networks (Xie and Levinson 2007). Indeed, the degree distribution is very

peaked around 3–4, and an interesting information is given by the ratio

$$r_N = \frac{N(1) + N(3)}{\sum_{k \neq 2} N(k)} \quad (4)$$

where $N(k)$ is the number of nodes of degree k. If this ratio is small, the number of dead ends and of "unfinished" crossing ($k = 3$) is small compared to regular crossing with $k = 4$, signalling a more organized city. In the opposite case of large $r_N \simeq 1$, there is dominance of $k = 1$ and $k = 3$ nodes which signals a more "organic" city.

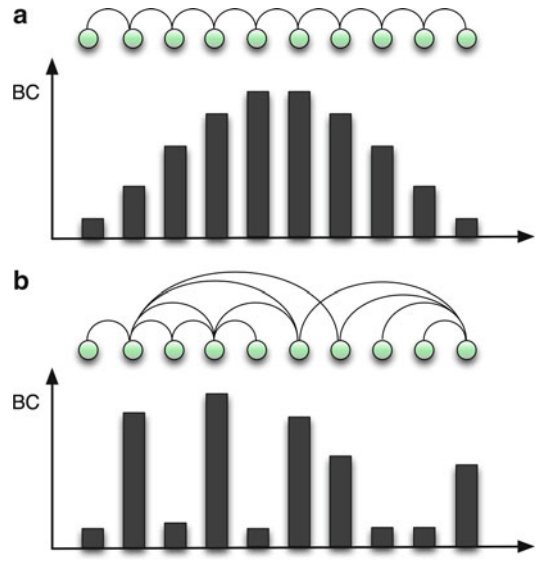
Betweenness Centrality

Anomalies

The betweenness centrality (BC) of a vertex (Freeman 1977) is determined by its ability to provide a path between separated regions of the network. Hubs are natural crossroads for paths, and it is natural to observe a marked correlation between the average $g(k) = \sum_{i/k_i=k} g(i)/N(k)$ and k as expressed in the following relation:

$$g(k) \sim k^{\eta} \quad (5)$$

where η depends on the characteristics of the network. We expect this relation to be altered when spatial constraints become important, and in order to understand this effect, we consider a one-dimensional lattice which is the simplest case of a spatially ordered network. For this lattice the shortest path between two nodes is simply the euclidean geodesic, and for two points lying far from each other, the probability that the shortest path passes near the barycenter of the network is very large. In other words, the barycenter (and its neighbors) will have a large centrality as illustrated in Fig. 1a. In contrast, in a purely topological network with no underlying geography, this consideration does not apply anymore, and if we rewire more and more links (as illustrated in Fig. 1b), we observe a progressive decorrelation of centrality and space while the correlation with degree increases. In a lattice, it is easy to show that the BC depends on space and is maximum at the barycenter, while in a network the BC of a node depends on its degree. When the network is constituted of long links superimposed on a lattice, we then expect the appearance of "anomalies" characterized by large deviations around the behavior $g \sim k^{\eta}$.



Spatial Networks, Fig. 1 (**a**) Betweenness centrality for the (one-dimensional) lattice case. The central nodes are close to the barycenter. (**b**) For a general graph, the central nodes are usually the ones with large degree

Betweenness Centrality Impact

When studying the time evolution of networks, it is important to be able to characterize quantitatively new links. This is particularly true for spatial networks, but what follows could also be applied to general, complex networks.

We consider a time-evolving graph G_t described by a set of nodes V_t and edges E_t at time t. In order to evaluate the impact of a new link on the overall distribution of the betweenness centrality in the graph at time t, we first compute the average betweenness centrality of all the links of G_t as

$$\overline{b}(G_t) = \frac{1}{(N(t)-1)(N(t)-2)} \sum_{e \in E_t} b(e) \quad (6)$$

where $b(e)$ is the betweenness centrality of the edge e in the graph G_t. Then, for each new link e^* added in the time window $[t-1, t]$, we consider the new graph obtained by removing the link e^* from G_t, denoted by $G_t \setminus \{e^*\}$. The impact $\delta_b(e^*)$ of edge e^* on the betweenness centrality of the network at time t is then defined as (Strano et al. 2012)

$$\delta_b(e^*) = \frac{\left[\overline{b}(G_t) - \overline{b}(G_t \setminus \{e^*\})\right]}{\overline{b}(G_t)} \tag{7}$$

The betweenness centrality impact is thus the relative variation of the graph average betweenness due to the removal of the link e^* and can thus help to characterize quantitatively the various mechanisms at play during the evolution of the network (Strano et al. 2012).

Mixing Space and Topology

All the previous indicators describe essentially the topology of the network, but are not specifically designed to characterize spatial networks. We will here briefly review other indicators which provide useful information about the spatial structure of networks. Different indices were defined a long time ago mainly by scientists working in quantitative geography since the 1960s and can be found in Haggett and Chorley (1969) (see also the more recent paper by Xie and Levinson (2007)). Most of these indices are relatively simple but still give important information about the structure of the network in particular if we are interested in planar networks. These indices were used so far to characterize transportation networks such as highways or railway systems.

Alpha and Gamma Indices

The most important indices are called the "alpha" and the "gamma" indices. The simplest index is called the gamma index and is simply defined by

$$\gamma = \frac{E}{E_{\max}} \tag{8}$$

where E is the number of edges and $E_{\max}$ is the maximal number of edges (for a given number of nodes N). For nonplanar networks, $E_{\max}$ is given by $N(N-1)/2$ for nondirected graphs and for planar graphs $E_{\max} = 3N - 6$ leading to

$$\gamma_P = \frac{E}{3N-6} \tag{9}$$

The gamma index is a simple measure of the density of the network, but one can define a similar quantity by counting not the edges but the number of elementary cycles. The number of elementary cycle for a network is known as the cyclomatic number (see, e.g., Clark and Holton 1991) and is equal to

$$\Gamma = E - N + 1 \tag{10}$$

For a planar graph this number is always less or equal to $2N - 5$ which leads naturally to the definition of the alpha index (also coined as meshedness in Buhl et al. 2006)

$$\alpha = \frac{E - N + 1}{2N - 5} \tag{11}$$

This index belongs to $[0, 1]$ and is equal to 0 for a tree and equal to 1 for a maximal planar graph.

Cell Area and Shape

For planar spatial networks, we have faces or cells which have a certain area and shape. In certain conditions, it can be interesting to characterize statistically these shapes, and various indicators were developed in this perspective (see Haggett and Chorley 1969 for a list of these indicators).

The first, simple important information is the distribution of the area $P(A)$ which for many cases follows a power law (Lammer et al. 2006; Barthelemy and Flammini 2008):

$$P(A) \sim A^{-\tau} \tag{12}$$

where $\tau \approx 2$. We can note here that a simple argument on node density fluctuation leads indeed to this value $\tau = 2$ and further empirical analysis is needed to test the universality of this result.

In addition to the area of the cell, its shape distribution is also interesting and contains a large part of the information about the structure of the network. A simple way to characterize the shape is given by the form factor ϕ. If we denote by L the major axis, the shape ratio is defined as A/L^2 (or equivalently, we can define the elongation ratio $\sqrt{A}/L$). In the paper (Lammer et al. 2006) on the road network structure, Lämmer et al. use another definition of the form factor and define it as

$$\phi = \frac{4A}{\pi D^2} \tag{13}$$

where πD^2 is the area of the circumscribed circle. If this ratio is small, the cell is very anisotropic, while on the contrary if ϕ is closer to one, the corresponding cell is almost circular. In many cases where rectangles and squares predominate (Lammer et al. 2006; Strano et al. 2012), we have $\phi \approx 0.5 - 0.6$.

Detour Index

When the network is embedded in a two-dimensional space, we can define at least two distances between the pairs of nodes. There is of course the natural euclidean distance $d_E(i, j)$ which can also be seen as the "as crow flies" distance. There is also the total "route" distance $d_R(i, j)$ from i to j by computing the sum of lengths of segments belonging to the shortest path between i and j. The detour index – also called the route factor – for this pair of nodes (i, j) is then given by (see Fig. 2 for an example)

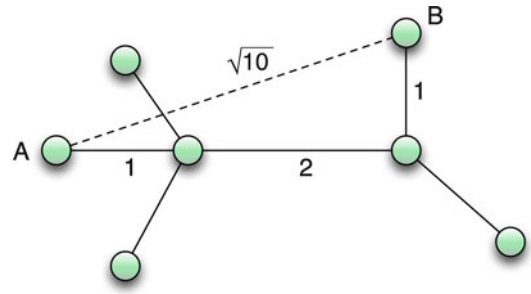
$$Q(i, j) = \frac{d_R(i, j)}{d_E(i, j)} \tag{14}$$

This ratio is always larger than one, and the closer to one, the more efficient the network. From this quantity, we can derive another one for a single node defined by

$$\langle Q(i) \rangle = \frac{1}{N-1} \sum_j Q(i, j) \tag{15}$$

which measures the "accessibility" for this specific node i. Indeed the smaller it is, the easier it is to reach the node i. This quantity is related to the quantity called "straightness centrality" (Crucitti et al. 2006):

$$C^S(i) = \frac{1}{N-1} \sum_{j \neq i} \frac{d_E(i, j)}{d_R(i, j)} \tag{16}$$



Spatial Networks, Fig. 2 Example of detour index calculation. The "as crow flies" distance between the nodes A and B is $d_E(A, B) = \sqrt{10}$, while the route distance over the network is $d_R(A, B) = 4$ leading to a detour index equal to $Q(A, B) = 4/\sqrt{10} \simeq 1.265$

And if one is interested in assessing the global efficiency of the network, one can compute the average over all pairs of nodes:

$$\langle Q \rangle = \frac{1}{N(N-1)} \sum_{i \neq j} Q(i, j) \tag{17}$$

The average $\langle Q \rangle$ or the maximum $Q_{\max}$, and more generally the statistics of $Q(i, j)$, is important and contains a lot of information about the spatial network under consideration (see Aldous and Shun 2010 for a discussion on this quantity for various networks). For example, one can define the interesting quantity Aldous and Shun (2010)

$$\rho(d) = \frac{1}{N_d} \sum_{ij/d_E(i,j)=d} Q(i, j) \tag{18}$$

(where N_d is the number of nodes such that $d_E(i, j) = d$) whose shape can help in characterizing combined spatial and topological properties.

Cost and Efficiency

The minimum number of links to connect N nodes is $E = N - 1$ and the corresponding network is then a tree. We can also look for the tree which minimizes the total length given by the sum of the lengths of all links:

$$\ell_T = \sum_{e \in E} d_E(e) \tag{19}$$

where $d_E(e)$ denotes the length of the link e. This procedure leads to the minimum spanning tree (MST) which has a total length ℓ_T^{MST} (see, e.g., Clark and Holton 1991). Obviously the tree is not a very efficient network (e.g., from the point of view of transportation), and usually more edges are added to the network, leading to an increase of accessibility but also of ℓ_T. A natural measure of the "cost" of the network is then given by

$$C = \frac{\ell_T}{\ell_T^{\mathrm{MST}}} \tag{20}$$

We note here that we easily estimate the total length if the segment length distribution is peaked around its average ℓ_1, and if the node distribution is uniform, $\ell_1 \sim 1/\sqrt{\rho}$ where $\rho = N/A$ is the average node density (A is the area of the system). In this case, the total length is given by $\ell_T = E\ell_1$ leading to

$$\ell_T = \frac{\langle k \rangle}{2}\sqrt{AN} \tag{21}$$

where $\langle k \rangle$ is the average degree of the graph. Adding links thus increases the cost but improves accessibility or the *transport performance* P of the network which can be measured as the minimum distance between all pairs of nodes, normalized by the same quantity computed for the minimum spanning tree:

$$P = \frac{\langle \ell \rangle}{\langle \ell_{\mathrm{MST}} \rangle} \tag{22}$$

Another measure of efficiency was also proposed in Latora and Marchiori (2001) and is defined as

$$E = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{1}{\ell(i,j)} \tag{23}$$

where $\ell(i,j)$ is the shortest path distance from i to j. Combination of these different indicators and comparisons with the MST or the maximal planar network can be constructed in order to characterize various aspects of the networks under consideration (see, e.g., Buhl et al. 2006).

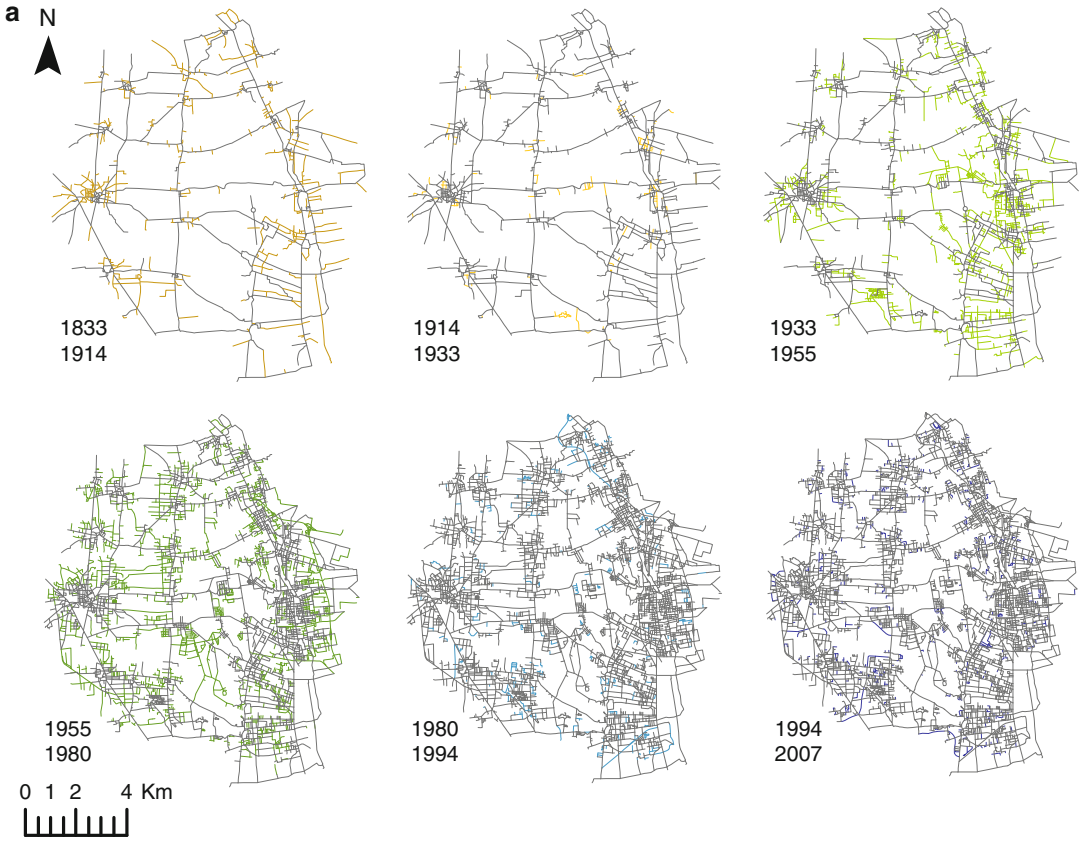
Finally, adding links improves the resilience of the network to attacks or dysfunctions. A way to quantify this is by using *fault tolerance (FT)* (see, e.g., Tero et al. 2010) measured as the probability of disconnecting parts of the network with the failure of a single link. The benefit/cost ratio could then be estimated by the quantity FT/ℓ_T^{MST} which is a quantitative characterization of the trade-off between cost and efficiency (Tero et al. 2010).

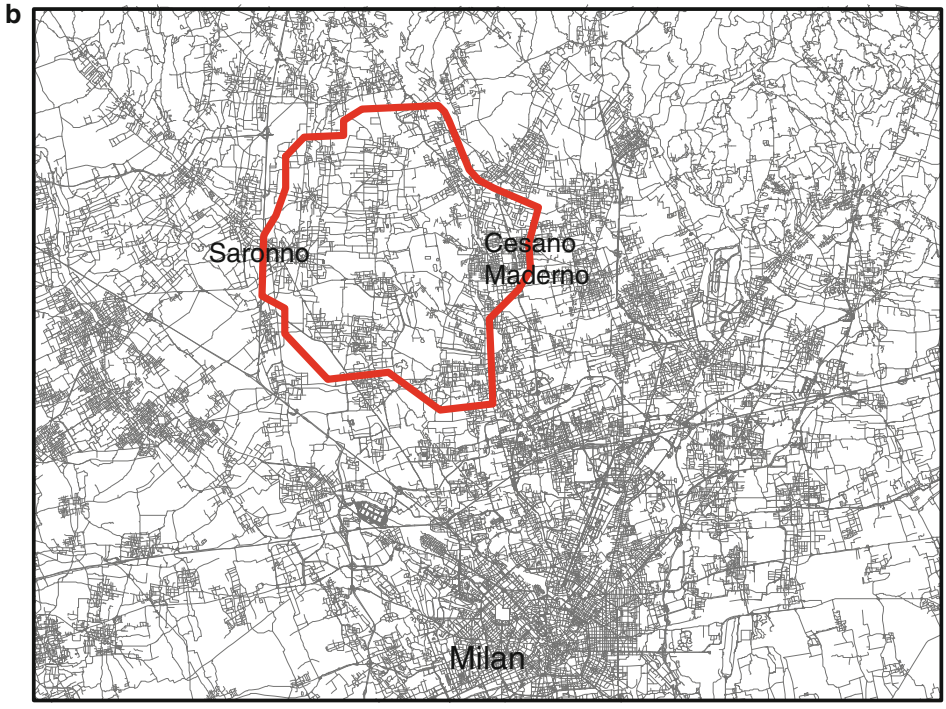
Future Directions

In this final section, we discuss briefly two directions for future research which seem very promising. Both directions come from the fact that ever more data are available, opening the path for new measures, new models, and new understanding of the formation and evolution of spatial networks.

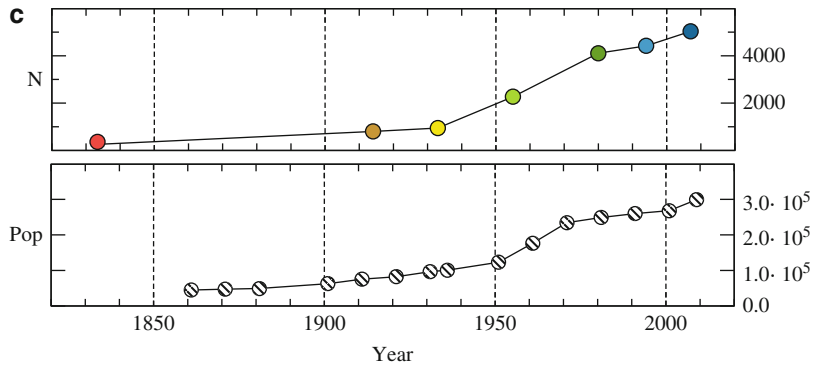
Measuring and Modeling the Time Evolution of Spatial Networks

Thanks to the efforts of GIS scientists (Batty 2005), we now have digitalized maps, combined with data from remote sensing, which allows for studying the time evolution of spatial networks such as roads and streets over long periods. Understanding the evolution of transportation networks (Xie and Levinson 2009) is important from a fundamental point of view but also sheds some light on the crucial problem of understanding the time evolution of a city. Recent studies (Xie and Levinson 2009; Strano et al. 2012; Barthelemy et al. 2013) started to quantify the evolution of spatial networks, and more empirical results are certainly to come (Fig. 3).





Spatial Networks, Fig. 3 (continued)



Spatial Networks, Fig. 3 (**a**) Evolution of the road network from 1833 to 2007 (for each map we show in *grey* all the nodes and links already existing in the previous snapshot of the network and in colors the new links added in the time window under consideration). (**b**) Map showing the location of the studied area (Groane area in the metropolitan region of Milan). (**c**) Time evolution of the total number of nodes N in the network and of the total population in the area (obtained from census data) (Figure taken from Strano et al. 2012)

At the time this article is written, we are still in the process of collecting data, processing them, and extracting stylized facts. The next important step will be the modeling of the evolution of these systems. There are already some simplified models, but we will now be able to confront theoretical models with stylized fact and hopefully converge to simple realistic models of spatial network evolution. In particular, all these studies will have to address the issue of self-organization versus centralized planning for different time scales, a crucial problem in the modeling of urban systems.

Connecting Spatial Networks with Socioeconomical Indicators

Revealing the relationships of network topology to socioeconomical features is not a new project. There is indeed a wealth of papers in quantitative geography of the 1960s–1970s (see, e.g., Haggett and Chorley 1969; Radke 1977 and references therein). In 1969, for example, (Kissling 1969) concludes that the analysis of the network structure is "likely to reveal probable growth points in the system." However, the recent availability of spatial data on networks and on socioeconomical indicators reinvigorates this direction of research. This can even be done at various scales. At large scales, for example, one can try to understand the relation between population, activity densities, and the structure of transportation networks. At a smaller scale, one can try to understand crime rates and activity density fluctuation in terms of topological properties of the transportation network.

This problem will also require a lot of efforts from the modeling side. In particular, we know that there is strong coupling between the population density and the network structure, but we still need a modeling framework for describing such a coupling and coevolution. From a longer time scale perspective, these studies on spatial networks belong to the more general problem of understanding the time evolution of a city. So far, modeling a city has mostly been done in the field of spatial economics (Fujita et al. 1999). However most of these studies consider monocentric structures and static properties, and their predictions are not compared with empirical data. Gathering various data, proposing simple dynamical models integrating the most relevant economical ingredients, and confronting their prediction to data will certainly lead in some future to a wealth of new and original results about this very complex system that is a city.

Cross-References

▸ Analysis and Planning of Urban Networks
▸ Community Identification in Dynamic and Complex Networks
▸ Location-Based Social Networks

- ▸ Networks in Geography
- ▸ Spatial Statistics
- ▸ Temporal Networks
- ▸ Tools for Networks

References

Albert R, Barabasi AL (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47

Aldous DJ, Shun J (2010) Connected spatial networks over random points and a route-length statistic. Stat Sci 25:275–288

Barrat A, Barthelemy M, Pastor-Satorras R, Vespignani A (2004) The architecture of complex weighted networks. Proc Natl Acad Sci USA 101:3747

Barthelemy M (2011) Spatial networks. Phys Rep 499:1

Barthelemy M, Flammini A (2008) Modelling urban street patterns. Phys Rev Lett 100:138702

Barthelemy M, Bordin P, Berestycki H, Gribaudi M (2013) Self-organization versus top-down planning in the evolution of a city. Nat Sci Rep 3:2153

Batty M (2005) Network geography: relations, interactions, scaling and spatial processes in GIS. In: Fisher PF, Unwin DJ (eds) Re-presenting GIS. Wiley, Chichester, pp 149–170

Buhl J, Gautrais J, Reeves N, Solé RV, Valverde S, Kuntz P, Theraulaz G (2006) Topological patterns in street networks of self-organized urban settlements. Eur Phys J B-Condens Matter Complex Syst 49(4):513–522

Bullmore E, Sporns O (2009) Complex brain networks: graph theoretical analysis of structural and functional systems. Nat Rev Neurosci 10(3):186–198

Clark J, Holton DA (1991) *A first look at graph theory* (vol. 6). Teaneck, NJ: World Scientific.

Crucitti P, Latora V, Porta S (2006) Centrality in networks of urban streets. Chaos Interdiscip J Nonlinear Sci 16(1):015113–015113

Freeman LC (1977) A set of measures of centrality based on betweenness. *Sociometry*, 35–41

Fujita M, Krugman PR, Venables AJ (1999) The spatial economy: cities, regions and international trade, vol 213. MIT, Cambridge

Haggett P, Chorley RJ (1969) Network analysis in geography. Edward Arnold, London

Kissling CC (1969) Linkage importance in a regional highway network. Can Geogr 13:113–129

Lammer S, Gehlsen B, Helbing D (2006) Scaling laws in the spatial structure of urban road networks. Phys A Stat Mech Appl 363(1):89–95

Latora V, Marchiori M (2001) Efficient behavior of small-world networks. Phys Rev Lett 87:198701

Liben-Nowell D, Novak J, Kumar R, Raghavan P, Tomkins A (2005) Geographic routing in social networks. Proc Natl Acad Sci USA 102:11623–11628

Radke JD (1977) Stochastic models in circuit network growth. Thesis and dissertations (Comprehensive). Paper 1450, Wilfrid Laurier University

Strano E, Nicosia V, Latora V, Porta S, Barthelemy M (2012) Elementary processes governing the evolution of road networks. Nat Sci Rep 2:296

Tero A, Takagi S, Saigusa T, Ito K, Bebber DP, Fricker MD, Yumiki K, Kobayashi R, Nakagaki T (2010) Rules for biologically inspired adaptive network design. Sci Signal 327:439

Watts D, Strogatz S (1998) Collective dynamics of small-world networks. Nature 393:440–442

Xie F, Levinson D (2007) Measuring the structure of road networks. Geograph Anal 39:336–356

Xie F, Levinson D (2009) Topological evolution of surface transportation networks. Comput Environ Urban Syst 33:211–223

Spatial Scan Statistic

- ▸ Disease Surveillance, Case Study

Spatial Statistics

Victor De Oliveira[1] and A. Alexandre Trindade[2]
[1]Department of Management Science and Statistics, The University of Texas at San Antonio, San Antonio, TX, USA
[2]Department of Mathematics & Statistics, Texas Tech University, Lubbock, TX, USA

Synonyms

Geocomputation; Geostatistics; Spatial analysis

Glossary

Correlation/Covariance Measures of similarity between observations

Geostatistics A branch of spatial statistics

Isotropy Property of covariance and variogram functions that make them is invariant under rotation of locations

Kriging Method for linear unbiased prediction

Random Field A collection of random variables indexed by location

Stationarity Property of random fields in which their mean and covariance functions are invariant under translation of locations

Variogram/Semivariogram Measures of dissimilarity between observations

Definition

Spatial statistics is a branch of statistics that studies methods to make inference based on data observed over spatial regions. In typical applications these regions are either 2- or 3-dimensional. The methodology is mostly aimed at accounting and modeling aspects of the so-called First Law of Geography: attributes from locations that are closer together are more closely related than attributes from locations that are farther apart. This is accomplished through appropriate measures of *spatial association*. An overview of models and methods is given for the three main types of spatial data: *geostatistical, lattice*, and *point pattern*.

Introduction

Spatial data refer to measurements of phenomena that vary over a region of space $D \subset \mathbb{R}^d$, $d \geq 1$, which would be called the *region of interest*. Each datum is associated to a subset of D that indicates *where* it was collected, often called the datum's *support*. This may be a single point or a larger subset, depending on the context.

There are three basic types of spatial data: geostatistical (or point referenced), lattice (or areal), and point pattern. The three types may be viewed as pairs $\{(s_i, z_i) : i = 1, \ldots, n\}$ where the interpretation and characteristics of the data components vary from type to type. For geostatistical data $z_1, \ldots, z_n$ are measurements or observations of a phenomenon of interest taken at sampling locations $s_1, \ldots, s_n \in D$, which are single *points*. In the models to be described later, the z_is are random, while n (the sample size) and the s_is are known and fixed. For lattice data $s_1, \ldots, s_n$ are *subregions* that form a partition of D, such as counties or postal codes, and $z_1, \ldots, z_n$ are averages or summaries of the phenomenon of interest over these subregions. For this type of data, it also holds that the z_is are random, while n and the s_is are known and fixed. For point pattern data $s_1, \ldots, s_n$ are points where a certain *event* of interest occurs, such as the presence of a type of tree or the epicenter of an earthquake, and $z_1, \ldots, z_n$ are a feature of the aforementioned events, such as the diameter of the tree at breast height or the magnitude of the earthquake. In the models for point pattern data to be described later, all components n, s_i and, z_i are random. Often the z_is are absent when interest centers only on the pattern of occurrences. For all three types of spatial data, additional variables could also be available, that serve as explanatory variables. Comprehensive treatments of statistical models and methods for all three types of data appear in Cressie (1993), Schabenberger and Gotway (2005), and the recent edited volume by Gelfand et al. (2010). Table 1 summarizes the key concepts and gives an overview of models and examples.

Key Points

Random Fields

A random field $\{Z(s) : s \in D\}$ on the region $D \subset \mathbb{R}^d$ is a collection of random variables indexed by the elements of D, where D can be finite or infinite. These random variables are often nonidentically distributed and *dependent*, so modeling these aspects is a key starting point. The simplest way to do this is through the mean and covariance functions of the random field, defined as

$$\mu(s) := \mathbb{E}\{Z(s)\} \quad \text{and}$$
$$C(s,u) := \text{cov}\{Z(s), Z(u)\}, \quad s, u \in D.$$

The former determines the *spatial trend*, a measure of variation over large distances, while the latter determines the *spatial association*, a measure of variation over small distances.

Spatial Statistics, Table 1 Summary and overview of concepts, models, and examples in the three types of spatial data

	Geostatistical	Lattice	Point pattern
Domain D	Fixed, continuous	Fixed, discrete	Random, continuous
Observation sites $\{s_i : i = 1, \ldots, n\}$	s_i fixed n fixed	s_i fixed n fixed	s_i random n random
Inference for	$Z(s)$ only	$Z(s)$ only	Both $Z(s)$ and D
Main models for $Z(s)$	Sum of regression trend and stationary random field	Simultaneous Autoregressive Model (SAR)	Poisson process (homogeneous and inhomogeneous)
Key aims, concepts	Kriging (minimum MSE prediction)	Spatial proximity matrix (W)	Assess tendency for clustering
Examples	Meteorological and geological variables	Geographic and demographic variables	Location and intensity of events

Other features of a random field also related with spatial association are the correlation function and semivariogram function, defined respectively as

$$K(s,u) := \text{corr}\{Z(s), Z(u)\} = \frac{C(s,u)}{\sigma(s)\sigma(u)}$$

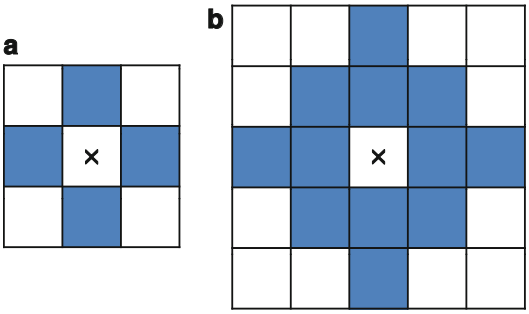
$$\gamma(s,u) := \frac{1}{2}\text{var}\{Z(s) - Z(u)\} = \frac{1}{2}\big(\sigma^2(s) + \sigma^2(u) - 2C(s,u)\big),$$

where $\sigma^2(s) := \text{var}\{Z(s)\}$ is the variance function. The functions $C(s,u)$ and $\gamma(s,u)$ provide similar information about the spatial association of the random field, with the former being a measure of *similarity* between $Z(s)$ and $Z(u)$, while the latter is a measure of *dissimilarity*. When choosing the aforementioned functions, it is important to note that *any* function can be used as a mean function, but *not* any function can be used as a covariance function. The latter needs to be *positive semi-definite*, meaning that for any $m \in \mathbb{N}$, $s_1, \ldots, s_m \in D$ and $a_1, \ldots, a_m \in \mathbb{R}$ it holds that

$$\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j C(s_i, s_j) \geq 0.$$

This is a difficult condition to verify, but fortunately the literature provides many functions known to be positive semi-definite; see Cressie (1993) and Chilès and Delfiner (1999) for examples. These references also provide an intermediate treatment on the theory and methods of random fields and their application to spatial statistics, while Matérn (1986), Yaglom (1987), and Stein (1999) provide more mathematical treatments.

Lattice data usually represent averages or summaries of a quantity of interest over subregions, so covariance and semivariogram functions are not the most suitable to quantify spatial association among this type of data. Instead, neighborhood relations and weight matrices are used. In this case the collection of subregions $\{s_1, \ldots, s_n\}$ is endowed with a *neighborhood system* $\{N_i : i = 1, \ldots, n\}$, where N_i denotes the subregions that are, in a precisely defined way, neighbors of subregion s_i. For rectangular regular lattices where the subregions may be thought of as pixels, it is common to use *first-order* neighborhood systems, where the neighbors of a pixel are the pixels adjacent to the north, south, east, and west; see Fig. 1a. *Second-order* neighborhood systems are also used, where the neighbors of a pixel are its first-order neighbors and their first-order neighbors; see Fig. 1b. In these cases all pixels have the same number of neighbors, except for pixels at (or near) the boundary of D. For regions divided in unequally shaped subregions (like counties in a state), a commonly used neighborhood system



Spatial Statistics, Fig. 1 Examples of first-order (**a**) and second-order (**b**) neighborhood systems. Pixels in *blue* are the neighbors of the pixel marked with an "x"

is defined in terms of geographic adjacency, $N_i = \{s_j : \text{subregions } s_i \text{ and } s_j \text{ share a boundary}\}$; other examples not based on geographic adjacency are also possible. In these cases the number of neighbors for each subregion usually differs.

In addition a weight (or neighborhood) matrix $W = (w_{ij})$ is specified, where w_{ij} measures the strength of direct association between sites s_i and s_j. It must satisfy that $w_{ij} \geq 0$, $w_{ii} = 0$ and $w_{ij} > 0$ if and only if s_i and s_j are neighbors (i.e., $s_j \in N_i$). The most common example of weight matrix is $w_{ij} = 1$ if s_i and s_j are neighbors and $w_{ij} = 0$ otherwise, but other more refined specifications are also possible, e.g., based on distance between subregions' centroids. Anselin (1988), Cressie (1993), Rue and Held (2005), and LeSage and Pace (2009) provide ample treatments of models and methods for the analysis of lattice data.

For geostatistical and lattice data, the sampling locations $s_1, \ldots, s_n$ are fixed and known, so these types of data are usually written as $\mathbf{z} = (z_1, \ldots, z_n)^\intercal$ ($^\intercal$ denotes transpose of a vector or matrix). The stochastic approach for modeling and inference assumes the data are a part of a *realization* of a random field $Z(\cdot)$, so datum z_i is the realized value of the random variable $Z(s_i)$.

Stationarity and Ergodicity

Spatial data typically contain no replicates as usually a single observation is available at each location, so some assumptions on the random field are needed to make statistical inference feasible. To illustrate this point consider the conceptual decomposition $Z(s_i) = \mu(s_i) + \varepsilon(s_i)$, with $\varepsilon(\cdot)$ a random field with mean zero and covariance function $C(s, u)$. Without some extra assumptions it is not possible to identify both $\mu(s_i)$ and $\varepsilon(s_i)$ with a single observation at s_i. This is so because a term can be added to $\mu(s_i)$ and subtracted from $\varepsilon(s_i)$ in infinitely many ways, any of which will not change the datum $Z(s_i)$ but will change the components that seek to be identified.

The assumptions alluded above are those of *stationarity* and *ergodicity*. A random field $Z(s)$ is said to be (second-order or weakly) stationary if

$$\mu(s) = \mu \text{ (constant) and}$$
$$C(s, u) = \tilde{C}(s - u), \quad s, u \in D,$$

where $\tilde{C}(\cdot)$ is a function of a single spatial variable. The above means that the mean and covariance functions are invariant under translations of the spatial locations. From these follow that the variance, correlation, and semivariogram functions are also invariant under translations of the spatial locations, and we have

$$\sigma^2(s) = \sigma^2, \; C(s, u) = \sigma^2 \tilde{K}(s - u),$$
$$\gamma(s, u) = \sigma^2\big(1 - \tilde{K}(s - u)\big).$$

An important and commonly used special case of stationarity is called *isotropy*, meaning that $C(s, u) = \bar{C}(\|s - u\|)$, where $\|h\| := (h_1^2 + \cdots + h_d^2)^{1/2}$ is the Euclidean norm of $h \in \mathbb{R}^d$ and $\bar{C}(\cdot)$ is a function of a single real variable. In this case the covariance function is also invariant under rotations of the spatial locations, so the nature of spatial association is the same in all directions; see Ripley (1981), Cressie (1993), and Schabenberger and Gotway (2005) for further discussion on stationarity.

A precise definition of ergodicity is somewhat technical (see Cressie 1993, pp. 53–58), but this assumption is key to make statistical inference based on spatial data feasible. This is so because the meaning and interpretation of many features of a random field, such as the mean function, are based on *ensemble* (i.e., population) averages,

namely, averages over the possible realizations of the random field. Ergodicity requires that spatial averages computed from a single realization converge to their respective ensemble averages as the sample size increases to infinity.

A complete description of a random field requires specifying its family of finite-dimensional distributions, namely, the family of joint distributions

$$F_{s_1,\ldots,s_m}(x_1,\ldots,x_m) = P\{Z(s_1) \le x_1,\ldots,Z(s_m) \le x_m\},$$

$\forall\ m \in \mathbb{N}$ and $s_1,\ldots,s_m \in D$. The simplest and most commonly used of such specification is that of *Gaussian* random fields, meaning that all the aforementioned distributions are multivariate normal. Gaussian random fields are completely specified by their mean and covariance functions, and when they are stationary, a sufficient condition for them to be ergodic is that $\lim_{\|h\|\to\infty} \tilde{C}(h) = 0$. Gaussian random fields are the most commonly used models because of their convenient mathematical properties and wide applicability, as well as their use as "building blocks" for more complex random fields models. Examples of the latter are hierarchical models used to describe discrete spatial data; see Banerjee et al. (2004) and Diggle and Ribeiro (2007).

Models and Inference

Geostatistical Data Models

The basic geostatistical model is based on the conceptual decomposition of the random field of interest as

$$Z(s) = \mu(s) + \varepsilon(s), \quad s \in D,$$

where $\mu(s)$ is the mean function (spatial trend) and $\varepsilon(\cdot)$ is a zero-mean random field that describes the short-range variation, with the same covariance function as $Z(\cdot)$. The usual model for the spatial trend is similar to that used in linear regression models

$$\mu(s) = \sum_{j=1}^{p} f_j(s)\beta_j = \boldsymbol{f}(s)^{\intercal}\boldsymbol{\beta},$$

where $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)^{\intercal}$ are unknown regression parameters and $\boldsymbol{f}(s) = (f_1(s),\ldots,f_p(s))^{\intercal}$ are known location-dependent covariates. The latter may include related spatially varying processes. For instance, if $Z(s)$ = rainfall amount that fell over a period of time at locations s, then $f(s)$ = altitude at location s may be a useful explanatory variable. More often a spatial trend is described in terms of a polynomial in the spatial coordinates. For the case when $d = 2$ and $s = (x, y)$, this would be

$$\mu(s) = \sum_{0\le i+j\le p} \beta_{ij} x^i y^j, \text{ for some } p \ge 1 \text{ known.}$$

Many examples of stationary covariance models have been proposed in the literature (see Cressie 1993; Chilès and Delfiner 1999). An example of a flexible family of isotropic covariance functions is the so-called *Matérn* family (Matérn 1986; Stein 1999)

$$\tilde{C}(t) = \frac{2\sigma^2}{\Gamma(\nu)}\left(\frac{t}{2\phi}\right)^{\nu} \mathcal{K}_{\nu}\left(\frac{t}{\phi}\right), \quad t \ge 0,$$

where $\Gamma(\cdot)$ is the gamma function and $\mathcal{K}_{\nu}(\cdot)$ is the modified Bessel function of the second kind and order ν. For such model $\phi > 0$ (mainly) controls how fast the correlation decreases with distance, and $\nu > 0$ controls the smoothness of the realizations of the random field. The commonly used exponential and Gaussian covariance functions are special cases obtained, respectively, by setting $\nu = 1/2$ and $\nu \to \infty$.

The above description assumes the process of interest is measured exactly (or nearly so), but more often the data contain measurement error; see Le and Zidek (2006) for an extensive discussion. In this case the simplest model for the observed data is

$$Z_{i,\text{obs}} = Z(s_i) + \epsilon_i, \quad i = 1,\ldots,n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are assumed i.i.d with mean 0, variance $\tau^2 > 0$ and independent of $Z(\cdot)$. Under the above model the data $\mathbf{Z}_{\text{obs}} = (Z_{1,\text{obs}}, \ldots, Z_{n,\text{obs}})^\intercal$ follow the general linear model

$$\mathbf{Z}_{\text{obs}} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where X is the n by p matrix with entries $(X)_{ij} = f_j(s_i)$ and $\boldsymbol{\varepsilon}$ is a random vector with $\mathbb{E}\{\boldsymbol{\varepsilon}\} = \mathbf{0}$ and $\text{var}\{\boldsymbol{\varepsilon}\} = \text{var}\{\mathbf{Z}_{\text{obs}}\} = \Sigma_{\boldsymbol{\theta}}$, with the n by n matrix $\Sigma_{\boldsymbol{\theta}}$ having entries $(\Sigma_{\boldsymbol{\theta}})_{ij} = \sigma^2(2/\tau(\gamma))(t_{ij}/2\phi)^\gamma \mathcal{K}_\gamma(t_{ij}/\phi)$ and $t_{ij} = \|s_i - s_j\|$; $1(A)$ denotes the indicator function of A. This basic specification of a geostatistical model depends on unknown regression parameters $\boldsymbol{\beta}$ and covariance parameters $\boldsymbol{\theta} = (\sigma^2, \phi, \nu, \tau^2)$.

Parameter Estimation

The classical geostatistical method of estimation uses a distribution-free approach (Journel and Huijbregts 1978; Cressie 1993; Chilès and Delfiner 1999). First, the regression parameters are estimated by least squares, resulting in

$$\hat{\boldsymbol{\beta}} = (X'QX)^{-1}X^\intercal Q\mathbf{Z}_{\text{obs}},$$

where $Q = I_n$ (ordinary least squares) or $Q = \Sigma_{\boldsymbol{\theta}}^{-1}$ (generalized least squares); the latter requires an estimate of $\Sigma_{\boldsymbol{\theta}}$. In both cases X is assumed to have full rank. The second choice of Q results in a more efficient estimator, but often in practice there is little difference between them. The resulting trend surface estimate is $\hat{\mu}(s) = \boldsymbol{f}(s)^\intercal\hat{\boldsymbol{\beta}}$.

Second, when the mean function is constant, the covariance parameters are estimated by the following two-stage approach: For selected distances $t_1 < \ldots < t_k$, the (model-free) semivariogram estimates are first computed

$$\hat{\gamma}(t_j) = \frac{1}{2|N(t_j)|}\sum_{N(t_j)}\left(z_{i,\text{obs}} - z_{j,\text{obs}}\right)^2,$$

where $N(t) = \{(i,j) : t - \Delta t < \|s_i - s_j\| < t + \Delta t\}$, with $\Delta t > 0$ fixed and $|N(t)|$ the number of elements in $N(t)$. A proposed semivariogram model, say $\gamma(t;\boldsymbol{\theta})$, is then fitted to the above semivariogram estimates $\hat{\gamma}(t_1), \ldots, \hat{\gamma}(t_k)$ using (nonlinear) least squares, so the covariance parameter estimates are

$$\hat{\boldsymbol{\theta}} = \arg\min\sum_{j=1}^{k}\left(\hat{\gamma}(t_j) - \gamma(t_j;\boldsymbol{\theta})\right)^2.$$

The resulting semivariogram function estimate is $\gamma(\cdot\,;\hat{\boldsymbol{\theta}})$. When $\mu(s)$ is not constant a similar procedure is done using the residuals $\mathbf{e} = \mathbf{z}_{\text{obs}} - X\hat{\boldsymbol{\beta}}$, rather than the observed data. This estimation method is popular among practitioners, but the statistical properties of the resulting estimators are not well understood.

When the random field $Z(\cdot)$ is Gaussian, all the parameters can be jointly estimated by maximum likelihood (Cressie 1993; Stein 1999), resulting in the estimators

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) = \arg\max L(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{z}_{\text{obs}}), \qquad (1)$$

where

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}; \mathbf{z}_{\text{obs}}) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}|\Sigma_{\boldsymbol{\theta}}|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{z}_{\text{obs}} - X\boldsymbol{\beta})^\intercal\Sigma_{\boldsymbol{\theta}}^{-1}(\mathbf{z}_{\text{obs}} - X\boldsymbol{\beta})\right\}. \qquad (2)$$

This method is more statistically satisfactory than the two-stage approach described above but is also more computationally demanding, to the point of not being feasible for very large datasets (n very large) due the need of storing and numerically inverting the n by n matrix $\Sigma_{\boldsymbol{\theta}}$; see Cressie (1993), Schabenberger and Gotway (2005), and Chap. 4 in Gelfand et al. (2010) for other methods of estimation.

Spatial Prediction (Kriging)

The primary task in the analysis of geostatistical data is often spatial prediction, also known as *kriging*, which consists of making inference about $Z(s_0)$ where $s_0 \in D$ is an unsampled location. The classical approach uses optimal linear unbiased prediction and only requires knowledge of the mean and covariance

(or semivariogram) functions. Specifically, the method seeks to minimize the mean squared prediction error

$$\mathrm{MSPE}(\hat{Z}(s_0)) = \mathbb{E}\{\big(Z(s_0) - \hat{Z}(s_0)\big)^2\},$$

over the class of linear unbiased predictors, that is, predictors of the form $\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i(s_0) z_{i,\mathrm{obs}}$ that satisfy $\mathbb{E}\{\hat{Z}(s_0)\} = \mathbb{E}\{Z(s_0)\}$. Under the aforementioned linear model, the optimal coefficients are obtained as the solution of a linear system of equations, and the resulting optimal predictor is

$$\hat{Z}^K(s_0) = \Big(\boldsymbol{\sigma}_0 + X(X^\intercal \Sigma_{\boldsymbol{\theta}}^{-1} X)^{-1}(\boldsymbol{f}(s_0) - X^\intercal \Sigma_{\boldsymbol{\theta}}^{-1} \boldsymbol{\sigma}_0)\Big)^\intercal \mathbf{Z}_{\mathrm{obs}},$$

where $\boldsymbol{\sigma}_0 = \mathrm{cov}\{\mathbf{Z}_{\mathrm{obs}}, Z(s_0)\}$; this is called the *best linear unbiased predictor* (BLUP) or kriging predictor of $Z(s_0)$. The usual uncertainty measure associated with the kriging predictor is $\mathrm{MSPE}(\hat{Z}^K(s_0))$, which is given by

$$\sigma^{2K}(s_0) = C(\mathbf{0}) - \boldsymbol{\sigma}_0^\intercal \Sigma_{\boldsymbol{\theta}}^{-1} \boldsymbol{\sigma}_0 + (\boldsymbol{f}(s_0) - X^\intercal \Sigma_{\boldsymbol{\theta}}^{-1} \boldsymbol{\sigma}_0)^\intercal (X^\intercal \Sigma_{\boldsymbol{\theta}}^{-1} X)^{-1} (\boldsymbol{f}(s_0) - X^\intercal \Sigma_{\boldsymbol{\theta}}^{-1} \boldsymbol{\sigma}_0).$$

When the random field $Z(\cdot)$ is Gaussian, then $\hat{Z}^K(s_0)$ is also the best unbiased predictor (it minimizes $\mathrm{MSPE}(\cdot)$ over the class of *all* unbiased predictors), and a 95 % prediction interval for $Z(s_0)$ is $\hat{Z}^K(s_0) \pm 1.96\sigma^K(s_0)$; see Cressie (1993), Chilès and Delfiner (1999), and Schabenberger and Gotway (2005) for methodological details and Stein (1999) for theoretical underpinnings.

The computation of kriging predictors and the validity of their optimality properties require the covariance parameters $\boldsymbol{\theta}$ to be known, which is certainly not the case in practice. The simplest and most commonly used practical solution is to use empirical or plug-in predictors and mean squared prediction errors obtained by replacing in the above formulas unknown covariance parameters with their estimates. But the properties of the resulting plug-in predictors and mean squared prediction errors differ from those of their known covariance parameters counterparts since the former do not take into account the sampling variability of parameter estimators. As a result plug-in mean square prediction errors tend to underestimate the true mean square prediction errors of plug-in predictors, and the true coverage probability of plug-in prediction intervals tends to be smaller than nominal. Possible approaches to account for parameter uncertainty when performing predictive inference include using *bootstrap* (Sjöstedt-De Luna and Young 2003) and the Bayesian approach (Banerjee et al. 2004; Diggle and Ribeiro 2007), where the latter approach appears to be the most effective.

Lattice Data Models

The starting point in the construction of models for lattice data is to empirically assess the existence of spatial association, which as mentioned in a previous section is usually specified in terms of neighborhood systems and weight matrices. The two most common statistics to diagnose spatial association among lattice data are *Moran's I* (an analogue of the lagged autocorrelation used in time series) and *Geary's c* (an analogue of the Durbin-Watson statistic used in time series). For random fields with constant mean, these statistics are defined as

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \big(Z(s_i) - \bar{Z}\big)\big(Z(s_j) - \bar{Z}\big)}{S_0 \sum_{i=1}^{n} \big(Z(s_i) - \bar{Z}\big)^2}$$

$$c = \frac{(n-1) \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \big(Z(s_i) - Z(s_j)\big)^2}{2 S_0 \sum_{i=1}^{n} \big(Z(s_i) - \bar{Z}\big)^2},$$

where $\bar{Z} = \frac{1}{n} \sum_{i=1}^{n} Z(s_i)$ and $S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$. For Gaussian processes, $\mathbb{E}\{I\} = -(n-1)^{-1}$ and $\mathbb{E}\{c\} = 1$ when observations are independent. Hence, observed values of I substantially below/above $-(n-1)^{-1}$ indicate negative/positive spatial association, while for Geary's c the interpretation is reversed, with observed values of c substantially above/-below 1 indicating negative/positive association.

When the random field has a nonconstant mean, the above statistics are computed using residuals; see Cressie (1993) and Cliff and Ord (1981) for further details.

A large number of models for lattice data have been proposed in the literature (see Cressie 1993 and LeSage and Pace 2009), where most of them involve the specification of a neighborhood system $\{N_i\}$ and weight matrix W. One of the most common models for lattice data is the *Simultaneous Auto-regressive* (SAR) model specified by a set of autoregressions

$$Z(s_i) = \boldsymbol{f}(s_i)^\intercal \boldsymbol{\beta} + \rho \sum_{j=1}^{n} w_{ij}\left(Z(s_j) - \boldsymbol{f}(s_j)^\intercal \boldsymbol{\beta}\right) + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\boldsymbol{f}(s_j)$ and $\boldsymbol{\beta}$ have the same interpretation as in models for geostatistical data and $\epsilon_i \sim N(0, \xi_i)$ are independent errors. This is a spatial analogue of autoregressive time series models, but unlike the latter the response and error vectors are correlated. Provided $I_n - \rho W$ is non-singular, it follows that $\mathbf{Z} \sim N_n\left(X\boldsymbol{\beta}, (I_n - \rho W)^{-1} M (I_n - \rho W^\intercal)^{-1}\right)$, where $M = \text{diag}(\xi_1, \ldots, \xi_n)$. It is common to assume $\xi_i = \xi$ for all i, in which case the model parameters are $\boldsymbol{\beta}$ and $\boldsymbol{\theta} = (\xi, \rho)$.

Another large class of models for lattice data is that of *Markov random fields* (Rue and Held 2005; Li 2009). These models construct the joint distribution for the data by specifying the set of all *full conditional* distributions, namely, the conditional distributions of $Z(s_i)$ given $\mathbf{Z}_{(i)}$, $i = 1, \ldots, n$, where $\mathbf{Z}_{(i)} = \left(Z(s_j) : j \neq i\right)$. In addition, these models assume a Markov property stating that the distribution of each datum depends on the rest only through its neighbors. An example of this is the class of *Conditional Auto-regressive* (CAR) models with full conditional distributions

$$\left(Z(s_i) \mid Z(s_j),\ j \neq i\right) \sim N\left(\boldsymbol{f}(s_i)^\intercal \boldsymbol{\beta} + \rho \sum_{j=1}^{n} w_{ij}\left(Z(s_j) - \boldsymbol{f}(s_j)^\intercal \boldsymbol{\beta}\right),\ \sigma_i^2\right), \quad i = 1, \ldots, n.$$

To guarantee the above set of full conditional distributions determines a unique joint distribution, it is required that $\sigma_j^2 w_{ij} = \sigma_i^2 w_{ji}$ for all i, j, and $M^{-1}(I_n - \rho W)$ be positive definite, with $M = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, in which case $\mathbf{Z} \sim N_n\left(X\boldsymbol{\beta}, (I - \rho W)^{-1} M\right)$. It is common to assume that $\sigma_i^2 = \sigma^2$ for all i, in which case the model parameters are $\boldsymbol{\beta}$ and $\boldsymbol{\theta} = (\sigma^2, \rho)$. An extensive comparison between the SAR and CAR models is given in Cressie (1993, Chap. 6).

The most commonly used method for parameter estimation in these models is maximum likelihood. As for geostatistical models, the resulting estimators are given by (Eq. 1) where in the likelihood (Eq. 2) $\Sigma_{\boldsymbol{\theta}}^{-1} = (I_n - \rho W^\intercal) M^{-1} (I_n - \rho W)$ for SAR models and $\Sigma_{\boldsymbol{\theta}}^{-1} = M^{-1}(I - \rho W)$ for CAR models. For both models (as for geostatistical models) the computation of these estimators requires the use of numerical iterative methods.

A point worth noting is that, unlike in geostatistical models, in SAR and CAR models the spatial association structure is specified in terms of the inverse covariance matrix, rather than the covariance matrix, so the interpretation of parameters controlling spatial association is less straightforward than that in geostatistical models.

Point Process Models

A *point process* on $D \subset \mathbb{R}^d$ is a random field whose realizations are sets of points in D, called point patterns (events). In the most general case attributes may also be observed along with the location of the events, resulting in a *marked* point process. For any $A \subset D$, let $N(A)$ denote the number of events in A and $\nu(A)$ the *size* of A $\left(= \int_A ds\right)$. The *intensity* function of a point process is the function $\lambda : D \to [0, \infty)$ with the property that $\mathbb{E}\{N(A)\} = \int_A \lambda(s) ds$. Alternatively, using an "infinitesimal disc" ds centered at s the intensity function can be defined as the ratio of the expected number of points in ds to its size, that is,

$$\lambda(s) = \lim_{\nu(ds) \to 0} \frac{\mathbb{E}\{N(ds)\}}{\nu(ds)}.$$

The most fundamental point process model is the *Poisson* process with intensity function $\lambda(s)$, which satisfies the following: For any $n \in \mathbb{N}$ and $A_1, \ldots, A_n$ disjoint subsets of D, it holds that (i) $N(A_i)$ has Poisson distribution with mean $\int_{A_i} \lambda(s)ds$, and (ii) $N(A_1), \ldots, N(A_n)$ are independent random variables. When the intensity function is constant, $\lambda(s) \equiv \lambda$, the above is called a *homogeneous Poisson process* (HPP), and otherwise it is called an *inhomogeneous Poisson process* (IPP). Point patterns from HPP have the property of *complete spatial randomness* (CSR): given the number of events in a set A, these events are independently and identically distributed over A, so there is no "interaction" between events. Poisson processes are often used on their own for the analysis of point patterns, or as "building blocks" for more complex models; see Diggle (2003) and Illian et al. (2008) for introductory treatments and Cressie (1993) and Daley and Vere-Jones (2003, 2007) for more mathematical treatments.

A basic question in the analysis of point patterns is to assess whether the events have the CSR property. Departures from this comprise either *clustering* (events tend to aggregate) or *regularity* (events tend not to aggregate). The standard model by which to assess the CSR property is the HPP. Testing for CSR is based on either counts of events in regions (quadrants) or distance-based measures using the event locations. Focusing on the former, the distributions of some test statistics are known (usually only asymptotically), which allows for closed-form tests. The default is the chi-square test, whereby the region D is bounded by a rectangle and divided into r rows and c columns. If n_{ij} denotes the number of events in the quadrant corresponding to the i-th row and j-th column, and $\bar{n}$ is the expected number of events in any quadrant, then under CSR the statistic

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \bar{n})^2}{\bar{n}},$$

follows a χ^2_{rc-1} distribution, asymptotically. Tests based on more complex nonstandard statistics can be carried out by resorting to Monte Carlo simulation.

Rejection of CSR may lead one to consider modeling a possibly nonconstant intensity function. This can be done either parametrically, by proposing a specific function for the intensity whose parameters are then estimated via maximum likelihood, or nonparametrically by means of kernel smoothing. For example, under an IPP $\lambda(s)$ can be estimated as a function of coordinates or covariates by fitting a log-linear model of the form

$$\log\big(\lambda(s)\big) = \beta_0 + \sum_{j=1}^{p} \beta_j f_j(s),$$

which provides a way to accommodate departures from CSR based on changes in the mean structure.

Alternatively, rejection of CSR may lead one to consider modeling interactions between events, when for non-overlapping regions A and B, $N(A)$, and $N(B)$ are correlated. The *second-order intensity* function, $\lambda_2(s, u)$, extends the definition of $\lambda(s)$ to measure the covariance between points at s and u, defined as

$$\lambda_2(s, u) = \lim_{\nu(ds),\nu(du)\to 0} \frac{\mathbb{E}\{N(ds)N(du)\}}{\nu(ds)\nu(du)}.$$

For stationary and isotropic processes, where $\lambda(s) \equiv \lambda$ and $\lambda_2(s, u) = \lambda_2(\|s - u\|) \equiv \lambda_2(t)$, the *K-function* is a more informative tool for assessing dependence defined, when $d = 2$, as

$$K(t) = \frac{2\pi}{\lambda^2} \int_0^t x\lambda_2(x)dx.$$

Then, $\lambda K(t)$ represents the expected number of extra events within a distance t from the origin, given that there is an event at the origin. For a HPP one has $K(t) = \pi t^2$; values larger (smaller) than this being indicative of clustering (regularity) on that distance scale. Plotting the estimated $K(t)$ vs. t, or the closely related *L-function*, $L(t) = \sqrt{K(t)/\pi}$, enables one to glean the degree of dependence with reference to the HPP for which $L(t) = t$; see Diggle (2003) and Illian et al. (2008) for further details.

Key Applications

Example 1 As an illustration of a geostatistical dataset Fig. 2a displays pH measurements of wet deposition (acid rain) at 39 rainfall stations taken in April 1987 over the Lower Saxony state in northwest Germany (Berke 1999). Each datum is associated with the sampling location where the pH measurement was taken. For instance, a pH value of 4.63 was observed at the sampling location $s = (0.61, 0.1)$ (the southernmost station). For this dataset the coordinates of the sampling locations were provided without units and are all between 0 and 1, which (presumably) mean they were scaled by the maximum distance between stations. A key characteristic of this phenomenon is that a pH value is associated with each location. A typical goal in the analysis of such datasets is the prediction of pH values over a dense grid of prediction locations, which together provide an estimated map of pH over the entire region.

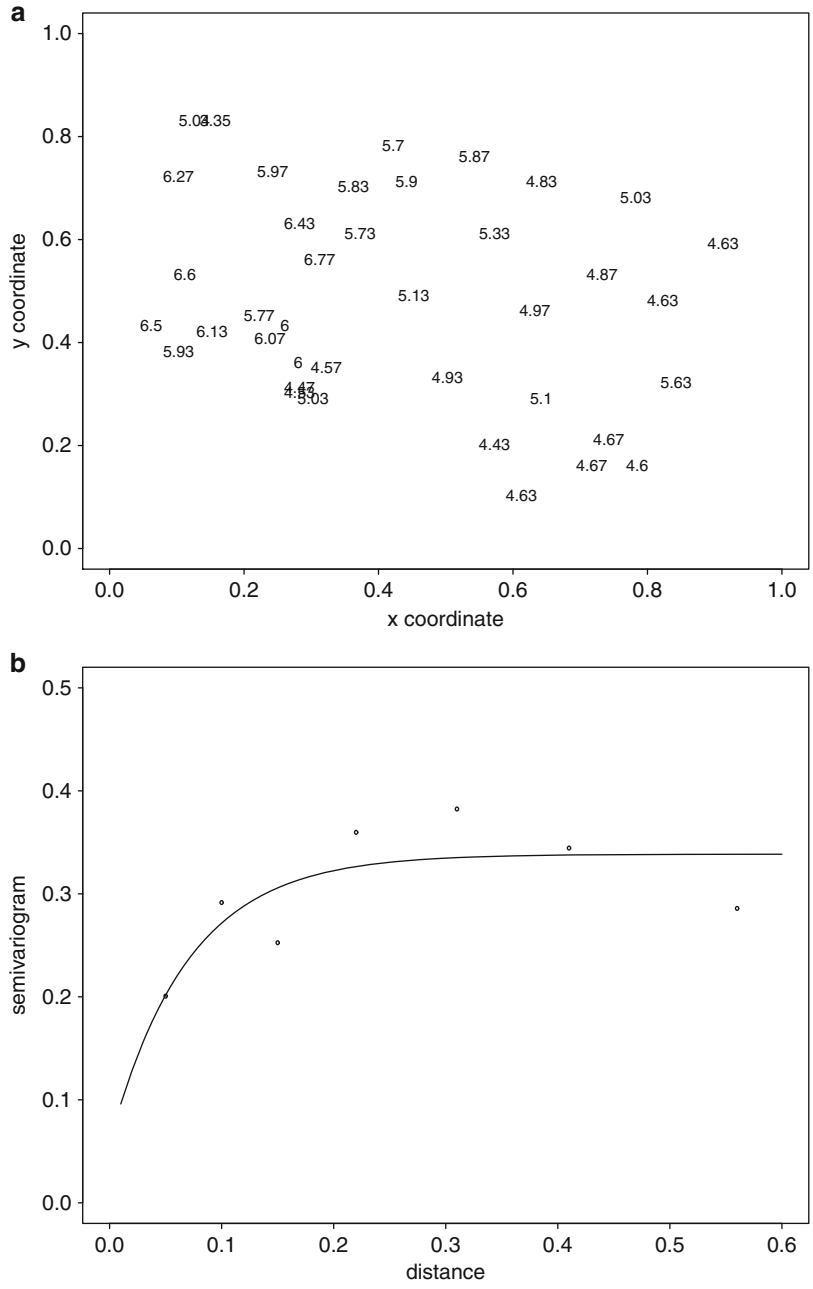
By plotting the pH values against the spatial coordinates, it can be seen that the pH values tend to decrease in the eastward direction and increase in the northward direction. We use a model with $\mu(s) = \beta_1 + \beta_2 x + \beta_3 y$, with $s = (x, y)$, for which the OLS estimates are $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (5.627, -1.440, 0.761)$. The second-order specification is completed by assuming the covariance function of the true pH process is isotropic and exponential. Figure 2b shows empirical semivariogram estimates at a few selected distances (dots) based on the OLS residuals. It displays an apparent discontinuity at the origin, suggesting the data contain measurement error, so the covariance function of the pH data is $C(h) = \sigma^2 \exp(-h/\phi) + \tau^2 1\{h = 0\}$. The estimated semivariogram function is also displayed in Fig. 2b (line), obtained using the parameters $(\hat{\sigma}^2, \hat{\phi}, \hat{\tau}^2) = (0.270, 0.070, 0.059)$, estimated by least squares.

Figure 3a shows a map of estimated pH values obtained by computing the kriging predictor with estimated parameters at about 4,200 prediction locations located inside the convex hull of the sampling locations. Except for the northwest corner of the prediction region that correspond to a group of islands, the pH values are high in the northwest of the state and decrease toward the south and east. Figure 3b shows a map of the square root of the kriging variance at the prediction locations, displaying the typical behavior of having small values at prediction locations close to some sampling location and larger values away from sampling locations.

Example 2 As an illustration of a lattice dataset, we study the relation between poverty level (POV) and total population (POP) at the county level in 2009 in the US state of Texas, using data obtained from the US Census Bureau. Figure 4 displays the state of Texas, composed of 254 counties color-coded by the 2009 logarithm of poverty levels. By plotting the data it can be seen that the logarithm of poverty level is closely linearly related with the logarithm of total population, where the least squares fit is $\hat{E}\{\log(\text{POV}) \mid \text{POP}\} = -1.741 + 0.992 \cdot \log(\text{POP})$. Based on the residuals from this fit, we have that Moran's and Geary's statistics are $I = 0.391$ and $c = 0.568$, respectively, which are both highly significant for the hypothesis of no spatial association (p-values $< 10^{-15}$). Hence, there is substantial spatial association among county log poverty levels, even after accounting for log total population.

We fitted both CAR and SAR models using log poverty level as the response and log total population as the explanatory variable, and the neighborhood system based on geographic adjacency: two counties are neighbors if and only if their boundaries intersect. As for the weights we assume that $w_{ij} = 1$ for any two neighbors s_i and s_j. The SAR model is fit by maximum likelihood, resulting in the estimates $\hat{E}\{\log(\text{POV}) \mid \text{POP}\} = -2.123 + 1.034 \cdot \log(\text{POP})$, and $(\hat{\sigma}^2, \hat{\rho}) = (0.067, 0.116)$. The estimated mean for the CAR model is similar, but the fit is slightly inferior.

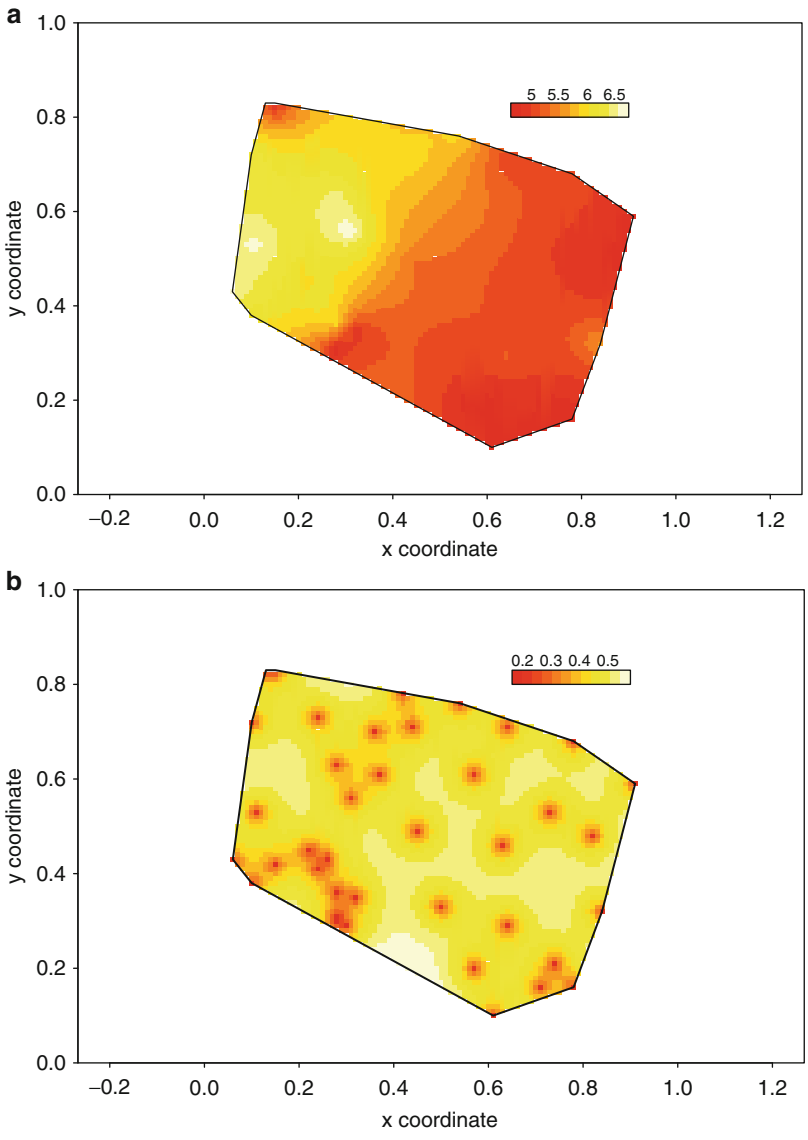
Example 3 As an illustration of a point pattern data, we consider earthquakes (with magnitude 1.0 or more on the Richter scale) that occurred worldwide in 2011 over the 8 consecutive days beginning at 00:00 h UTC on May 20. Figure 5a displays the locations of the 981 events as a "bubble map" with respect to magnitude (size of bubble is proportional to square root of earthquake's



Spatial Statistics, Fig. 2 (**a**) pH measurements and sampling locations and (**b**) empirical and fitted semivariogram function of pH measurements

magnitude) and so provides a fair visual comparison of the relative sizes (magnitudes) among events. The color-coding scheme renders earlier events in lighter shades of orange and later events in darker shades of red. Since *magnitude* is an attribute recorded along with each event's location, this is a marked point pattern.

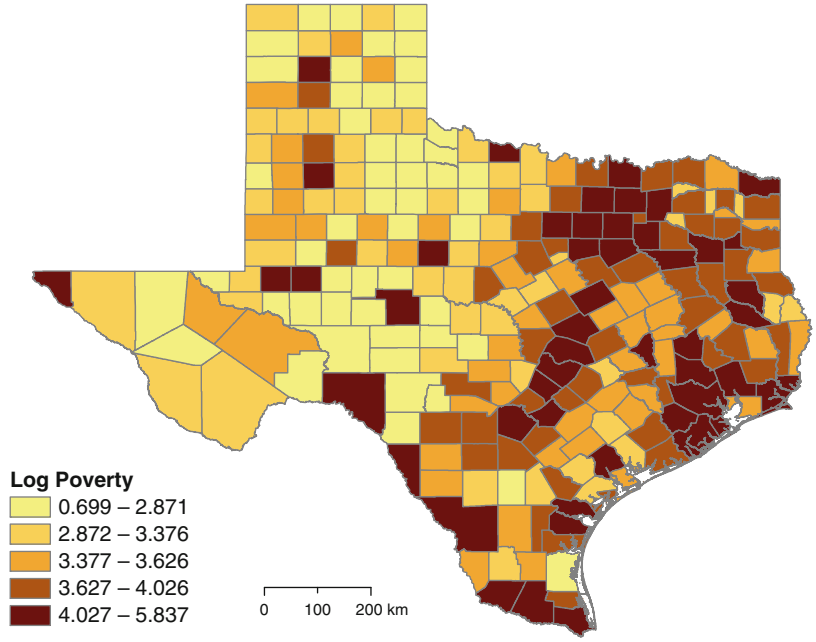
We focus merely on assessing tendency for clustering and disregard magnitude. It is obvious in the current context that there is clustering as geology informs us that this tends to occur at the junction of tectonic plates and fault zones. The Aleutian Islands/Bering Strait and southern Alaska are prominent "hot spots."



Spatial Statistics, Fig. 3 (**a**) Map of kriging predictor of pH and (**b**) map of square root of kriging variance of pH

In fact, a chi-square test strongly rejects CSR (p-value $\approx 10^{-16}$). Assuming (for the sake of illustration) stationarity and isotropy, the estimated L-function reveals a pronounced upward bow that falls well outside the 95 % confidence envelopes for a HPP, thus further confirming the strong tendency for clustering on this spatial scale.

Since a constant intensity function is an inadequate hypothesis, we continue the analysis by producing an estimate of the intensity function in the context of an IPP. The result is displayed in Fig. 5b which shows a kernel smoothing estimate (with bandwidth selected by cross-validation; see Diggle 2003). Since intensity is the expected number of (random) points per unit area, the units are "earthquakes per unit area." The two Alaskan hot spots alluded to earlier are clearly visible. Interestingly, the central Caribbean emerges as a third hot spot.



Spatial Statistics, Fig. 4 Choropleth map of county log poverty level in the US state of Texas in 2009

Historical Background and Final Remarks

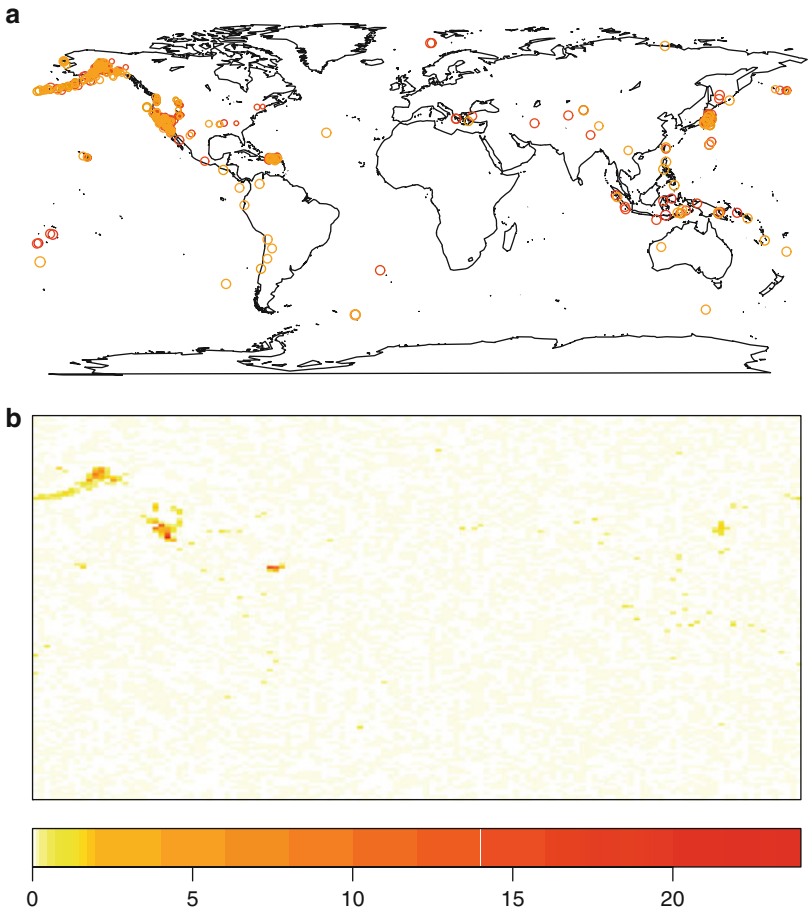
Early pioneers of statistical inference (e.g., Fisher, Gossett, Pearson) alluded to issues arising from the correlation of observations due to spatial proximity in designed experiments and proposed methods to account for it. Some of the history and early developments in spatial statistics is reviewed in Chap. 1 of Gelfand et al. (2010). Since some areas of spatial statistics have not been included in this brief overview, we end with some additional pointers to the literature. A review of non-stationary spatial processes is given in Chap. 9 of Gelfand et al. (2010). The problems of spatial sampling and design (how and where to collect the data) are treated in Cressie (1993), Le and Zidek (2006), Müller (2007), and Chap. 10 of Gelfand et al. (2010). Multivariate methods in spatial statistics are treated in Banerjee et al. (2004), Le and Zidek (2006), Wackernagel (2010), and Chap. 21 of Gelfand et al. (2010). Hierarchical models for the modeling of non-Gaussian spatial data, specially models for discrete spatial data, are discussed in Banerjee et al. (2004) and Diggle and Ribeiro (2007), where the Bayesian approach is featured prominently. Models for more complex types of spatial random objects are treated in Matheron (1975), Cressie (1993), and Nguyen (2006). Finally, an extensive discussion of available software written in R that implements the methods described here for the statistical analysis of the three types of spatial data appears in Bivand et al. (2008).

Acknowledgments

The authors thank Edgar Muñoz for producing Fig. 4. The first author was partially supported by National Science Foundation Grant HRD-0932339.

Cross-References

▸ Distance and Similarity Measures
▸ Least Squares
▸ Modeling and Analysis of Spatiotemporal Social Networks
▸ Regression Analysis
▸ Spatial Networks



Spatial Statistics, Fig. 5 (**a**) Worldwide earthquakes, May 20–27, 2011, and (**b**) Corresponding estimated intensity function

▸ Theory of Probability, Basics and Fundamentals
▸ Theory of Statistics, Basics, and Fundamentals
▸ Univariate Descriptive Statistics

References

Anselin L (1988) Spatial econometrics: methods and models. Kluwer, Dordrecht

Banerjee S, Carlin BP, Gelfand AE (2004) Hierarchical modeling and analysis for spatial data. Chapman & Hall/CRC, Boca Raton

Berke O (1999) Estimation and prediction in the spatial linear model. Water Air Soil Pollut 110:215–237

Bivand RS, Pebesba EJ, Gómez-Rubio V (2008) Applied spatial data analysis with R. Springer, New York

Chilès J-P, Delfiner P (1999) Geostatistics: modeling spatial uncertainty. Wiley, New York

Cliff AD, Ord JK (1981) Spatial processes: models and applications. Pion, London

Cressie NAC (1993) Statistics for spatial data. Wiley, New York

Daley D, Vere-Jones DJ (2003) Introduction to the theory of point processes, volume I: elementary theory and methods, 2nd edn. Springer, New York

Daley D, Vere-Jones DJ (2007) Introduction to the theory of point processes, volume II: general theory and structure, 2nd edn. Springer, New York

Diggle PJ (2003) Statistical analysis of spatial point patterns, 2nd edn. Arnold, New York

Diggle PJ, Ribeiro PJ (2007) Model-based geostatistics. Springer, New York

Gelfand AE, Diggle PJ, Guttorp P, Fuentes M (eds) (2010) Handbook of spatial statistics. Chapman & Hall/CRC, Boca Raton

Illian J, Penttinen A, Stoyan H, Stoyan D (2008) Statistical analysis and modelling of spatial point patterns. Wiley, Chichester

Journel AG, Huijbregts CJ (1978) Mining geostatistics. Academic, London

Le ND, Zidek JV (2006) Statistical analysis of environmental space-time processes. Springer, New York
LeSage JP, Pace RK (2009) Introduction to spatial econometrics. Chapman & Hall/CRC, Boca Raton
Li SZ (2009) Markov random field modeling in image analysis, 3rd edn. Springer, London
Matérn B (1986) Spatial variation. Lecture notes in statistics, 2nd edn. Springer, Berlin
Matheron G (1975) Random sets and integral geometry. Wiley, New York
Müller WG (2007) Collecting spatial data: optimum design of experiments for random fields, 3rd edn. Springer, Heidelberg
Nguyen HT (2006) An introduction to random sets. Chapman & Hall/CRC, Boca Roton
Ripley BD (1981) Spatial statistics. Wiley, New York
Rue H, Held L (2005) Gaussian Markov random fields: theory and applications. Chapman & Hall/CRC, Boca Raton
Schabenberger O, Gotway CA (2005) Statistical methods for spatial data analysis. Chapman & Hall/CRC, Boca Raton
Sjöstedt-De Luna S, Young A (2003) The bootstrap and kriging prediction intervals. Scand J Stat 30:175–192
Stein ML (1999) Interpolation of spatial data: some theory for kriging. Springer, New York
Wackernagel H (2010) Multivariate geostatistics: an introduction with applications, 3rd edn. Springer, Berlin
Yaglom AM (1987) Correlation theory of stationary and related random function I: basic results. Springer, New York

Spatial-Textual Web Search

▸ Spatiotemporal Information for the Web

Spatiotemporal Analysis

▸ Modeling and Analysis of Spatiotemporal Social Networks
▸ Social Networks in Emergency Response

Spatiotemporal Collaborative Filtering

▸ Spatiotemporal Personalized Recommendation of Social Media Content

Spatiotemporal Footprints in Social Networks

Linna Li and Michael F. Goodchild
Department of Geography, University of California, Santa Barbara, Santa Barbara, CA, USA

Synonyms

Check-in; Location; Movement; Spatial interaction; Time; Trajectory

Glossary

GPS Global Positioning System, a satellite-based navigation system that provides location and time almost anywhere near the Earth's surface

VGI Volunteered geographic information, a type of user-generated content with a spatial component (Goodchild 2007)

Flickr A popular photo-sharing website allowing people to upload and share photos that may be tagged with location

Twitter A popular microblogging and social networking service that supports sending text messages (which are called tweets) of less than 140 characters. Tweets may be associated with location

Spatial Interaction Models Models describing interaction between two locations as a variable dependent on distance

Definition

Spatiotemporal footprints discussed in this chapter are locational and temporal information regarding people's occurrences that are digitally recorded. Spatial location may be automatically captured as latitude and longitude by a GPS receiver or provided by a user as a place name, e.g., a city or neighborhood. Along with spatial information, time is usually automatically recorded,

too. Spatiotemporal footprints may be intentionally collected as a series of point locations when people move around, such as GPS tracks. They may also be attached to instant messages, associated with telephone calls, tagged to photos, or linked with other forms of human activities. Footprints are basically location and time that indicate where a person went and when. Several other terms are also partially synonymous with spatiotemporal footprints. Check-in is used by social networking services, such as Foursquare and Google Latitude, to allow users to instantly share their physical locations. Concatenation of spatial footprints in the order of time generates trajectories of human movements, which have been a long-lasting research topic in the social sciences, where they are studied in the context of migration, travel, commuting, etc.

Introduction

Society is comprised of many different types of social networks on various levels. Social networks play a critical role in achieving goals and solving problems. While traditional social networks only existed within a very limited geographical distance (e.g., villages) constrained by temporal factors, modern technologies – especially the growth of the Internet, the wide adoption of cell phones, and the support of Web 2.0 technologies – have greatly reduced spatiotemporal limitations on human communication. People who live on different continents in different time zones can interact with each other using phones, emails, and websites. Particularly, online social networking services provide an effective channel to enhance existing social networks and to initiate new ones. Facebook, for example, offers services to create profiles, add friends, and exchange information. Twitter, as another example, provides a platform to share and discover "what is happening right now (at where)?" These services offer an alternative and complementary form of social networks with a growing number of users.

Human activities take place in particular locations at specific times; activities in online social networks may reflect activities in the physical world. For instance, people may comment on an event that they are currently experiencing, such as a football game or a fire. On the other hand, activities in the physical world may generate a virtual social network. To facilitate organization of regular gatherings, local friends may create a virtual group on Facebook to publish information and to share photos. One of the major reasons for people to record spatiotemporal footprints of activities is to share them with family and friends, basically with people within their social networks; such sharing has generated large amounts of locational and temporal data. This phenomenon has attracted increasing attention from both academia and industry, because the prevalence of such information provides a great potential to study human mobility, human activities, and the composition of large-scale social networks, using vast volumes of geospatial data on large samples of people, for the first time in history.

Uncertainty in Footprints

Spatiotemporal footprints are voluminous with very rich information; however, we need to be aware of the inherent uncertainty associated with them in order to validate conclusions based on these data. Spatial uncertainty is the difference between a recorded position and the corresponding position in reality. It is critical to understand uncertainty when dealing with locational information, so it has been studied extensively in the field of GIScience. For example, Goodchild and Gopal (1989) compiled a set of papers that address accuracy of spatial data from a wide range of applications, including both physical and social phenomena. Zhang and Goodchild (2002) systematically examined spatial uncertainty modelling in continuous and categorical variables. Footprints generated in social media and social networks are a type of VGI. Unlike location recorded in a scientific database that aims to minimize spatial uncertainty and inaccuracy as much as possible with standard quality control procedures, uncertainty associated with footprints may vary from case to case due to the nature of

tweet_text	created_at	geo_lat	geo_long
@Rocknrealty volunteering for Alejandro's high sch...	2011-01-20 17:49:24	30.49307	–97.77580
I'm at El Kartel (1025 Robson St., at Burrard St.,...	2011-01-20 17:49:24	49.28649	–123.12775
I'm at Phillips Seafood (900 Water St SW, Washingt...	2011-01-20 17:49:24	38.88053	–77.02669
@ThenAndreaSaid 'Just keep swimming, just keep swi...	2011-01-20 17:49:24	26.21436	–98.13900
Bear Grylls is the funniest person ever !!!! Haha	2011-01-20 17:49:24	35.78715	–78.59540

photo_title	description	photo_tags	date_taken	geo_lat	geo_long
Seattle, WA	Pike Place Corner Market		2008-08-21 12:12:22	47.60892	–122.34058
Happy New Year! 2011 will rock.		square squareformat iphoneography instagramapp upl...	2011-01-01 00:01:10	45.53898	–122.63066
Olivia for President 2012	A paraody sure to please a child in the next elect...	olivia president 2012	2010-12-31 23:56:46	47.62035	–122.34900

Spatiotemporal Footprints in Social Networks, Fig. 1 Latitude and longitude associated with tweets in Twitter and photos in Flickr

VGI (Goodchild and Li 2012). Users may choose to disclose or hide footprints and select the level of disclosure. For example, they can reveal their exact point location as coordinates, or they can show only the city name as the location.

When users of social networking services or social media share their locations, they usually have several options. If GPS is enabled in a mobile device such as a smart phone, location is recorded automatically as latitude and longitude. Otherwise, users can select a city or neighborhood name from a set of provided place names that are usually reverse geocoded based on an estimate of their device location. The degree of uncertainty in footprints depends on the mechanism used to record location. There are two major categories of footprints: those recorded by a digital system automatically and those provided by a user manually. Automatic footprints may be generated by GPS, relative location of cell phone towers, or IP address. Location produced by the same method has a similar level of uncertainty, but it varies from one method to another. Spatial footprints captured by GPS (either a GPS unit or built-in GPS in mobile phones or cameras) in the form of latitude and longitude are supposed to be the most accurate means to record location. Uncertainty of footprints recorded by GPS is usually several meters, depending on the particular device and the surrounding conditions (e.g., satellite visibility). However, the number of decimal digits of stored coordinates may not reflect their actual degree of uncertainty. For example, latitude and longitude associated with tweets in the Twitter database and photos in the Flickr database both have 5 decimal places (Fig. 1), indicating that spatial precision should be around 1 m, which is not the correct expectation of GPS uncertainty in most devices. In addition, approximate location may be determined by the relative position of a user's equipment in a cellular network, leading to uncertainty as large as a cell area, ranging from 150 to 30,000 m (Zhao 2000). Physical location inferred from an IP address may be at the level of ZIP code, city, state, or even country and thus show varying accuracy. Databases have been established to map the correspondence between IP address and physical address (e.g., http://whatismyipaddress.com/ip-lookup), and efforts have been made to increase accuracy of IP address locators (Guo et al. 2009). Uncertainty of temporal footprints is less complex. Time of users' interaction with the Web or mobile services is always automatically recorded with good accuracy. However, temporal information provided by a user may be arbitrary with an uncertainty that is difficult to estimate. For instance, there are various degrees of uncertainty in the times associated with photos in Flickr.

Key Applications

Numerous questions that were not answerable before due to the lack of data can now be investigated using spatiotemporal footprints. In general, footprints have been utilized to study place and people. Place refers to geographic features that are present at particular locations on the Earth's surface. A place could be a simple feature such as a road or a restaurant with a clear boundary or a vague feature without an exact agreed-upon location, such as "downtown." The focus of this type of research is on the geographic landscape associated with footprints without explicit consideration of the people who provide them. The second category of research on footprints emphasizes the people who record them, their behavior, and the relationships that may be inferred from the pattern of their whereabouts. Here are some example research questions. How do people move around a city? What percentage of their trips is captured by footprints recorded in social networking services? What relationships can be extracted from the pattern of spatiotemporal footprints of two people or a group of people in a social network?

Locational information in footprints can be used to characterize the position and shape of geographic features. This type of footprint is usually called VGI – a special form of user-generated content with a geographic component (Goodchild 2007). A typical example is OpenStreetMap (http://www.openstreetmap.org/), with a goal to create a free editable map of the world. Originally, map data were contributed by volunteers using a handheld GPS to record their walking or biking paths. People purposefully collect continuous footprints to produce geographic infrastructure data, mostly of roads and points of interest (POIs). In this case, spatial footprints are used to identify the location of geographic features: what is available at that particular location on the Earth's surface? People may choose to map whatever features that are interesting to them. For instance, SeeClickFix, a Web service supported by both Web browsers and mobile apps, enables citizens to report the location of nonemergency issues within their communities (e.g., a broken traffic light), while governments use this information to respond more promptly to the problems and take actions to fix them. In addition to geographic features created by individual users, places may also be inferred by aggregating spatial footprints generated by multiple people. Clusters of spatial footprints suggest popular places, and clusters of footprints in both space and time may indicate events. For example, location and spatial extent of specific places can be constructed as a probability-density surface by extracting and summarizing footprints of photos tagged with the same place name (Li and Goodchild 2012). Spatial boundaries of city cores (e.g., downtown, CBD) may also be defined based on photo footprints obtained from Flickr (Hollenstein and Purves 2010). Lee and Sumiya (2010) proposed a method to detect unusual geo-social events (e.g., local festivals) by comparing spatiotemporal patterns of footprints in the study area with the distribution of footprints in normal times.

Spatiotemporal footprints can also be used in combination with other information to gain knowledge about the Earth. Together with visual information in photos, representative scenes at different locations were automatically selected from vast volumes of geotagged photos in Flickr (Crandall et al. 2009). In another type of application, location is attached to collected data about some natural phenomenon as a systematic spatial sampling strategy to facilitate scientific data analysis. One famous example is the Audubon Society's Christmas Bird Count project that started in 1900. Volunteer birdwatchers are divided into small groups to follow assigned routes and to count birds they see along the routes. Moreover, locational information with regard to disaster status contributed by citizens has been proved very helpful in emergency response (Li and Goodchild 2010), such as wildfires (Goodchild and Glennon 2010). A spatial model was proposed to estimate the location of an earthquake and the trajectory of a typhoon in Japan based on georeferenced tweets (Sakaki et al. 2010).

Furthermore, footprints are also used to study the people who generate them. Concatenation of spatiotemporal footprints of a single user provides a trajectory of the places he or she has

visited at specific time of the day, which may shed light on people's daily activities. We may collect data automatically on places people have been to or how long they stay at a particular place. For example, georeferenced tweets may provide a real-time record of people's activity episodes that is even more accurate than a travel diary recorded from memory recall. Besides, this data collection is nonintrusive, so subjects do not need to write down what they are doing and at what time, because the time of a tweet is recorded automatically by the online service and the tweet content may suggest their activities. Traditionally, travel behavior was studied using travel diaries that were a part of a travel survey, which is very expensive in terms of time and labor. Research has already been done on methods to use recorded footprints as complementary to traditional self-reported travel surveys in studying travel behavior. Murakami and Wagner (1999) discussed the use of GPS to collect automatically date, start time, end time, and vehicle position in trips at frequent intervals. A comparison between GPS-recorded trips and self-reported trips shows that self-reported distances are much longer than the actually traveled distances. In addition to locational and temporal information stored by GPS in trips, even purposes of trips may be inferred from footprints with auxiliary land-use data (Wolf et al. 2001). Crandall et al. (2009) reconstructed the pathways of people who visit Manhattan and the San Francisco Bay area based on footprints associated with photos uploaded to Flickr. Although footprints recorded by GPS have been used to study travel behavior and trajectories may be extracted from photo footprints, not much research has been done to investigate detailed travel behavior at the level of traditional travel surveys using footprints collected in social networks and social media.

Comparison of spatiotemporal footprints between different users or groups of people may indicate the relationship between them in social networks. According to the first law of geography, "Everything is related to everything else, but near things are more related than distant things" (Tobler 1970). If this is true for social phenomena, we can infer the strength of social interactions between people at two places based on their spatial footprints. Distance between people may signify the probability of them being friends or acquaintances. Spatial interaction models have been developed to describe this relationship (Isard 1960). The distance-decay effect is characterized this way: as the distance between two locations increases, the interaction between them decreases. A typical example is a gravity model (Abler et al. 1971):

$$I_{ij} = a\frac{M_i M_j}{d_{ij}^b} \tag{1}$$

where I_{ij} is the interaction between i and j, a is a constant, M_i and M_j are properties associated with i and j, d_{ij} is the distance between i and j, and b is another constant, dependent on the phenomenon.

Researchers have started to investigate the role of distance derived from footprints in studying social relations. A collection of maps were produced to represent cyberspace, including online communications and connections between people located in different places (Dodge and Kitchin 2001). Strong connections may exist between people who visit the same place at the same time regularly. Using geotagged photos from Flickr, the probability of a social tie between people is calculated based on the co-occurrences of their spatial and temporal footprints (Crandall et al. 2010). Cho et al. (2011) studied the relationship between social ties and people's movement patterns using data collected from public check-ins in online social networks and cell phone location trace data. A study on mobile phone data demonstrates that distance decay is present in the number of calls and the number of co-locations, defined as people sharing the same location at the same time (Calabrese et al. 2011). Hardy et al. (2012) applied a gravity model to describe the decrease of the likelihood of a person to contribute to a georeferenced article in Wikipedia when the distance between the user and the subject place in the article increases.

Future Directions

Although voluminous amounts of footprints are generated in social networks every day, data discovery and data access are still two important considerations. How can we find relevant geographic data in social media? How can we collect more spatial and temporal footprints of social networks from spatially embedded populations, both online and off-line? How can geography promote the "human as sensor" paradigm in spatial data generation? How can we harvest spatio-temporal footprints about involved people from existing sources?

Another critical question is the synthesis of footprints with various accuracies generated in different contexts. How can we quantify the uncertainty in a particular footprint? Can we apply mathematical models of uncertainty developed in the GIScience literature to study uncertainty in footprints? How can we use footprints that are available as both coordinates and place names to do cross-validation, so as to increase spatial accuracy? Currently, research has been done with only a single source of footprint data (i.e., footprints created in one social networking service). It would be valuable to investigate the potential for using footprints collected from different sources to improve the data quality and quantity. Geographic data conflation has been applied to merge spatial data from multiple sources (Saalfeld 1988; Li 2010; Li and Goodchild 2011). Can these techniques be used in integration of spatiotemporal footprints generated in various social networks? What type of new methodologies might be required in footprint data synthesis?

Furthermore, representativeness of the available footprints is an interesting yet challenging research area. Since systematic sampling strategies are not applied in the collection of footprints, how representative are these data compared to the total population under study? (Li et al. 2013) What are the major types of motivation of people who join online social networks? The usual users of social media are undoubtedly self-selected. What characteristics cause them to join online social networks and to leave spatiotemporal footprints? Is there a way to measure the bias in this type of data source?

More analyses could be performed using footprints generated in social networks. How can we identify social relations and mobile patterns from heterogeneous footprint data? Social networks are embedded in space and time. However, that embedding may not always be relevant to specific analysis. How can we incorporate spatial interaction functions into different types of network, particularly when space and time vary significantly? For example, spatial dependence and distance decay are valid in many processes in geographic space but are less relevant when all people in the social network are in the same room. While it is known that social network links decay with distance and that new technologies do not completely overcome this decay, what is not known is the circumstances where the technologies overcome the decay (e.g., where in a task cycle). Moreover, the predictive value of knowing that network links fall off with distance seems low. What can be predicted with a better understanding of the relationship between space and networks? What are the types of activities in social networks that are strongly constrained by space and time? For what type of groups is group maintenance and persistence dependent on spatial locations? How can we use spatiotemporal footprints to infer missing network data, select specialized social network subgroups, and forecast change? How are network-mediated processes (e.g., information diffusion, VGI) influenced by spatial and temporal relations (e.g., nearness in space or time)?

Finally, privacy in social network studies has attracted much attention (Li and Goodchild 2013). When is it appropriate to collect information on people's footprints and to study them without their knowledge? Revelation of locations may lead to crimes, such as stalking and burglary. Is there a way to preserve spatiotemporal patterns of social networks and to protect privacy simultaneously? What types of generalization and aggregation from statistics and cartography can be adapted to achieve the two objectives? What would be an appropriate

level of generalization of locational data for a particular application? How does the level of abstraction limit the types of network questions that can be answered?

Cross-References

- ▸ Spatiotemporal Information for the Web
- ▸ Spatio-Temporal Outlier and Anomaly Detection
- ▸ Spatiotemporal Proximity and Social Distance
- ▸ Spatiotemporal Reasoning and Decision Support Tools

References

Abler R, Adams J, Gould P (1971) Spatial organization—the geographer's view of the world. Prentice-Hall, Englewood Cliffs

Calabrese F, Smoreda Z, Blondel VD, Ratti C (2011) Interplay between telecommunications and face-to-face interactions: a study using mobile phone data. PLoS ONE 6(7):e20814

Cho E, Myers S, Leskovec J (2011) Friendship and mobility: user movement in location-based social networks. In KDD, San Diego. ACM, pp 1082–1090

Crandall DJ, Backstrom L, Huttenlocher D, Kleinberg J (2009) Mapping the world's photos. In: Proceedings of the 18th international conference on World Wide Web, Madrid, 20–24 Apr 2009

Crandall DJ, Backstrom L, Cosley D, Suri S, Huttenlocher D, Kleinberg J (2010) Inferring social ties from geographic coincidences. Proc Natl Acad Sci 107(52):22436–22441

Dodge M, Kitchin R (2001) Mapping cyberspace. Routledge, New York

Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69: 211–221

Goodchild MF, Glennon JA (2010) Crowdsourcing geographic information for disaster response: a research frontier. Int J Digit Earth 3(3):231–241

Goodchild MF, Gopal S (1989) Accuracy of spatial databases. Taylor and Francis, New York

Goodchild MF, Li L (2012) Assuring the quality of volunteered geographic information. Spat Stat 1:110–120

Guo C, Liu Y, Shen W, Wang HJ, Yu Q, Zhang Y (2009) Mining the web and the internet for accurate ip address geolocations. In: Infocom mini conference, 2009, Rio de Janeiro

Hardy D, Frew J, Goodchild M (2012) Volunteered geographic information production as a spatial process. Int J Geogr Inf Sci 26:1191–1212

Hollenstein L, Purves R (2010) Exploring place through user-generated content: using Flickr to describe city cores. J Spat Inf Sci 1(1):21–48

Isard W (1960) Methods of regional analysis. MIT, Cambridge

Lee R, Sumiya K (2010) Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In: Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks (LBSN 2010), San Jose, pp 1–10

Li L (2010) Design of a conceptual framework and approaches for geo-object data conflation. PhD dissertation, Department of Geography, University of California, Santa Barbara

Li L, Goodchild MF (2010) The role of social networks in emergency management: a research agenda. Int J Inf Syst Crisis Response Manag 2(4):49–59

Li L,Goodchild MF (2011) An optimization model for linear feature matching in geographical data conflation. Int J Image Data Fusion 2(4):309–328

Li L, Goodchild MF (2012) Constructing places from spatial footprints. In: Proceedings of ACM SIGSPATIAL GEOCROWD'12, Redondo Beach, 6 Nov 2012

Li L, Goodchild MF, Xu B (2013) Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. Cart Geo Inf Sci 40(2):61–77

Li L, Goodchild MF (2013) Is privacy still an issue in the era of big data? Location disclosure in spatial footprints, In: Proceedings of the 21st International Conference on Geoinformatics, June 20–22, 2013. Kaifeng, Henan, China.

Murakami E, Wagner D (1999) Can using global positioning system (GPS) improve trip reporting? Transp Res C 7:149–165

Saalfeld A (1988) Conflation automated map compilation. Int J Geogr Inf Syst 2(3):217–228

Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on World Wide Web, Raleigh, 26–30 Apr 2010

Tobler W (1970) A computer movie simulating urban growth in the Detroit region. Econ Geogr 46(2): 234–240

Wolf J, Guensler R, Bachman W (2001) Elimination of the travel diary: an experiment to derive trip purpose from GPS travel data. In: Proceedings from the transportation research board 80th annual meeting, Washington, DC

Zhang J-X, Goodchild MF (2002) Uncertainty in geographical information. Taylor and Francis, New York

Zhao Y (2000) Mobile phone location determination and its impact on intelligent transportation systems. IEEE Trans Intell Transp Syst 1(1):55–64

Spatiotemporal Information for the Web

Peiquan Jin, Sheng Lin, and Qingqing Zhang
School of Computer Science and Technology, University of Science and Technology of China, Hefei, China

Synonyms

Spatial-textual Web search; Spatiotemporal Web; Temporal-textual Web search

Glossary

Location A site name or geographic scope mentioned in Web pages
Time One or more units of chronons. It can be a time instant or a time period
OID Object identifier
AD Attribute descriptor
LD Location descriptor
TD Time descriptor
Search Engine Search engine is a popular tool to find information in the Web
Primary Time The most appropriate time associated with a Web page
Primary Location The most appropriate location associated with a Web page
GEO/GEO Ambiguity it refers that many locations can share a single place name
GEO/NON-GEO Ambiguity it refers that a location name can be used as other types of names
NER Named entity recognition
GRT Global reference time
LRT Local reference time

Definition

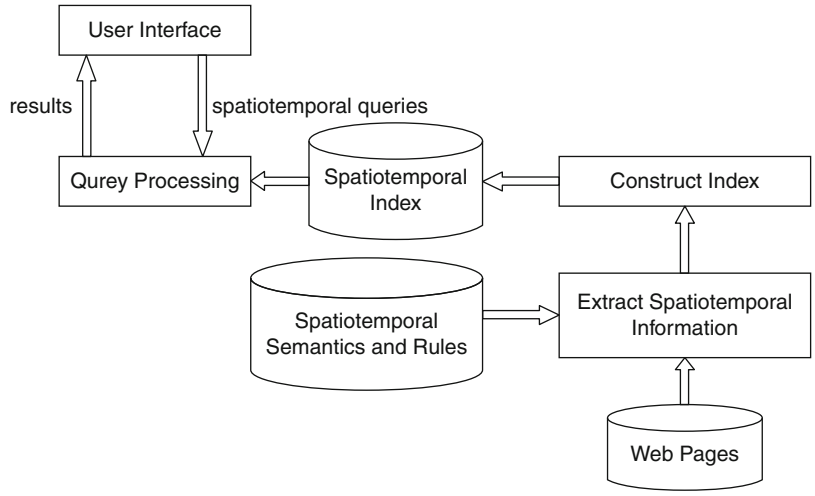
This subject is mainly towards the spatiotemporal information involved in the Web, particularly in Web pages. Typical spatiotemporal information in the Web includes the locations and time mentioned in Web pages, the update date of Web pages, and the Web server locations. As we know, location and time are the essential dimensions of information including Web information. However, they are usually ignored in traditional keyword-based Web search engines.

Traditional search engines are basically based on keyword-based approaches or content-based methods. Though many contributions have been presented in both directions, in some cases users are still difficult to express their search needs. For example, more than 70 % Web queries are related with time and locations (Setzer and Gaizauskas 2002; Sanderson and Kohler 2004), but spatiotemporal Web queries such as "*to get the news about Olympic Beijing in recent three days*" or "*to get the sales information about Nike in Beijing in this week*" are often with bad results in traditional search engines. One reason is that such queries are difficult to express in keyword-based search engines. Moreover, traditional search engines also lack of the ability to process such spatiotemporal queries.

Aiming at improving the effectiveness and efficiency of spatiotemporal queries in search engines, many researchers began to study the spatiotemporal information in the Web. However, most of previous researches focused on time-based Web search (Nunes et al. 2008) and location-based Web search (Wang et al. 2005; Ding et al. 2000; Zhou et al. 2005; Markowetz et al. 2005) separately. And few works considered the temporal information of the content in Web pages. In this entry, we will describe the semantics of spatiotemporal information in the Web and try to present a framework for spatiotemporal information extraction under the Web context.

Introduction

The main goal of incorporating spatiotemporal semantics into search engines is to develop a search engine that is able to express and process spatiotemporal queries. Figure 1 shows the



Spatiotemporal Information for the Web, Fig. 1 The framework of Web search system based on spatiotemporal semantics

framework of a spatiotemporal Web search engine. In this framework, spatiotemporal information is first extracted from Web pages based on spatiotemporal semantics and rules, and then we use them to construct a spatiotemporal index for Web pages, using the extracted spatiotemporal information. Users can input location- or time-related queries through the user interface, and the query processing engine will interpret the queries and perform an index-based search on archived Web pages. Finally, the resulted Web pages are returned to users according to an improved ranking algorithm, which combines text ranking techniques with new temporal and spatial ranking mechanisms.

Spatiotemporal information has been deeply studied in spatiotemporal database area, in which moving geographic objects are concentrated. However, they are not popular in Web context. So at present, the main focus on spatiotemporal information in the Web is to integrate location and time information into search process, such as information extraction, indexing, querying, ranking, and visualization.

In this entry, we focus on the spatiotemporal semantics of Web information, mainly of Web pages, and present a framework to represent and extract the spatiotemporal information in the Web.

Key Points

The spatiotemporal semantics of Web pages refer to the ontological meaning of the time and location information of Web pages. The Web can be regarded as a database of Web pages, so a Web page can be looked as an object in the Web. According to the object-oriented theory, an object consists of a unique identifier and other attributes that describe the properties of the object. Based on this view, a Web page is a spatiotemporal object which contains the following parts:

Identifier: The identifier of a Web page is usually the URL.

Locations: The spatial information of a Web page may consist of two types of locations, which are provider location and content locations (Wang et al. 2005). The provider location refers to the physical location of the provider who owns the Web resource. The content locations are the geographic locations that are described in the content of a Web page.

Time: The temporal information of a Web page has two types: update time and content time. The update time is the latest modified time of a Web page. The content time is the time that the content of a Web page indicates. The content time may contain implicit time such as "Today" and "Three Days Ago".

Non-spatiotemporal attributes: The non-spatiotemporal attributes of a Web page refer to the traditional keywords set of the Web page.

Historical Background

Most Web pages contain location and time information. Previous works regard the locations in Web pages as geographic scope (Ding et al. 2000), which can be determined by analyzing the content and links in the Web page. The locations in a Web page usually have spatial containment relationships. For example, "China" contains "Beijing." In the literatures (Zhou et al. 2005), a classification framework for Web locations is presented, and an algorithm to extract the locations in Web pages is further proposed. In order to support spatial computation, they use MBRs (Minimal Bounding Rectangle) to represent the geographic scope of Web pages. There are also some other methods proposed to represent geographic scopes, such as raster-based representation (Markowetz et al. 2005). Generally, the MBR-based method is widely used (Lee et al. 2003; Ma and Tanaka 2004). One problem of those previous works is that they treat the geographic scope of a Web page as exact one MBR, which is not very precise for many Web pages.

Temporal information is also very common in Web pages, especially in news pages. Temporal information extraction first appeared in MUC-5 whose task was to extract from business news when a joint venture took place. In MUC-6 some research was done on extracting absolute time information as part of general tasks of named entity recognition (Sundheim and Chinchor 1995). In MUC-7, the notion of temporal information extraction was expanded to include relative time in named entities (Chinchor 1998). MUC is practically the pioneer and prime driver of temporal information extraction research.

The temporal information of a Web page refers to the time related with it, e.g., the created date of the Web page and the date of an event reported in the Web page. There are many representation forms for the temporal information in Web pages, such as yesterday, Christmas, and August 15, 2012. Besides, many Web queries are time sensitive. Fresh Web pages have more important roles when users are searching news or sales information.

Proposed Solution and Methodology

In this section, we introduce the approach in capturing spatiotemporal information in Web search. The proposed approach consists of three main components: (1) Semantic Modeling for Spatiotemporal Information in the Web, (2) Extracting Primary Location from the Web, and (3) Extracting Primary Time for Web Pages.

Semantic Modeling for Spatiotemporal Information in the Web

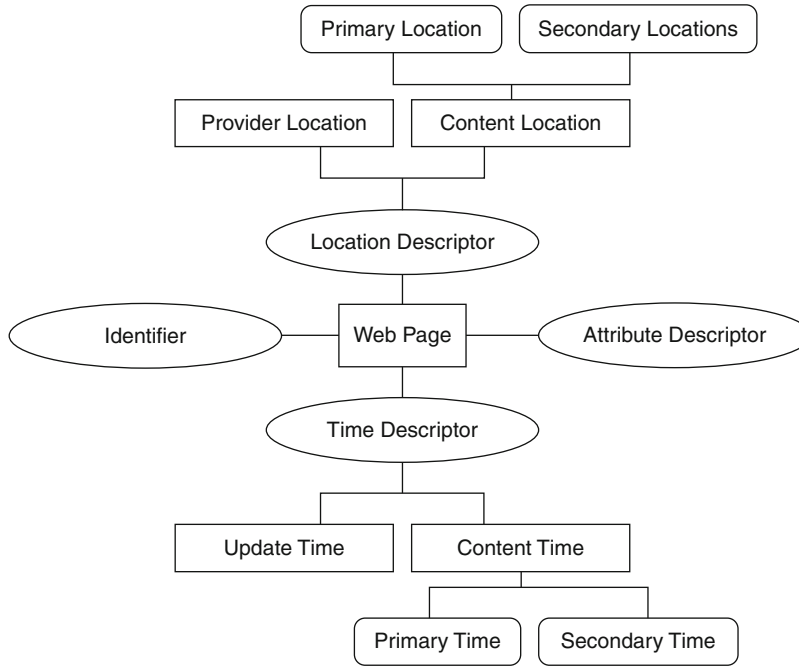
From an object-oriented perspective, a Web page can be defined as follows:

Definition 1 *A Web page is a quintuple* $O = < OID, LD, TD, AD >$*, where OID (Object IDentifier) is the identifier of the Web page, LD (Location Descriptor) is the location descriptor describing the location information of the Web page, TD (Time Descriptor) is the time descriptor describing the temporal information of the Web page, and AD is the attribute descriptor which describes the non-spatiotemporal properties of the Web page.*

Figure 2 shows the spatiotemporal semantic model of Web pages.

Location Descriptor

Location descriptor represents the location information of a Web page. A Web page has a unique *provider location* which is the geographic location of the Web server containing the Web page. The location information that described in the content of a Web page is called *content locations*. For example, in a company's homepage, the provider location may be "Beijing," since the Web server containing the homepage is located in Beijing, while the content locations may include the address of the company and other locations. As many locations may be involved in the content of a Web page, we should define a



Spatiotemporal Information for the Web, Fig. 2 Spatiotemporal semantic modeling for a Web page

primary location for the content of a Web page. The primary location is the most appropriate location that describes the location information of a Web page. In the previous example, the primary location of the Web page could be the address of the company. However, how to compute the primary location of a Web page is an unrevealed issue in location-based Web search area.

Time Descriptor

Time descriptor represents the temporal information of a Web page. There are two types of temporal information related to Web pages:

Update Time. This refers to the update time of the corresponding file of a Web page. For a given Web page, the update time is unique and can be regarded as the timestamp of the Web page. Whenever a Web page is updated, the update time is also renewed.

Content Time. This refers to the involved temporal information in the text content of a Web page. Compared with update time, which is unique and explicit for a specific Web page, the content time is a set of time instant or time period which may be explicit or implicit. For example, a news page may contain the explicit published time "2008-1-24" of the news in the title. Meanwhile, in the news body, there may have some temporal keywords such as "three days ago" and "today." The implicit content time should be translated into calendar time. Among the many time instants and periods described in the content of a Web page, we also need to define and compute the *primary time* of the Web page. The primary time of a Web page is the most appropriate time related to the Web page. In time-based Web search engine, primary time and secondary time should be treated and searched in different ways.

The above classification on the Web page time mainly considers the role of time in Web pages. Upon another view on time structure, there are two types of time: instant and period.

Instant. Instant is a specific point in the timeline. An instant may be a second, e.g., "2008-04-01 11:59:59." It also can be a time point related to current time, e.g., "one hour ago" means the time instant which is one hour before current time.

Period. Period is time duration. It contains a pair of instants and represents the time duration between the instants. For example, "[2000-09-01 00:00:00, 2003-02-01 00:00:00]" represents the time duration from "2000-09-01 00:00:00" to "2003-02-01 00:00:00," and "[2002-09-01 00:00:00, NOW]" indicates the time duration since "2002-09-01 00:00:00."

Another issue when considering the temporal semantics of Web pages is the granularity of the time. Different events in Web pages will have different granularities, e.g., the foundation event of a company may use "day" as the granularity, while a news report about earthquake may use "second." How to set up a unified referential framework for the temporal granularity is a critical issue in the spatiotemporal information modeling of Web pages.

Attribute Descriptor

Attribute descriptor describes the text keywords that mostly depict the content of a Web page. Generally, it consists of a set of keywords which are extracted from the Web page. Many traditional technologies can be used to construct the attribute descriptor of a Web page, such as word segment and keyword extraction in commercial search engines.

Extracting Primary Locations from the Web

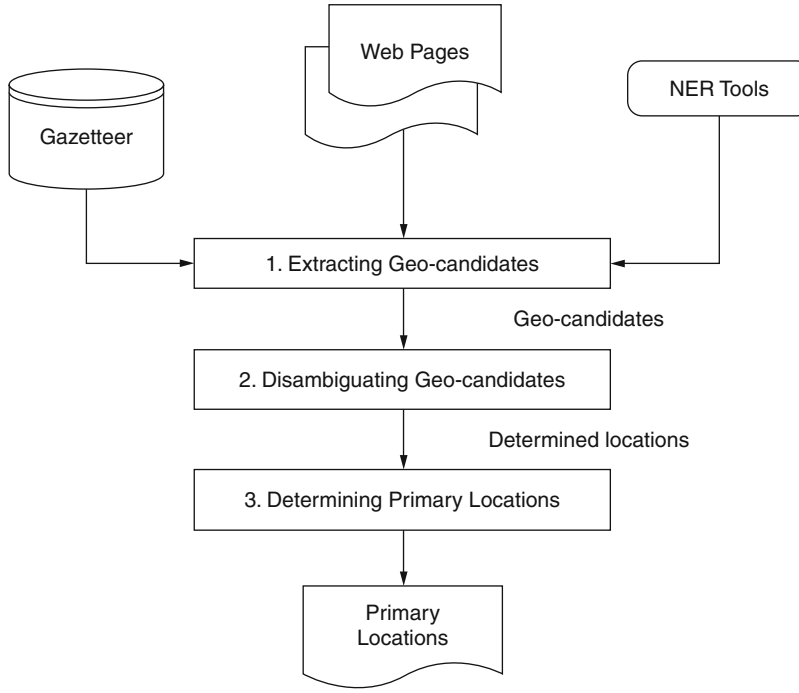
Most Web pages are associated with certain locations, e.g., news report and retailer promotion. Therefore, how to extract locations for Web pages and then use them in Web search process has been a hot and critical issue in current Web search.

As a Web page usually contains two or more location words, it is necessary to find the primary locations of the Web page. The primary locations represent the most appropriate locations associated with contents of a Web page. Generally, we assume that each Web page has several primary locations. The most difficult issue in determining primary locations is that there are GEO/GEO and GEO/NON-GEO ambiguities existing in Web pages. The GEO/GEO ambiguity refers that many locations can share a single place name. For example, Washington can be 41 cities and communities in the USA and 11 locations outside. The GEO/NON-GEO ambiguity refers that a location name can be used as other types of names, such as person names. For example, Washington can be regarded as a person name as George Washington and as a location name as Washington, D.C. Mark Sanderson's work (2000) shows that 20–30 % extent of error rate in location name disambiguation was enough to worsen the performance of the information retrieval methods. Due to those ambiguities in Web pages, previous research failed to reach a satisfied performance in primary location extraction.

On the other side, it is hard to resolve the GEO/GEO and GEO/NON-GEO ambiguities as well as to determine the primary locations of Web pages through the widely studied named entity recognition (NER) approaches. Current NER tools in Web area aim at annotating named entities including place names from Web pages. However, although some of the GEO/NON-GEO ambiguities can be removed by NER tools, the GEO/GEO disambiguation is still a problem. Furthermore, NER tools have no consideration on the extraction of the primary locations of Web pages. Basically, the NER tools are able to extract place names from Web pages, which can be further processed to resolve the GEO/GEO ambiguities as well as the GEO/NON-GEO ones. Thus, we will not concentrate on the NER approaches but on the following disambiguation and primary location determination. Those works differ a lot from traditional NER approaches.

The General Framework

Figure 3 shows the general process to extract primary locations from Web pages, in which we first extract geo-candidates based on Gazetteer and NER (named entity recognition) techniques. After this procedure, we get a set of geo-candidates. In this set, the relative order of candidates is the same as that in the text. Here, geo-candidates are just possible place names, e.g., "Washington."



Spatiotemporal Information for the Web, Fig. 3 The general process to extract primary locations from Web pages

Then, we run the disambiguation procedure to assign a location for each GEO/GEO ambiguous geo-candidate and remove GEO/NON-GEO ambiguous geo-candidates. A location means a concrete geographic place in the world, e.g., USA/Washington, D.C. As a geo-candidate may refer to many locations in the world, the GEO/GEO disambiguation will decide which is the exact location that the geo-candidate refers to, and the GEO/NON-GEO disambiguation is going to determine whether it is a location or not. Finally, we present an effective algorithm to determine the primary locations among the resolved locations.

Geo-candidates Disambiguation

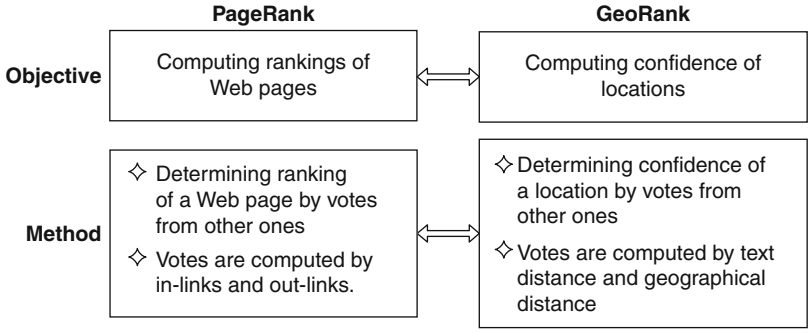
As Fig. 3 shows, we get a set of geo-candidates before the disambiguation procedure. We assume that all geo-candidates are associated with the locations in the Web page.

Basically, we assume there are n geo-candidates in a Web page and totally N locations that those geo-candidates may refer to. Then the GEO/GEO disambiguation problem can be formalized as follows:

Given a specific geo-candidate G, determining the most appropriate location among its possible locations.

We use a basic idea similar to PageRank (Brin and Page 1998) to resolve the GEO/GEO ambiguity, which is named GeoRank. The PageRank algorithm introduced an iterated voting process to determine the ranking of a Web page. We also regard the GEO/GEO disambiguation in a Web page as a voting process. Figure 4 shows the similar problem definition between PageRank and our GeoRank algorithm. Specially, in GeoRank, nodes are the possible locations corresponding to geo-candidates, and a linkage from one node A to another node B is marked with a score of evidence which represents A's voting for B to be the right location for the given geo-candidate.

In detail, as a geo-candidate can give more evidence to the one near to it in a Web page (text contribution) and a location can give more evidence to the one near to it in the geographic context (geographic contribution), we first construct a matrix M involving all locations (with each location occupies one row and one column),



Spatiotemporal Information for the Web, Fig. 4 PageRank vs. GeoRank

whose values are scores of each location of each geo-candidate voted by other ones that belong to different geo-candidates. This procedure is much like the voting process in PageRank, except that the items in M are locations but not Web pages and the scoring policy is based on text contribution and geographic contribution but not based on Web links.

Named entity recognition tools usually can remove some types of the GEO/NON-GEO ambiguities in a Web page. In order to get an improved performance, we propose two additional heuristics to further resolve GEO/NON-GEO ambiguities.

Rule 1: In the matrix M, if a location of a geo-candidate gets score averagely from all locations of other geo-candidates, it is not considered as a location, because none of any possible location of any other geo-candidate can give evidence to locations of this geo-candidate.

Rule 2: After removing the GEO/GEO ambiguity, if a non-country location does not have the same country with any other location, it is considered not a location. Here we get the rule from our observation that a Web page is unlikely to mention a non-country location that does not share a same country with any other locations.

Determining Primary Locations

In this stage, we calculate the scores of all the locations after disambiguation and then return the focused ones for the Web page. We consider three aspects when computing the scores of a location, namely, the term frequency, position, and geographic contributions (the contributions from locations geographically contained by the location). The motivation of the geographic contribution is that if there are many states of USA in a Web page, the location "USA" will receive contributions from those states, as those states are all geographically contained in the USA. As a result, we use an explicit score to represent the term frequency of a location name and an implicit score for the geographic contribution. The score of a location is determined by its explicit score and implicit score.

For a location D_i its explicit score, denoted as $ES(D_i)$, is defined as the term frequency of D_i in the Web page.

Then we use the following heuristics to modify $ES(D_i)$:

1. If D_i follows on the heels of the other location D_j and D_i has some relationship with D_j, suppose D_j is contained in D_i, then we think the appearance of D_i in the page will emphasize D_j, so we take 0.5 away from D_i and add it to D_j, i.e., $ES(D_i) = ES(D_i) - 0.5$, $ES(D_j) = ES(D_j) + 0.5$.
2. If D_i appears in the title of a Web page, then we add half of SUM to D_i to emphasize this appearance, where SUM is the sum of all the ES values, as defined in the formula (1):

$$\text{SUM} = \sum_{i=1}^{n} ES(D_i) \qquad (1)$$

For the implicit scores, since many locations that appear in one Web page usually have some geographic relationships, we take this feature when computing the implicit score of a location. In particular, we add some contributions from those locations contained by the given location into the score. Suppose a location D_i contains n sub-locations in the Gazetteer: $S_1, S_2, \ldots, S_n$, and the former m sub-locations appear along with D_i in the Web page, then those m sub-locations will provide geographic contributions to D_i. The implicit score of D_i is defined in the formula (2) and (3)

$$IS(D_i) = \sum_{k=1}^{m} (ES(S_k) + IS(S_k)) * \frac{m}{n * diff} \quad (2)$$

$$diff = \frac{\text{avg}(S_1, S_2, \ldots, S_m)}{\max(S_1, S_2, \ldots, S_m)} \quad (3)$$

Here, diff refers to the score difference among S_1, S_2,…,S_m. The average value of S_1, $S_2, \ldots, S_m$ must be less than or equal the maximum value of them, so $diff <= 1$. If D_i contains no sub-locations, then $IS(D_i) = 0$.

Based on a Gazetteer, we can build a hierarchy location tree. Then we start from the leaf nodes and compute the scores of all locations. After that, we sort all locations according to their scores and partition locations into three groups based on the scores. The first group with the highest scores is determined as the primary locations.

Extracting Primary Time from the Web

The various forms of temporal expressions in Web pages impose some challenging issues to temporal information extraction within the scope of Web search:

1. How to determine the right temporal information for implicit expressions contained in Web pages? Differing from the explicit expressions, which can be directly found in a calendar, the implicit expressions need a transformation process and usually a referential time is required.
2. How to determine the primary time for a Web page? A Web page may contain a lot of temporal information, but which ones are the most appropriate times associated with the Web page? This is very important to temporal-textual Web search engines which support both term-based and time-based queries, as they aim at finding "*the Web pages associated with the given terms and under the given temporal predicate*." For instance, to answer the query specifying "*finding the information about tourism during the National Day*," the search engines have to first determine which Web pages are mostly related with "*the National Day*."

For the first issue, namely, implicit time resolution, the difficult part is to select the referential time which is used to resolve implicit expressions. For example, to determine the exact time of the implicit expression "Yesterday" in a Web page, we must know the date of NOW under the context.

For the second issue, namely, primary time determination, the difficult part is to develop an effective scoring technique to measure the importance and relevance of the extracted temporal information. As there may be some containment relationship among temporal information, the time ranking task has to consider both frequency and the temporal containment. For instance, suppose "April, 2011" and "17 April, 2011" are two extracted time words, and "17 April, 2011" is contained in "April, 2011." Therefore, even "April, 2011" rarely appears in the Web pages, it will still be the primary time for the page in case that there are a great number of extracted time words contained by "April, 2011."

We focus on the above two issues and aim to propose effective solutions to the resolution of implicit expressions and the extraction of the primary time for Web pages. The main ideas can be summarized as follows:

1. We propose a new dynamic approach to resolve the implicit temporal expressions in Web pages. We classify the implicit expressions into global and local temporal expressions

and then use different methods to determine the referential time for global expressions and local expressions.
2. We present a score model to determine the primary time for Web pages. Our score model takes into account both the frequency of temporal information in Web pages and the containment relationship among temporal information.

Temporal Expressions Extraction

Temporal expressions in Web pages can be generally classified into two categories.

Explicit Temporal Expressions. These temporal expressions directly describe entries in some timeline, such as an exact date or year. For example, the token sequences "December 2004" or "September 12, 2005" in a document are explicit temporal expressions and can be mapped directly to chronons in a timeline.

Implicit Temporal Expressions. These temporal expressions represent tempo-ral entities that can only be anchored in a timeline in reference to another explicit or implicit, already anchored temporal expression. For example, the expression "today" alone cannot be anchored in any timeline. However, it can be anchored if the document is known to have a publication date. This date then can be used as a reference for that expression, which then can be mapped to a chronon. There are many instances of implicit temporal expressions, such as the names of weekdays (e.g., "on Thursday") or months (e.g., "in July") or references to such points in time like "next week" or "last Friday."

The explicit temporal expressions can be recognized by many time annotation tools, such as TempEx and GUTime (GUTime 2012). The temporal expressions in the GUTime output are annotated with TIMEX3 tags, which is an extension of the ACE 2004 TIMEX2 annotation scheme (tern.mitre.org).

For the extraction of implicit temporal expressions, the biggest difference of recognition between the explicit and implicit temporal expressions is that the implicit temporal expressions need to determine a reference time, so choosing the right reference is the key to the identification of the implicit temporal expression. The reference time can either be the publication time or another temporal expression in the document. Although the GUTime has a good performance in the extraction of explicit temporal expressions, it does not perform very well in dealing with the implicit temporal expressions, especially in the case of lacking of the document publication time. To improve the GUTime performance, we need to improve the reference-choosing mechanism of GUTime.

In this entry, we suppose that an implicit time expression consists of a modifier and a temporal noun which is modified by the modifier. For instance, a news report is as follows:

> (Beijing, May 6, 2009) B company took over A company totally on March 8, 2000″After one week, B company listed in Hong Kong, and became the first listed company in that industry. However, owing to the decision-making mistakes in the leadership and the company later poor management, B company got into debt for several hundred million dollars, and was forced to announce bankruptcy this Monday.

In this news report, "ten days" is a temporal noun, but "ten days ago" is modified after adding the modifier "ago." For the two temporal expressions that hold reference relations in this text, "after one week" and "this Monday," we can achieve the anchor direction easily from the modifiers through some mapping rules. Meanwhile, the offsets are able to understand directly by machine with pattern matching. But for the anchor points (referents), we must build the context-dependent reference reasoning to trace them. The full temporal reference comes from two parts: modifier reference and temporal noun reference. Because the former is inferred from the latter, the temporal noun reference reasoning plays more important roles in normalizations. Actually, we notice that the temporal noun can be classified into two classes according to the reference attributes. One is called Global Time (GT) whose temporal semantics is independent with the current context and takes the report time or publication time as the referent. Another one, Local Time (LT),

makes reference to the narrative time in text above on account of depending on the current context.

In our approach, there is a reference time table which is used to hold full reference time for the whole text, and we need to update and maintain it dynamically after each normalizing process. The time table consists of two parts: Global Reference Time and Local Reference Time.

Global Reference Time: Global Reference Time (GRT) is a type of reference time which is referred to by the Global Time. Specifically, it is the report time or the publication time of the document.

Local Reference Time: Local Reference Time (LRT) is referenced by the Local Time. It will be updated dynamically after each normalizing.

Different classes of time will dynamically and automatically choose references based on their respective classes rather than doing it using the fixed value or the inconsiderate rule under the static mechanism. And the reference time table is updated in real time finishing each normalizing, which makes the temporal situation compliable with dynamically changeable contexts.

Determining the Primary Time

We proposed a score model to calculate the score of each temporal expression. In detail, we consider two aspects when calculating the score of a temporal expression, namely, the term frequency of the temporal expression and the relevance between temporal expressions. It is easy to understand that the term frequency is related to the score of a temporal expression. Here we focus attention on introducing the relevance between temporal expressions. We make an assumption that there is an article which contains some temporal expressions, and most of them refer to a certain day in March. In this case, we tend to choose March as the primary time rather than any one of them. Based on this view, we think that a temporal expression will make a contribution to its parent temporal expression. For example, the expression March 7, 2012 makes a contribution to its parent expression March 2012, and the expression 1983 contribute to its parent expression 1980s.

Here, we define the score of a temporal expression as a combination of an explicit score and an implicit score. The explicit score is related to the term frequency of a temporal expression, and accordingly, the implicit score is related to the contribution made by all its children expressions. The score of T_i, denoted as $S(T_{\neg i})$, is the sum of its explicit score, denoted as $ES(T_i)$, and its implicit score, denoted as $IS(T_i)$.

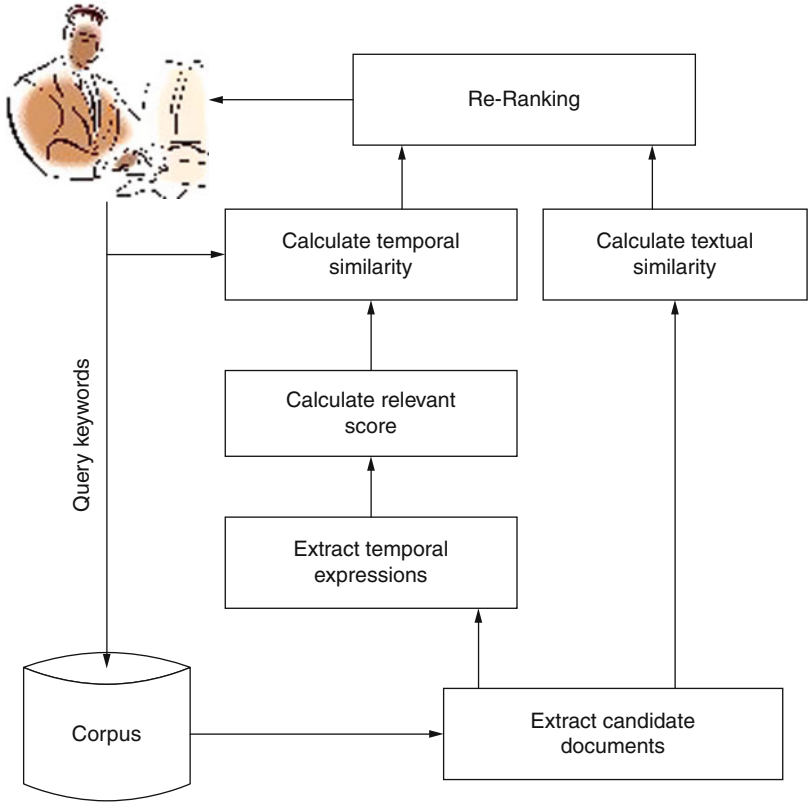
The explicit score $ES(T_i)$ is defined as the term frequency of Ti in the article. As compared to implicit temporal expressions, the explicit temporal expressions are more accurate in the extraction. In other words, the explicit temporal expressions are more credible, so we add a weighting factor d to the implicit temporal expressions. The explicit score of T_i is defined as formula (4):

$$ES(T_i) = TF_{ETE}(T_i) + d * TF_{ITE}(T_i) \quad (4)$$

Here, $TF_{ETE}(T_i)$ refers to the term frequency of the explicit temporal expressions which are recognized as T_i. $TF_{ITE}(T_i)$ refers to the term frequency of the implicit temporal expressions which are calculated as T_i. d is the weighting factor; if d is set to 1, it means that the explicit and implicit temporal expression have the same credible level; if d is set to 0, it means that we take no account of implicit temporal expressions.

The implicit score $IS(T_i)$ is related to all the scores of its children, we denoted as C_1, C_2, …, C_n, respectively, and we use the symbol N to represent the number of children that T_i contains. For example, if the granularity of T_i is MONTH, then the value of N is 30 because a month contains about 30 days. Likewise, if the granularity of T_i is QUARTER, the value of N should be 3 because a quarter contains 3 months. Here, we use the factor α to represent how much contribution the children of T_i make. So the implicit score $IS(T_i)$ can be defined as formula (5):

$$IS(T_i) = \frac{1}{\alpha \times N} \sum_{i=1}^{n} S(C_i) \quad (5)$$



Spatiotemporal Information for the Web, Fig. 5 The architecture of TASE

Finally, we can compute the scores of each time expression based on its explicit score and implicit one and then choose the Top-K time expressions as the primary time of the Web page.

Illustrative Example

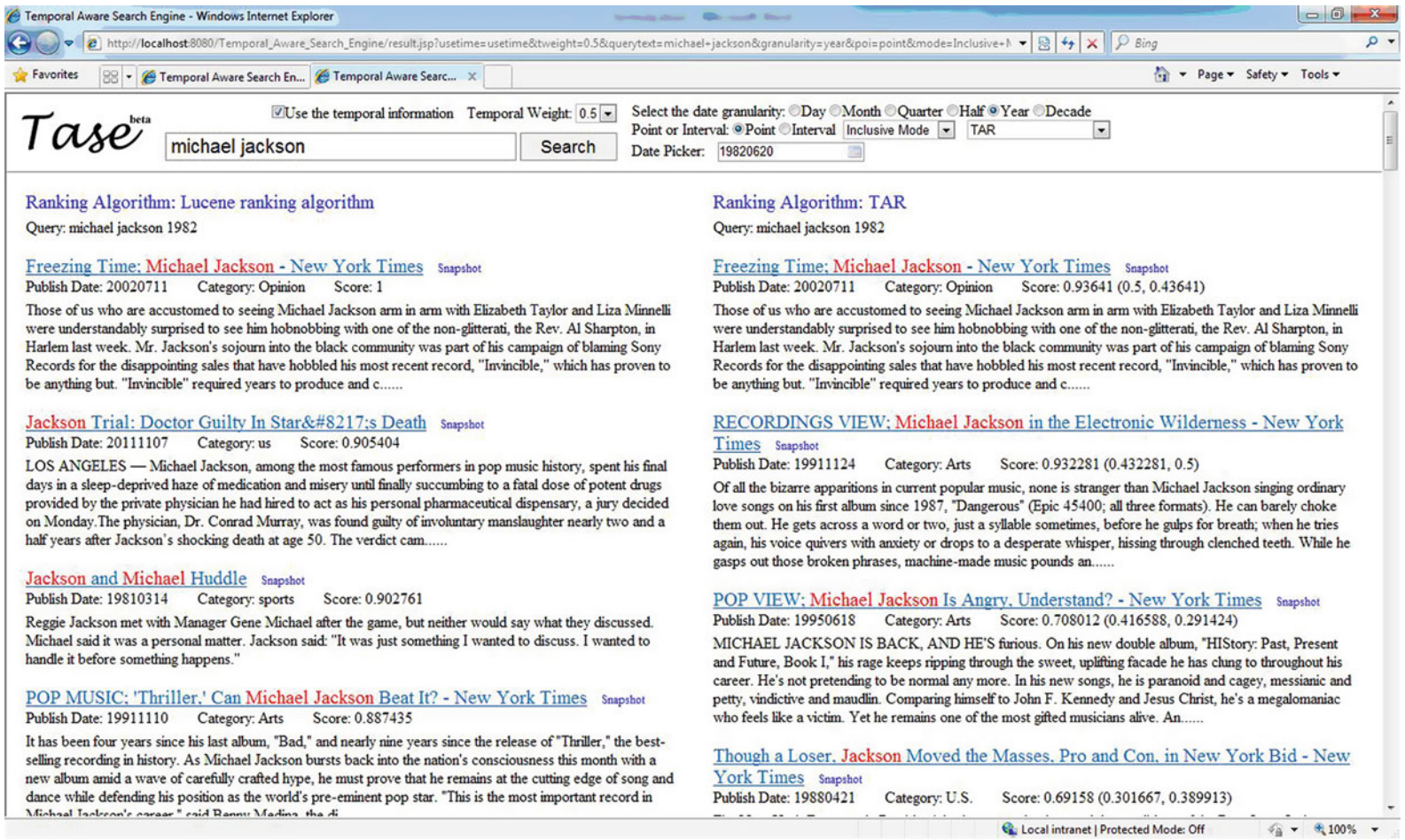
In order to show the usability of spatiotemporal information in Web search, we present and implement a prototype system for temporal-sensitive queries called TASE (Time-Aware Search Engine) (Lin et al. 2012). The major features of TASE can be described as follows:

1. TASE extracts the temporal expressions for each Web page and calculates the relevant score between the Web page and each temporal expression. Compared with traditional approaches, TASE uses a new reference time dynamic-choosing approach to extract implicit temporal expressions in Web pages. Besides, it distinguishes the temporal expressions with their relevant score and takes the containment relationship among the temporal expressions into consideration.
2. TASE combines the temporal similarity and the textual similarity to re-rank the search results. Our experiments demonstrate its effectiveness in dealing with temporal-sensitive Web queries.

Figure 5 shows the architecture of TASE, and the interface of TASE is shown in Fig. 6. The four major modules in TASE are described as follows:

Extract Candidate Documents. This module extracts the original Top-K documents from the search results which are used as the candidate documents.

Extract Temporal Expressions. This module extracts all the temporal expressions in each candidate document, including the explicit temporal expressions and the implicit temporal expressions.



Spatiotemporal Information for the Web, Fig. 6 The interface of TASE

Calculate Relevant Score. The relevant score between a temporal expression and a Web page will be calculated in this module.

Calculate Temporal Similarity. It calculates the similarity between the temporal expressions in a query and a document.

Calculate Textual Similarity. TASE is built on Lucene, an open-source search engine. Therefore, we use the textual similarity determined by Lucene as the original textual similarity.

Re-ranking. In this module, it used the temporal similarity and the original textual similarity to determine the final relevant score of a document.

Key Applications

Spatiotemporal information presented here may be used in many applications. First of all, spatiotemporal information can be used in search engines to improve the quality of results. By designing spatiotemporal indexes and ranking algorithms, search engines can be enhanced to process time-and-location-related Web queries effectively and efficiently. Secondly, our method is also useful in focused search engines, such as news search, product search, or stock search. In such applications, time and location information play an important role and our approach can be applied to offer better solutions to the search needs. Thirdly, spatiotemporal information can be utilized in question answering or automatic summarization in the Web. Many questions in the Web are related with time and location, which can be answered if we extract facts as well as their associated time and locations. For the automatic summarization, as events are usually described along a timeline, so it can be well done with the help of the extracted time of the specified topic.

Future Directions

This work can be extended to spatiotemporal analysis and mining in Web data, which may bring values for Web knowledge discovery. As Web has been regarded a major source of competitive intelligence, how to acquire competitive intelligence from the Web has been a hot topic. By using spatiotemporal information, we are able to find some historical information about interested competitors and further detect their future strategic planning in the near future. Spatiotemporal information can also be used to measure the credibility of Web information. With the rapid development of Web 2.0 and social network applications, there are many fakes and false information in the Web, which will introduce a lot of risks in decision making and other applications. Though information credibility involves many aspects of factors, spatiotemporal information can be used as one type of measurement to validate the credibility of specific information. For example, when we want to determine the credibility of a piece of news reporting "Apple iPhone 5 has been released," we can collect the Web pages or microblogs mentioning the news and perform spatiotemporal clustering process to detect its credibility.

Acknowledgments

We would like to thank the University of Science and Technology of China for providing the environment where the study described in this entry was completed. The work involved in this article is partially supported by the National Science Foundation of Anhui Province (NO. 1208085MG117) and the USTC Youth Innovation Foundation.

Cross-References

▸ Modeling and Analysis of Spatiotemporal Social Networks
▸ Social Web Search
▸ Spatiotemporal Personalized Recommendation of Social Media Content
▸ Spatiotemporal Reasoning and Decision Support Tools

References

Brin S, Page L (1998) The anatomy of a large-scale hyper textual web search engine. In: Proceedings of WWW, Brisbane, pp 107–117

Chinchor N (1998) MUC-7 information extraction task definition, version 5.1. In: Proceedings of the 7th message understanding conference (MUC-7), Fairfax

Ding J, Gravano L, Shivakumar N (2000) Computing geographical scopes of web resources. In: Proceedings of VLDB, Cairo, pp 545–556

GUTime (2012) http://www.timeml.org/site/tarsqi/modules/gutime/index.html. Accessed Aug 2012

Lee R et al (2003) Optimization of geographic area to a web page for two-dimensional range query processing. In: Proceedings of fourth international conference on web information systems engineering workshops (WISEW 2003), Roma. IEEE Computer Society, pp 9–17

Lin S, Jin P, Zhao X, Yue L (2012) TASE: a time-aware search engine. In: Proceedings Of CIKM'12, Maui. ACM

Ma Q, Tanaka K (2004) Retrieving regional information from web by contents localness and user location. In: Proceedings of AIRS, Beijing, pp 301–312

Markowetz A, Chen Y, Suel T, Long, X, Seeger B (2005) Design and implementation of a geographic search engine. Technical report TR-CIS-2005-03, Polytechnic University, Brooklyn

Nunes S, Ribeiro C, David G (2008) Use of temporal expressions in web search. In: Proceedings of ECIR'08, Glasgow, pp 580–584

Sanderson M (2000) Retrieving with good sense. Inf Retr 2(1):45–65

Sanderson M, Kohler J (2004) Analyzing geographic queries. In: Proceedings of GIR'04, Sheffield. ACM

Setzer A, Gaizauskas R (2002) On the importance of annotating event-event temporal relations in text. In: Proceedings of LREC'02, Paris

Sundheim B, Chinchor N (1995) Named entity task definition, version 2.0. In: Proceedings of the 6th message understanding conference (MUC-6), Columbia. Morgan Kaufman, pp 319–332

Wang C, Xie X et al (2005) Web resource geographic location classification and detection. In: Proceedings of WWW'05, Chiba. ACM

Zhou Y, Xie X, Wang C et al (2005) Hybrid index structures of location-based web search. In: Proceedings of CIKM'05, Bremen

Spatiotemporal Outlier

▸ Detection of Spatiotemporal Outlier Events in Social Networks

Spatiotemporal Personalized Recommendation of Social Media Content

Bee-Chung Chen
LinkedIn, Mountain View, CA, USA

Synonyms

Location-based recommendation; Positional or layout effect in recommender systems; Spatiotemporal collaborative filtering; Time-sensitive recommendation

Glossary

Recommender A system that recommends items (e.g., news articles, blog posts) to users

Response Rate The probability that a user would respond to (e.g., click, share) a recommended item

Feature Information (about a user, an item, and the context in which the item may be recommended to the user) that can be used to predict the response rate

Page A web page on which recommended items are placed

Context The situation (which includes time, geographical location, location of a web page, etc.) in which recommendations are made to a user

Graph A set of nodes connected by a set of edges

Definition

Social media sites (like twitter.com, digg.com, blogger.com) complement traditional media by incorporating content generated by regular people and allowing users to interact with content through sharing, commenting, voting, liking, and other actions. Since the number of content items is usually too large for a person to manually examine to find interesting ones, it is important

for social media sites to recommend a small set of items that are worth looking at for each user. To satisfy each individual user, recommended items have to match the user's personal interests and be relevant to the user's current spatiotemporal context. For example, a content item about the user's hometown is usually a better choice than an item about an unknown foreign country, and a content item on a fresh trending topic is usually more interesting than an item on a stale topic.

Spatiotemporal personalized recommendation of social media content refers to techniques used to make personalized recommendation based on:

- The geographical location of a user and an item (the location of an item can be the location that the item is about or the location of the author of the item)
- The location of a user in the social space (e.g., the neighborhood of a user in a friendship graph)
- The position of an item placed on a page and the layout of the page
- Temporal evolution of user interests
- Temporal behavior of the popularity of an item
- Identification of trending topics

Introduction

Social media usually refers to a group of Internet-based applications that allow creation and exchange of user-generated content (Kaplan and Haenlein 2010). For example, weblog sites like blogger.com provide regular people the ability of publishing any article (called blog) on the web, microblogging sites like twitter.com facilitate fast distribution of short messages of any topic posted by any one, and social news sites like digg.com allow their users to vote news articles (and other web content) up or down in order to present popular and interesting news stories based on the wisdom of the crowd (i.e., votes from users), just to name a few. Because of the success of such social media sites, almost all online media sites now provide their users with the functionality of sharing and commenting on content items (e.g., news articles, photos, songs, movies), no matter whether the content items are generated by regular users. Since sharing and commenting are usually considered as social activities, the distinction between social media and online media (which includes social media) blurs. Thus, in this article, we discuss recommendation methods suitable for any online media with a special emphasis on spatial, temporal, and social characteristics of users and content items.

The large amount of content generated by social media makes it difficult for users to find personally relevant content. To alleviate such information overload, many social media sites recommend a small set of content items to each user based on what they know about the user and the items. We use the term "item" to refer to any candidate objects to be recommended to users, which include (but are not limited to):

- Publisher-generated items like articles, songs, and movies, which are not generated by regular users, but are voted, shared, liked, or commented on by them
- User-generated items like blogs, tweets (short messages posted on twitter.com), photos, videos, status updates, and comments on other items

Good recommendations help social media sites keep their users engaged and interested.

Key Points

When recommending items to users, it is important to consider whether an item is relevant to a user in the *spatiotemporal context* in which recommendations are to be made. A few key reasons are listed below. Notice that we take a broad view of the spatial aspect that includes locations in geographical space, social space, and positions on a web page:

- Users are likely to be more interested in items about their current geographical location than items about a random location, which is especially true for mobile applications (see, e.g., Zheng et al. (2010)).
- In some applications, users tend to have similar preferences to those who are close to them in the social space, which is especially true

when closeness is defined based on a trust network (see, e.g., Jamali and Ester (2010)).
- It is generally true that items placed at prominent positions (e.g., top) on a page generate more responses from users than same items placed at non-prominent positions (see, e.g., Agarwal et al. (2009)).
- Users change their interests in topics over time (see, e.g., Ahmed et al. (2011)).
- Popularity of items also changes over time (see, e.g., Agarwal et al. (2009)).

Many methods have been developed to exploit these spatiotemporal characteristics to improve the performance of recommenders. A comprehensive review of these methods is beyond the scope of this article. Instead, after providing a brief historical background, we illustrate key ideas in spatiotemporal personalized recommendation through a generic supervised learning approach, which handles spatiotemporal characteristics by (1) defining features that capture those characteristics and (2) learning a function that predicts whether a user would respond to an item positively based on these features from a dataset that records users' past responses to items. This approach generally applies to recommendation of any kind of item.

Historical Background

There have been many approaches developed to make personalized recommendations. When the items to be recommended are text articles, which may be represented as a bag of words, an early approach is to also represent a user as a bag of words. The user's bag of words can be constructed by including representative words in the articles that the user likes to read. Then, we can recommend a user the articles which bags of words are most similar to the user's bag of words through Salton's vector space model (Salton et al. 1975). For items that are not easily representable as bags of words, how other users respond to an item may provide a clue as to whether to recommend the item to a user who has not yet responded to the item. Agrawal et al. (1993) proposed that, in a retail store setting, products can be recommended based on customers' co-buying behavior. For example, if the majority of customers who buy product A also buy product B, then we may recommend product B to a customer who only bought product A. This idea was then extended by incorporating a notion of similarity of users or items. For example, when we decide whether to recommend item B to user i, we look at whether users "similar" to user i respond to item B positively. Notice that Agrawal's method is based on the similarity definition that if two customers buy the same product, then they are similar. A different definition of similarity between users leads to a different method. Furthermore, we can also exploit similarity between items in a similar way – when deciding whether to recommend item B to user i, check whether user i liked items that are "similar" to B in the past. Here, similarity between two items can be defined by looking at whether most users responded to the two similarly. Adomavicius and Tuzhilin (2005) provided a good review of such methods. This kind of methods is generally referred to as collaborative filtering, because the recommendations that a user receives depend on other users' responses to candidate items – this process can be thought of as a collaboration among users to help one another find interesting items (although users may not be aware of the collaboration).

Conceptually, one can put users' past responses to items into a matrix. Since this matrix-oriented approach is popular in movie recommendation (Koren et al. 2009), we use it as an example in the following discussion. In a movie recommender system, users rate movies. Let y_{ij} denote the rating that user i gives to movie j. For example, y_{ij} may be a numeric value ranging from 1 to 5, representing 1 star to 5 stars. Let Y denote the $m \times n$ matrix such that the value in the (i, j) entry is y_{ij}, where m is the number of users and n is the number of movies in the system. Notice that there are many entries with missing (i.e., unknown) values in matrix Y because most users only rate a small number of movies. For user i, if we can predict the missing values in the ith row of matrix Y accurately (where the entries with missing values correspond to movies

that have not yet been rated by user i and are thus candidate items to be recommended to him/her), then we can recommend user i the movies having the highest predicted rating values. One popular way of making such predictions is through matrix factorization – approximate matrix Y as the product UV' of two low rank matrices U of size $m \times r$ and V of size $n \times r$, where V' denotes the transpose of matrix V and the rank r of matrices U and V is much smaller than the numbers m and n of users and items, respectively. Let $\boldsymbol{u}_i$ denote the ith row of matrix U, $\boldsymbol{v}_j$ denote the jth row of matrix V, and $\Omega = \{(i, j) : \text{user } i \text{ rated movie } j\}$ denote the set of observed entries in matrix Y. This approximation then can be mathematically formulated as the following optimization problem.

$$\text{Find } U \text{ and } V \text{ that minimize } \sum_{(i,j)\in\Omega} (y_{ij} - \boldsymbol{u}_i' \boldsymbol{v}_j)^2, \tag{1}$$

where $\boldsymbol{u}_i' \boldsymbol{v}_j$ is the inner product of two vectors $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$. Notice that $\boldsymbol{u}_i' \boldsymbol{v}_j$ is the (i, j) entry of matrix (UV') and is also the predicted value of y_{ij}. Thus, the above optimization seeks to minimize the difference between matrix Y and matrix (UV') over only the set Ω of observed entries of Y. Sum of squared differences is a common choice, while other choices are also available for different problem settings. Recent studies, such as Koren et al. (2009), Agarwal and Chen (2009), and many others, suggest that matrix factorization usually provides superior recommendations than more traditional methods.

A survey of a wide range of approaches to recommender systems can be found in Jannach et al. (2010) and Ricci et al. (2011). Here, we focus on how to make use of spatial, temporal, and social information to make good recommendations of social media content. In particular, we illustrate key ideas in spatiotemporal personalized recommendation through a general supervised learning (or statistical modeling) approach, which generally applies to recommendation of any kind of item.

Supervised Learning Approach

In general, a recommendation problem can be formulated as follows. A recommender is given:

- A user, who is associated with a vector of *user features*, e.g., age, gender, and location
- A context, which is associated with a vector of *context features*, e.g., day of week when the recommendation is to be made
- A set of candidate items, each of which is associated with a vector of *item features*, e.g., topics and keywords

The goal of the recommender is to rank and pick the top few items from the set of candidate items that best "match" the user's interests and information need in the context. The supervised learning approach exploits the fact that, in many recommenders, a dataset of users' past responses (e.g., click, share) to items can be collected and defines the degree that an item matches a user as the response rate of the user to the item (e.g., the probability that the user would click the item if he/she sees the item on a web page). Such predictions can be made by using a statistical (regression or machine learning) model, which "learns" the user and item behavior that allows accurate predictions from the dataset, where users' past responses in the dataset "supervise" the learning process via giving desired (e.g., click) and undesired (e.g., no click) examples. When such a model is available, recommendations for a user can be made by picking the top few items having the highest response rates among the set of candidate items. This supervised learning approach applies to recommendation of any kind of item, where spatiotemporal and other characteristics can be incorporated by defining features that capture those characteristics.

To use this supervised learning approach, a developer of a recommender needs to make the following three decisions:

- What response should the model try to predict?
- What features should the model use to capture the characteristics of users, items, and the spatiotemporal context?
- What class of model do we want to use?

After introducing a running example, we discuss how to choose the response, provide a number of useful features, and then introduce two commonly used classes of models, namely feature-based regression model and latent factor model. See Jannach et al. (2010) and Ricci et al. (2011) for other classes of models. See Hastie et al. (2009) for a general introduction to supervised learning.

Example Recommender

For concreteness, we use blog article recommendation as a running example. Consider that we want to develop a recommender for a blog service provider (e.g., blogger.com) that seeks to recommend each user with a set of interesting blog articles posted by other users. To make modeling more interesting, assume that a user can declare friendship with other users and such friendship connections between users are available to the recommender. In this example, the set of candidate items for each user consists of all of the articles posted within a 1-week time window (to ensure freshness) by any user of this service provider. Notice that the set of candidate items changes over time. For simplicity, we only need to recommend 10 articles for each user, once per day, and the recommended articles are displayed in a list on the sidebar of each user's homepage (they are only visible to the owner of the homepage, not the visitors of the homepage, since the recommendations are made to the owner).

Choice of Response

The choice of response depends on the objective that a recommender is developed for and availability of user feedback that the recommender receives. A common objective is to maximize clicks on recommended items because the fact that a user clicks an item indicates that the user is interested in knowing more about the item. Note that clicks are user feedback that can easily be made available to a recommender through logging whether each user clicks the recommended items. In this case, a natural choice of the response is whether a user would click an item if he/she sees the item being recommended. Here, the goal of learning is to predict the probability that a user would click an item based on a dataset that records what items each user clicked and what items each user did not click in the past.

Beyond clicks, a recommender may be developed for other objectives. For example, if the objective of recommendation is to encourage users to make comments on recommended items, then a natural choice of the response would be whether a user would comment on a recommended item or not. On some sites, users can explicitly rate items (e.g., using one star to five stars); then, a natural choice of the response would be the rating that a user would give to an item. For simplicity, we only consider methods that seek to achieve a single objective and model the response rate of a single type of choice (e.g., modeling either click rate or explicit star rating, but not both). See Agarwal et al. (2011a) for an example of multi-objective recommendation, and see Agarwal et al. (2011b) for an example of joint modeling of multiple types of responses.

Let y_{ijk} denote the response that user i gives to item j in context k. For concreteness, assume that we choose to model whether the user would click the item.

Feature Engineering

Having good features is essential to an accurate model, but one usually does not get good features automatically. It requires domain knowledge, good intuition, and experience in the application to define good features. Here, for illustration purposes, we only show a number of example features that can potentially capture different kinds of spatiotemporal characteristics for our example recommender. Real-life recommenders usually need to use much more features than the following ones.

User Features

Let $\boldsymbol{w}_i$ denote the vector of features of user i. For simplicity, we mostly consider binary features, meaning each element in the vector is either 0 or 1. Example features are as follows:

- **Gender:** From the user's registration record when he/she signed up on the site, the recommender obtains the gender of each user.

The numeric value of the feature is 1 if the user is a male and 0 if the user is a female.

- **Age:** Also from the user's registration record, the recommender obtains the age of each user. For example, we can group age values into 10 age groups, which give 10 age features. If the user's age is in an age group, the value of the feature corresponding to that age group is 1, and the rest age groups get feature value 0.
- **City:** From the IP address of a user, the recommender can guess the city that the user is in. Here, we use a set of features, one for each city, to represent the user's geographic location. For example, assume the user lives in New York City. Then, the value of the New York City feature is 1 and the values of the rest of the city features are all 0 for the user. It is common to only include cities that have at least n users, where n is a threshold that a developer of the recommender can choose to reduce the number of features.

Item Features

Let $\boldsymbol{x}_j$ denote the vector of features of item j. Example features are as follows:

- **Bag of words:** It is common to represent the text content of an article as a bag of words, which corresponds to a set of features, one for each keyword. For simplicity, we only consider binary keyword features. The value of a keyword feature is 1 if the article contains the keyword and 0 if the article does not contain the keyword. Since the total number of words in all articles is usually too large, it is also common to reduce the space of all keywords to a relatively small number of important words, e.g., location names or other named entities.
- **Topics:** Another way to reduce the space of words in articles is to group words into topics and then assign topics to articles based on the words in articles. This process can be automated through topic models like latent Dirichlet allocation (Blei et al. 2003). One output from such a model is a vector of topic membership for each article, where each element in the vector represents the probability that the article is about a particular topic.

Context Features

Let z_{ijk} denote the vector of features of the context in which user i is (to be) recommended with item j in context k (which include time and location). Example features are as follows:

- **Day of week:** This is the day of week (weekday vs. weekend) when the recommendation is to be made. User behavior during the weekday can be quite different from that during the weekend. The value of this feature is 1 for weekday and 0 for weekend.
- **Article age:** This is the age of an article (not to be confused with the age of a user), which is the number of days since the article was posted. We put it into the category of context features, instead of item features, because it depends on both the article and time, instead of the article alone. For example, assume the article was posted 2 days ago; then, the value of the feature corresponding to 2 days ago is 1, and the other days get feature value 0. To model finer-grained temporal effect, one may choose a finer time resolution (e.g., hour, instead of day).
- **Position on page:** It is well known that the click rate of an item put on the top of a list on a page is usually higher than that of the same item put in the middle or the bottom of the page. To capture this positional bias, we define a set of features, each of which corresponds to a position in the list. For example, assume the article is put at the third position, the value of the feature corresponding to the third position is 1 and all other positions have feature value 0.
- **Friendship:** This feature is 1 if user i is connected to the author of item j through a friendship connection and is 0 otherwise.
- **Same city:** This feature is 1 if user i is in the same city as the author of item j and is 0 otherwise.

Note that the above features are only simple examples. The goal here is to provide concrete examples of features for illustration purposes, instead of suggesting good features for practical implementation.

Feature-Based Regression Model

After defining the response and features, we have a standard supervised learning problem. When the response is binary (e.g., either click or no click), we can use logistic regression. See Hastie et al. (2009) for an introduction to logistic regression. Let p_{ijk} denote the probability that user i would respond to item j when he/she sees it in context k. There are many ways in which one can define a function that predicts p_{ijk} based on features. A useful prediction function is as follows:

$$p_{ijk} = \sigma(\boldsymbol{w}_i' \boldsymbol{A} \boldsymbol{x}_j + \boldsymbol{\beta}' \boldsymbol{z}_{ijk}), \qquad (2)$$

where $\sigma(a) = \frac{1}{1+\exp(-a)}$ is the sigmoid function that transforms an unbounded value a into a number between 0 and 1 (since p_{ijk} is a probability), $\boldsymbol{A}$ is a regression coefficient matrix, $\boldsymbol{\beta}$ is a regression coefficient vector, and $\boldsymbol{w}_i'$ and $\boldsymbol{\beta}'$ are the row vectors after transposing the two-column vectors $\boldsymbol{w}_i$ and $\boldsymbol{\beta}$, respectively. Given a dataset of users' past responses to items, where each record is in the form $(y_{ijk}, \boldsymbol{w}_i, \boldsymbol{x}_j, \boldsymbol{z}_{ijk})$, off-the-shelf logistic regression packages can be applied to learn the regression coefficients $\boldsymbol{A}$ and $\boldsymbol{\beta}$.

To better understand this model, we take a closer look at the prediction function. Let A_{mn} denote the (m, n) entry of matrix A, w_{im} denote the mth user feature in vector $\boldsymbol{w}_i$, and x_{jn} denote the nth item feature in vector $\boldsymbol{x}_j$. By definition, we have

$$\boldsymbol{w}_i' \boldsymbol{A} \boldsymbol{x}_j = \sum_m \sum_n A_{mn} w_{im} x_{jn}. \qquad (3)$$

For example, assume w_{im} is the feature that indicates whether user i lives in New York City and x_{jn} is the feature that indicates whether article j contains keyword "new york." Then, the regression coefficient A_{mn} would try to capture the propensity that users living in the New York City would click an article that contains keyword "new york" after adjusting for all other factors. Now, assume that the mth and nth context features in $\boldsymbol{z}_{ijk}$ indicate whether article j is posted 1 day ago and whether j is posted 5 days ago, respectively. Then, the difference between regression coefficients $\beta_m - \beta_n$ would try to quantify how much the popularity of an article drops from day 1 to day 5 when all other conditions being equal.

Latent Factor Model

Although feature-based regression models are useful for predicting users' response rates to items, they depend highly on the availability of predictive features, which usually requires a significant feature engineering effort with no guarantee of obtaining predictive features. Also, feature vectors may not be sufficient to capture the differences between users or items. For example, when two users have identical feature vectors, feature-based regression models would be unable to tell the differences between the two. One way of addressing these issues is to add *latent factors* into the prediction function; i.e.,

$$p_{ijk} = \sigma(\boldsymbol{w}_i' \boldsymbol{A} \boldsymbol{x}_j + \boldsymbol{\beta}' \boldsymbol{z}_{ijk} + \boldsymbol{u}_i' \boldsymbol{v}_j), \qquad (4)$$

where $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ are two r-dimensional vectors both to be learned from data like regression coefficients $\boldsymbol{A}$ and $\boldsymbol{\beta}$, where r is much smaller than the number of users and the number of items. Recall that we have seen $\boldsymbol{u}_i' \boldsymbol{v}_j$ in the matrix factorization method in the historical background section. The difference is that, instead of factorizing the response matrix, here we factorize the residual (i.e., prediction error) matrix of feature-based regression in order to capture the behavior of users and items that the features fail to capture.

Intuitively, one can think of $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ as "latent feature" vectors of user i and item j, respectively. We do not determine the values of these r latent features per user or item before learning the model. Instead, $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ are treated as variables that can be used to reduce the error of predicting the responses in the dataset used for learning. The inner product $\boldsymbol{u}_i' \boldsymbol{v}_j$ then represents the affinity between user i and item j; the larger the inner product value, the higher the probability that user i would click item j. After the learning process, we simultaneously obtain the values of these latent features and also the regression coefficients $\boldsymbol{A}$ and $\boldsymbol{\beta}$. See Agarwal et al. (2010) for an example of such a latent factor model.

Spatiotemporal contexts can also be involved in a latent factor model. For example, assume we

want to model a temporal effect through latent factors. Let context index k represent the kth time period (e.g., day). One way of capturing user or item behavioral changes over time is through the following model:

$$p_{ijk} = \sigma(\boldsymbol{w}_i' \boldsymbol{A} \boldsymbol{x}_j + \boldsymbol{\beta}' \boldsymbol{z}_{ijk} + \langle \boldsymbol{u}_i, \boldsymbol{v}_j, \boldsymbol{t}_k \rangle), \quad (5)$$

where $\langle \boldsymbol{u}_i, \boldsymbol{v}_j, \boldsymbol{t}_k \rangle = \sum_\ell u_{i\ell} v_{j\ell} t_{k\ell}$ is a form of tensor product of three vectors $\boldsymbol{u}_i$, $\boldsymbol{v}_j$ and $\boldsymbol{t}_k$. Note that $u_{i\ell}$ denotes the ℓth element of vector $\boldsymbol{u}_i$ and so on. Similar to the previous model, $\boldsymbol{u}_i$, $\boldsymbol{v}_j$, and $\boldsymbol{t}_k$ are all latent feature vectors, which values are to be learned from data. Unlike the previous model where the affinity $\boldsymbol{u}_i' \boldsymbol{v}_j$ between user i and item j is fixed over time, now the affinity $\langle \boldsymbol{u}_i, \boldsymbol{v}_j, \boldsymbol{t}_k \rangle$ is a function of time period k, which means this model captures the changing behavior of user-item affinity. Specifically, in this model, the user and item latent feature vectors are fixed over time, but the affinity between the two is a weighted sum of the element-wise product of the two latent feature vectors $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$, where the weight vector $\boldsymbol{t}_k$ changes over time. See Xiong et al. (2010) for an example of such a temporal latent factor model.

Summary

Personalized recommendation is an important mechanism for surfacing social media content. The spatiotemporal context in which a recommendation is made provides a key piece of information that helps a recommender to recommend the right item to the right user at the right time. While many methods have been proposed in the literature, the supervised approach is attractive because of its generality, where spatiotemporal characteristics can be incorporated as features or latent factors. In this article, we introduced a number of example features and two example models. In practice, many features need to be evaluated and a number of different models need to be tried, so that a good recommender can be built.

Future Directions

Personalized content recommendation is currently an active research area in data mining, information retrieval, and machine learning. A lot of progress has been made in this area, but challenges remain.

- *Improving response rate prediction accuracy*: Although many models have been proposed to predict response rates and we have seen prediction accuracy improve over time, accurate prediction of the probability that a user would respond to an item is still a challenging problem, especially for users and items that the recommender knows little about. What are the spatial, temporal, social, and other kinds of features that can further improve accuracy? How can a recommender actively collect data to achieve better model learning and evaluation?
- *Multi-objective optimization*: A recommender usually is designed to achieve multiple objectives. For example, many web sites put advertisements on article pages to generate revenue. In addition to recommend articles that users like to click, we may also want to recommend articles that can generate high advertising revenue. How can a recommender optimize multiple objectives in a principled way?
- *Multi-type response modeling*: In social media, users respond to items in multiple ways, e.g., clicks, shares, tweets, emails, and likes. How can we jointly model such different types of user responses in order to find out the items that a user truly want to be recommended?
- *Whole-page optimization*: On a web page, there can be multiple recommender modules. For example, one recommends news articles, another recommends updates from a user's friends, and yet another recommends online discussions the user may be interested in. How can we jointly optimize multiple recommender modules on a page to leverage the correlation among modules and to ensure consistency, diversity, and serendipity?
- *Collaborative content creation*: Wikipedia demonstrated high-quality content creation through massive collaboration. However, in most recommender systems, items to

be recommended are created by a single party (e.g., a publisher or a user). How can we synthesize items at the right level of granularity to recommend to users in a semi-automatic collaborative way?

Cross-References

- ▸ Data Mining
- ▸ Friends Recommendations in Dynamic Social Networks
- ▸ Link Prediction
- ▸ Matrix Decomposition
- ▸ Mining Trends in the Blogosphere
- ▸ Probabilistic Graphical Models
- ▸ Recommender Systems: Models and Techniques
- ▸ Recommender Systems Using Social Network Analysis: Challenges and Future Trends
- ▸ Recommender Systems, Semantic-Based
- ▸ Regression Analysis

References

Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng 17:734–749

Agarwal D, Chen BC (2009) Regression-based latent factor models. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. In: KDD'09, Paris, France, pp 19–28

Agrawal R, Imieliński T, Swami A (1993) Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD international conference on Management of data. In: SIGMOD'93, Washington, DC, USA, pp 207–216

Agarwal D, Chen BC, Elango P (2009) Spatio-temporal models for estimating click-through rate. Proceedings of the 18th international conference on World wide web. In: WWW'09, Madrid, Spain, pp 21–30

Agarwal D, Chen BC, Elango P (2010) Fast online learning through offline initialization for time-sensitive recommendation. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. In: KDD'10, Washington, DC, USA, pp 703–712

Agarwal D, Chen BC, Elango P, Wang X (2011a) Click shaping to optimize multiple objectives. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, KDD'11. ACM, New York, pp 132–140. doi:10.1145/2020408.2020435, http://doi.acm.org/10.1145/2020408.2020435

Agarwal D, Chen BC, Long B (2011b) Localized factor models for multi-context recommendation. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, KDD'11. ACM, New York, pp 609–617. doi:10.1145/2020408.2020504, http://doi.acm.org/10.1145/2020408.2020504

Ahmed A, Low Y, Aly M, Josifovski V, Smola A-J (2011) Scalable distributed inference of dynamic user interests for behavioral targeting. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, KDD'11. ACM, New York, pp 114–122. doi:10.1145/2020408.2020433, http://doi.acm.org/10.1145/2020408.2020433

Blei DM, Ng AY, Jordan MI (2003) Latent Dirichlet allocation. J Mach Learn Res 3:993–1022. http://dl.acm.org/citation.cfm?id=944919.944937

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer, New York, NY

Jamali M, Ester M (2010) A matrix factorization technique with trust propagation for recommendation in social networks. In: Proceedings of the fourth ACM conference on recommender systems, RecSys'10. ACM, New York, pp 135–142. doi:10.1145/1864708.1864736, http://doi.acm.org/10.1145/1864708.1864736

Jannach D, Zanker M, Felfernig A, Friedrich G (2010) Recommender systems: an introduction. Cambridge University Press. http://books.google.com/books?id=eygTJBd_U2cC

Kaplan AM, Haenlein M (2010) Users of the world, unite! the challenges and opportunities of social media. Bus Horiz 53(1):59–68. doi:10.1016/j.bushor.2009.09.003, http://www.sciencedirect.com/science/article/pii/S0007681309001232

Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. Computer 42(8):30–37

Ricci F, Rokach L, Shapira B, Kantor PB (eds) (2011) Recommender systems handbook. Springer, New York/London

Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. Commun ACM 18(11):613–620. doi:10.1145/361219.361220, http://doi.acm.org/10.1145/361219.361220

Xiong L, Chen X, Huang TK, Schneider JG, Carbonell JG (2010) Temporal collaborative filtering with Bayesian probabilistic tensor factorization. In: Proceedings of the SIAM international conference on data mining, SDM 2010, Columbus, pp 211–222, Apr 29–May 1 2010

Zheng VW, Cao B, Zheng Y, Xie X, Yang Q (2010) Collaborative filtering meets mobile recommendation: a user-centered approach. In: Proceedings of the 24th AAAI conference on artificial intelligence, Atlanta, pp 236–241

Spatiotemporal Proximity and Social Distance

Christoph Schlieder
Faculty for Information Systems and Applied Computer Sciences, Chair of Computing in the Cultural Sciences, University of Bamberg, Bamberg, Germany

Synonyms

Information filtering; Report confirmation

Glossary

LBSN Location-based social network
Heuristic Principle An experience-based, but fallible, problem-solving approach
Information Filtering An algorithm that aims at identifying relevant pieces of information
User-Generated Content Text, images, or other media published in a LBSN

Definition

Spatiotemporal proximity and social distance are two heuristic principles for filtering user-generated content produced by the members of a location-based social network (Schlieder and Yanenko 2010; Yap et al. 2012). Information filtering addresses the quality problem which arises when content is created by a large community of voluntary contributors as is the case in Web-based forms of participatory or citizen journalism. While the computational filtering approaches share some basic assumptions with the evaluation approach adopted in classical journalism, there are significant differences with respect to the scale of the problem and the methods for establishing confirmation relationships between reports.

User-Generated Content in Location-Based Social Networks

The idea of citizen reporters who complement the news coverage provided by professional journalists predates the Web by several decades (Deuze et al. 2007). With the diffusion of smart phones and the mobile access to the Web, however, it became much easier for eyewitnesses of events to report their observations to digital communities. Observation reports are published in location-based social networks (LBSN), that is, any type of Web-based social medium which provides geo-location metadata about its members and the user-generated content (Jensen et al. 2011). Such LBSN also makes available the temporal metadata used by conventional social media services. In other words, each observation reported in the LBSN comes with a time stamp and a place stamp. A prominent example of LBSN technology supporting citizen reporting is the Ushahidi platform which was originally created to collect and visualize reports about incidents of politically motivated violence (Okolloh 2009). Other scenarios include emergency response to natural disasters, the documentation of urban sprawl (Bishr and Mantelas 2008), and reports on wild animal sightings (Schlieder and Yanenko 2010).

Often it is useful to combine automatic filtering as a preprocessing step with manual post-processing by human experts who examine the remaining set of critical cases. Note, however, that in classical journalism, a small number of authors contribute articles each stating a large number of facts, whereas in citizen journalism, a large number of contributors publish reports that mostly state a single elementary fact such as a tweet about the sighting of a wood fire. Only the automatic approaches scale easily with the number of reports.

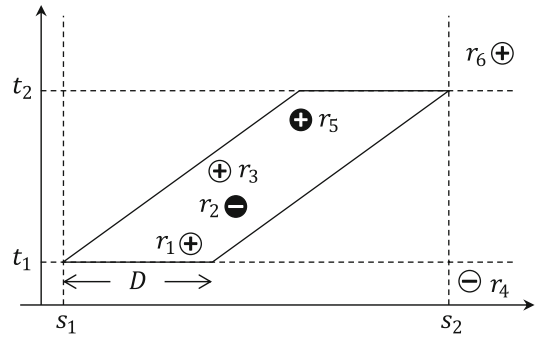
Constraint-Based Approaches for Report Confirmation

Different heuristic principles are used by the automatic filtering approaches. Bishr and Mantelas (2008) argue for using the spatial proximity of the observer to the object described in the report as a measure of the observer's reputation in

contributor status filtering. The rationale behind this heuristic *principle of spatial proximity* is that an eyewitness should have higher reputation than someone who reports from hearsay and that spatial proximity constitutes a necessary – though not sufficient – condition for observing the object. In scenarios such as reporting about natural disasters, however, eyewitnesses are often first-time or infrequent contributors which cannot be handled by reputation models. Report confirmation approaches have been proposed as an alternative by Schlieder and Yanenko (2010) and Yanenko and Schlieder (2012) to handle such scenarios.

Confirmation focuses on events instead of objects, that is, entities extended in time. As a consequence, the principle of spatial proximity needs to be complemented by a related *principle of temporal proximity*. A report of the sighting of a rare bird species, for instance, could be confirmed by a second report one hour later stating the sighting of the same species at a place nearby. Generally, a smaller distance corresponds to better confirmation. Both the spatial and the temporal proximity interact in confirmation and are therefore referred to as a single heuristic *principle of spatiotemporal proximity*.

In many application scenarios, observations are informed by the social role of the observer, by his or her affiliation to a subcommunity of the LBSN. An example is the competitive situation in a location-based game. The categorization of a game event as foul play is likely to be affected by which team the observer belongs to or supports (Yanenko and Schlieder 2012). In reporting about political events, such biases become even more important. The confirmation *principle of social distance* addresses such cases. It states that a report from an observer from a subcommunity of the LBSN which takes a different stance on the issue provides better confirmation than a report from an observer from the same subcommunity. According to this principle, a foul play reported by at least one member of both teams is considered having higher confirmation than a foul play reported only by members of the same team. This principle reflects the confirmation approach taken by classical journalism which requires at least two independent sources for each fact reported.

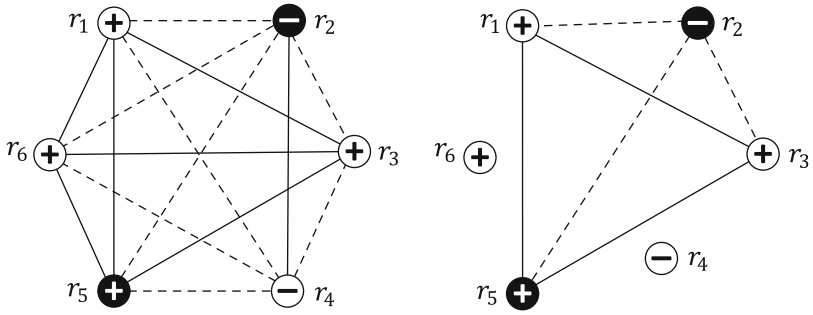


Spatiotemporal Proximity and Social Distance, Fig. 1
Space-time diagram

Space-Time Diagram

In terms of input and output, report confirmation approaches start from a collection of reports published in an LBSN where each report is represented by a tuple $r = (observer, time, place, observation)$. A confirmation value is computed for each ordered pair of reports as output. Consider the following example: six reports have been published by six different observers who are all positioned on the central avenue of the same city. The observations either affirm or deny that a protest march is moving along the avenue from point s_1 to point s_2 during the time period $[t_1, t_2]$. Since positions and movements refer to a linear geographic object, only a single spatial dimension needs to be represented in the example. In practice, however, confirmation filtering refers to two, sometimes to three, spatial dimensions (latitude, longitude, geoid height). Each of the reports $r_1, \ldots, r_6$ corresponds to a point in the space-time diagram (Fig. 1).

Affirmative observations are denoted by (+), negative ones which deny the event by (–). In the example, every observer either supports or opposes the issue of the protest march with social ties being established only within each of the two subcommunities. The space-time diagram distinguishes reports of supporters (white) from those of opposers (black). While reports appear as points in the diagram, events have a spatial and a temporal extension, that is, they cover regions. The spatiotemporal coverage of a protest march



Spatiotemporal Proximity and Social Distance, Fig. 2 Agreement graph and confirmation graph

of length D moving with constant speed is a parallelogram.

Without spatiotemporal and social context, it is only possible to determine agreement and disagreement between reports. An affirmative report r agrees with every report r_i affirming the same observation. This is expressed by the affirmation value ï£¡$a(r, r_i) = 1$. Similarly, the affirmative report r disagrees with any negative report r_j resulting in ï£¡$a(r, r_j) = -1$. The graph in Fig. 2 shows agreement by solid lines and disagreement by dashed lines. Agreement, however, is only a necessary, not a sufficient, condition for confirmation.

Confirmation Graph

Confirmation is modelled by a real-valued function which maps a pair of reports onto a confirmation value $c : (r_i, r_j) \to v \in [-1, 1]$ where $c(r_i, r_j) = 1$ corresponds to maximal confirmation, $c(r_i, r_j) = 0$ to unrelated reports, and $c(r_i, r_j) = -1$ to maximal conflict. The *spatial proximity* of two reports r_i and r_j is computed by considering the straight line distance or the street distance $d(r_i, r_j)$ between their place stamps and by taking into account additional spatial constraints. In the example, an additional constraint is the maximal length D, of a protest march in that city. If the place stamps are farther apart than D then it is rather unlikely that the two reports refer to the same event: $c_{\text{spa}}(r_i, r_j) = a(r_i, r_j)$ when $d(r_i, r_j) \leq D$, and $c_{\text{spa}}(r_i, r_j) = 0$ otherwise. Often, a logistic function is used to express the vagueness of the spatial threshold, $c(r_i, r_j) = a(r_i, r_j)/(1 + e^{d(r_i, r_j) - D})$.

In a comparable way, temporal constraints such as the maximal duration of an event are exploited to determine *temporal proximity*. Spatial and temporal constraints interact in the movement of the protest march. In such cases spatial distances and temporal distances cannot be considered independently. Measures of *spatiotemporal proximity* take account of the dependency of spatial and temporal constraints such as the assumption of uniform motion of the protest march in the example.

Figure 2 shows the agreement graph (left) and the result of confirmation filtering with the spatial confirmation function $c_{\text{spa}}(r_i, r_j)$ and an analogous temporal confirmation function. Generally, confirmation filtering removes edges. Report r_6, for instance, agrees with reports r_1, r_2, r_3, and r_5 but is not considered to confirm those reports because it is spatially farer than D from each of these reports. For the same reason, the negative report r_4 does not confirm the negative report r_2.

A further filtering step evaluates social distance and addresses the issue of (social) independence of sources. In the example, the authors of the reports r_1, r_3, r_4, and r_6 have established links in the social network forming a connected component isomorphic to K_4 while the authors of r_2 and r_5 form a second connected component isomorphic to K_2. There are only two social distances: between different nodes of the same component $d_{\text{soc}}(r_i, r_j) = 1$, otherwise $d_{\text{soc}}(r_i, r_j) = \infty$. The simplest filtering approach only identifies confirmation edges between socially distant nodes: $c_{\text{soc}}(r_i, r_j) = c(r_i, r_j)$ when

$c_{soc}(r_i, r_j) > 1$ and $c_{soc}(r_i, r_j) = 0$ otherwise. Applying this filter results in a further pruning of the confirmation graph. Only the edges between r_1 and r_5, r_1 and r_2, r_2 and r_3, as well as between r_3 and r_5 remain.

In application scenarios, the graphs are much larger and the distance measures reflect more complex modelling assumptions. Often it is possible to represent a further filtering step which exploits confirmation relations between more than just two reports as a finite domain constraint satisfaction problem which can be solved with algorithmic methods from qualitative spatiotemporal reasoning (Ligozat 2012; Yanenko and Schlieder 2012).

Cross-References

▸ Location-Based Social Networks

References

Bishr M, Mantelas L (2008) A trust and reputation model for filtering and classifying knowledge about urban growth. GeoJournal 72(3–4):229–237

Deuze M, Bruns A, Neuberger C (2007) Preparing for an age of participatory news. J Pract 1:322–338

Jensen C, Lee W-C, Zheng Y, Mokbel M (eds) (2011) Proceedings of the 3rd international workshop on location based social networks. ACM, New York

Ligozat G (2012) Qualitative spatial and temporal reasoning. Wiley, Hoboken

Okolloh O (2009) Ushahidi or 'testimony': Web 2.0 tools for crowdsourcing crisis information. In: Participatory learning and action, 59 (special issue on Web 2.0 for development). International Institute for Environment and Development, London, pp 65–70

Schlieder C, Yanenko O (2010) Spatio-temporal proximity and social distance: a confirmation framework for social reporting. In: Zhou X et al (eds) Proceedings of the 2nd international workshop on location based social networks, San Jose. ACM, New York, pp 60–67

Yanenko O, Schlieder C (2012) Enhancing the quality of volunteered geographic information: a constraint-based approach. In: Gensel J et al (eds) Bridging the geographic information sciences. Springer, Berlin, pp 429–446

Yap L, Bessho M, Koshizuka N, Sakamura K (2012) User-generated content for location-based services: a review. In: Lazakidou A (eds) Virtual communities, social networks and collaboration. Springer, Berlin, pp 163–179

Spatiotemporal Reasoning and Decision Support Tools

Chiara Renso[1] and Monica Wachowicz[2]
[1]KDD Lab, ISTI Institute of National Research Council, Pisa, Italy
[2]Geodesy and Geomatics Engineering, University of New Brunswick, Fredericton, Canada

Glossary

Computational Reasoning Is a process to solving problems, designing systems, and understanding human behavior that draws on the concepts fundamental to computer science

Deductive Reasoning Is a process based on a hierarchy of statements or truths in which it is thought that the observations provide a guarantee of the truth of the conclusion. Deductive reasoning arrives at a specific conclusion based on generalizations

Inductive Reasoning Is a process of creating probable true conclusions by starting from many specific observations. Inductive reasoning progresses from observations of individual cases to the development of a generality

Decision Support System According to Geoffrion's definition, a DSS has six characteristics (Geoffrion 1983): (1) is designed to solve ill- or semi-structured problems, i.e., where objectives cannot be fully or precisely defined; (2) has an interface that is both powerful and easy to use; (3) enables the user to combine models and data in a flexible manner; (4) helps the user explore the solution space (the options available to them) by using the models in the system to generate a series of feasible alternatives; (5) supports a variety of decision-making styles and easily adapted to provide new capabilities as the needs of the user evolves; and (6) allows an interactive and recursive process in which decision making proceeds by multiple passes, perhaps involving different routes, rather than a single linear path

Mobility Data Is any type of large volume datasets – structured and unstructured data – containing the information about the positions of a moving entity over time. It is usually represented as trajectories

Domain Knowledge Is the knowledge which is valid and directly used for a preselected domain of human or an autonomous computer activity. Different specialists and experts use and develop their own domain knowledge

Geographic Information System Integrates hardware, software, and data for capturing, managing, analyzing, and displaying all forms of geographically referenced information

Definition

One of the most universal theories about computational spatiotemporal reasoning is that the flow of inference is inherently unidirectional, moving from premises to be accepted as given to inferred conclusions. The direction of inference may vary depending on what is initially known, but it is generally assumed that in any reasoning task certain information constitutes the fixed premises (e.g., facts, observations, or events) from which certain other information can be derived as a tentative conclusion. In deductive reasoning, one moves from a general premise to a more specific conclusion. In contrast, inductive reasoning moves from specific premises to a general conclusion.

Premises based on experience or observations are best expressed inductively, while premises based on laws, rules, or other widely accepted principles are best expressed deductively. Spatial decision support systems (SDSS) have been largely developed based on the integration of deductive reasoning techniques and geographic information systems (GIS). The reasoning tasks of SDSS have been typically formulated as a set of rules, constraints, and multi-criteria matrices and/or functions representing possibilities (e.g., possible beliefs or actions over space and time), interconnected by links representing positive and negative support relations between pairs of these possibilities.

In contrast, social networks are transforming SDSS because their feeds (e.g., twits, microblogs, mobility traces) exist in data streams that are processed on the fly, producing a continuous flow of information for automated (near) real-time decision making. Indeed new models of scientific discovery are emerging from developments in rather focussed crowdsourcing, and these are applicable to how we might figure out good designs for the future generation of SDSS which must deliver real-time information and changing knowledge to decision makers.

Introduction

Currently, each individual of a social network is generating automatically sensed mobility data that is revolutionizing the traditional fields of spatiotemporal reasoning and decision-making analysis, not only to scale up to the large and near real-time data volumes but also to address complex questions related to change, trends, duration, and evolution. The mobility datasets are usually ground truthed: real trajectories are directly and continuously sampled as they occur in real time, but clearly they do not have any semantic annotation or context. Therefore, their interpretation requires rich domain background knowledge to fulfill meaningful reasoning tasks.

For example, behavior recognition in smart homes often employs graphical models like hidden Markov chains. By combining them with contextual information about space and time, the performance of these models can be boosted (e.g., see Chua et al. 2009). Such cross-fertilizations are clearly identifiable in the recent work in cognitive vision (Dubba et al. 2010), where the demonstrated interactions and integrations of techniques from machine learning, inductive logic programming, and spatiotemporal modelling may serve as a blueprint for the construction of hybrid intelligent systems dealing with real-time spatial information.

The combination of deductive and inductive reasoning tasks is also needed to extend the notion of SDSS, which will involve reasoning in real time on huge and possibly noise mobility

data generated by social networks. In perspective, by coupling the social and mobility networks with further context information, it will be possible to explore the evolution dynamics of the urban social sphere; to predict the spreading of sentiments, opinions, and diseases; and thus to understand in real time the evolving borders of the community structure of a city.

Key Points

The analysis, reasoning, and, consequently, taking decisions based on the mobility traces of any social network bring several challenges. First of all, there is a need to properly represent the multidimensional aspects that mobility data brings not only because space and time are inextricably linked but also for its inherent contextual dimension that comes from the semantics of a domain knowledge. We argue that an automated real-time decision-making process cannot be achieved based on the pattern discovery analysis of the *raw mobility data alone* but requires a platform for reasoning on massive heterogeneous information such as social media data. The platform might have a cloud architecture to exploit techniques and heuristics from diverse areas such as databases, machine learning, and the Semantic Web. Extending reasoning approaches to support such a platform is a known challenge for the reasoning community. In deductive reasoning, different approaches have been developed to revise beliefs based on recent information. In inductive reasoning, a body of research in data mining and machine learning already supports online data analysis (Giannotti et al. 2011). However, more research is needed for dealing with rich and unstructured data streams. Second, when dealing with mobility, domain knowledge, and social interactions, all the standard methods developed for computational reasoning are no longer sufficient to clearly separate the static and dynamic parts of a domain knowledge. Some examples include the development of new areas such as the emergence of integrated spatiotemporal calculi, spatiotemporal dynamics, commonsense reasoning about space, and the use of non-monotonic reasoning techniques for reasoning about spatial change (Galton 2000).

The set of spatial relations used in the 9-Intersection Model (Egenhofer and Franzosa 1991) has become part of the OpenGIS Implementation Specifications (ISO 19107) and are currently also supported by some commercial GIS products. In the domain of spatial computing for design (Freksa 1991), e.g., for architecture design assistance, the integration of spatial reasoning with other forms of reasoning such as conceptual/ontological and (spatio-) terminological inference and constraint logic programming (CLP) (Jaffar and Maher 1994) has led to encouraging results, interesting fundamental questions, and possibilities for the application of QSTR in an area (i.e., CAAD) with a potential industrial impact.

This new paradigm is providing new research directions for the development of new spatiotemporal reasoning and decision-making tools for a better understanding of social process over space and time, such as influence, trust, and information spreading.

Finally, the data mining community has seen a growing interest in providing several algorithms and techniques tailored on trajectory data. However, decision making needs to take into consideration some sort of interactions between moving entities, thereby extracting valid, novel, and useful patterns in networks ranging from transportation networks to World Wide Web and to social networks (Memom et al. 2010). This area is still in its infancy but rapidly evolving to provide examples of new techniques and applications, leading to future research directions.

An important research issue that may arise from this data combination is the impact of social networks on human mobility (and vice versa). Some example of research questions are the following: How do the social communities of a person "geographically move" across the time? Is there any relationship between the social contacts of a person and their location/co-location over time? Is there any relation between location visited by a person and the locations visited by his/her

friend? These are open questions that could give rise to a new research area at the cross line of spatiotemporal reasoning, mobility, and social networks.

Proposed Solution and Methodology

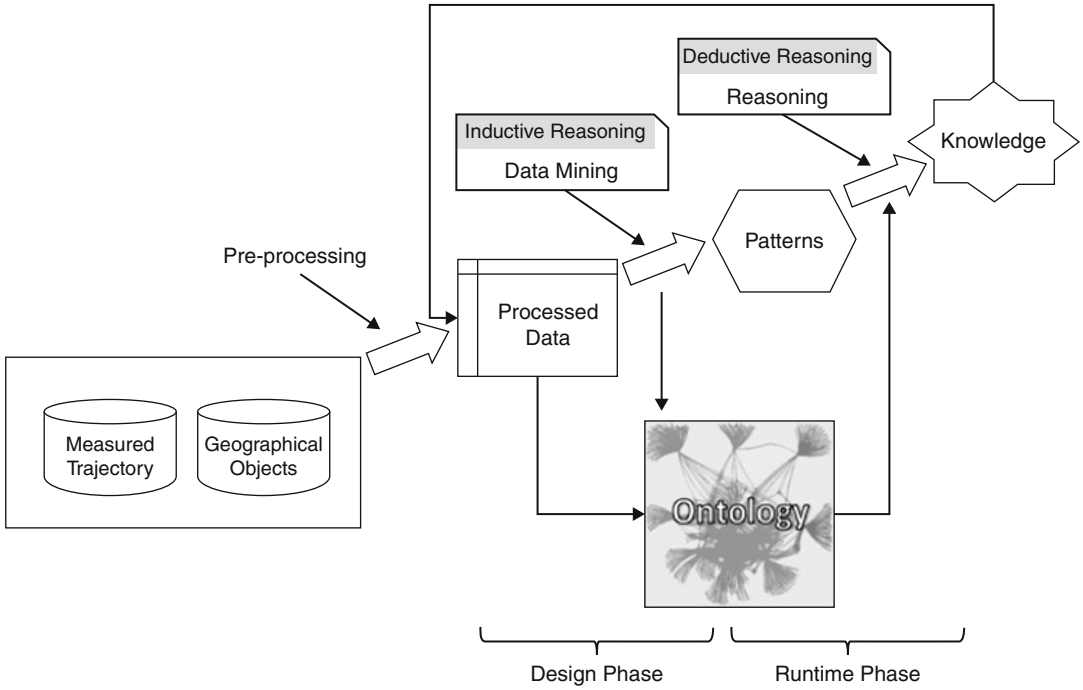
Analyzing mobility data can be rephrased in *how, when, where*, and *why* entities move. This knowledge gives to a user – who could be a traffic manager or an urban planner – a better understanding on how to increase the efficiency of energy systems and the delivery of services ranging from utilities to retailing in cities and to improve communications and transportation. However, the semantic lift from finding *how* movement happened to understanding *why* entities are moving in that way needs a new computational reasoning approach tailored on mobility and network characteristics and possibly enriched with contextual semantic information from a domain knowledge.

The classical knowledge discovery (KDD) process is mainly characterized by inductive reasoning, which begins with gathering data (i.e., facts) that are specific and limited in scope. Then, it proceeds to a generalized conclusion with a certain degree of uncertainty, depending on the accumulated evidences. By gathering data, seeking patterns, and building hypothesis, this process allows us to explain what has been observed, having the ultimate goal of enhancing a domain knowledge. However, the generalized conclusions are not absolutely certain, even after taking into account any premises on human behavior.

We propose a *semantic-enriched knowledge discovery process* that makes use of domain knowledge (i.e., the users' a priori knowledge on human behavior) by developing a mobility behavior ontology where trajectory data and movement patterns are to be classified. This process is based on the integration of inductive reasoning (pattern discovery) and deductive reasoning (behavior inference) that allows discovered mobility patterns to be understood in terms of human behavior. Essentially, this new process allows the integration of querying and mining tasks with reasoning tasks, and it is illustrated in Fig. 1 below. The trajectories measured from location devices along with contextual data such as geographical objects are first preprocessed to be adapted for the mining, or inductive, step. This reasoning step returns patterns representing models of the mobility, further post-processed in a deductive step to get meaningful knowledge. An ontology supports the steps of the semantic-enriched KDD process in defining the mobility behavior interesting for the mobility decisions and performing the final deductive reasoning step. The KDD process is said to be interactive and iterative. Interactive since the user is expected to interact with the process choosing the most appropriate algorithm and setting the best parameters. Iterative since the process has typically to be repeated in a progressive manner to refine the results. This semantic-enriched mobility knowledge discovery process was implemented as a tool called M-Atlas which provides the basic components for supporting a decision support system (Giannotti et al. 2011; Renso et al. 2013).

M-Atlas

M-Atlas handles all the steps of the mobility knowledge discovery process providing a SQL-based trajectory data mining query language. Besides the mechanisms for storing and querying trajectory data, M-Atlas has mechanisms for mining trajectory patterns and models that, in turn, can be stored and queried. The basic design choice is compositionality, i.e., querying and mining of trajectory data, patterns, and models may be freely combined, in order to provide the expressive power needed to master the complexity of the mobility knowledge discovery process. The conceptual model behind M-Atlas provides the interaction between two conceptual worlds: the *data world* and the *model world*. The former is a set of entities to be mined, trajectories in our case; the latter is a set of models and patterns inferred from the data, representing the result of mining tasks. Mining operators map data into models, or patterns, while entailment operators map models, patterns, and data into



Spatiotemporal Reasoning and Decision Support Tools, Fig. 1 The semantic enriched knowledge discovery process

the data that satisfy the property expressed in the given model or pattern. This view supports compositionality, since data can be mapped onto models and vice versa, coherently with inductive databases vision introduced in Imielinski and Mannila (1996).

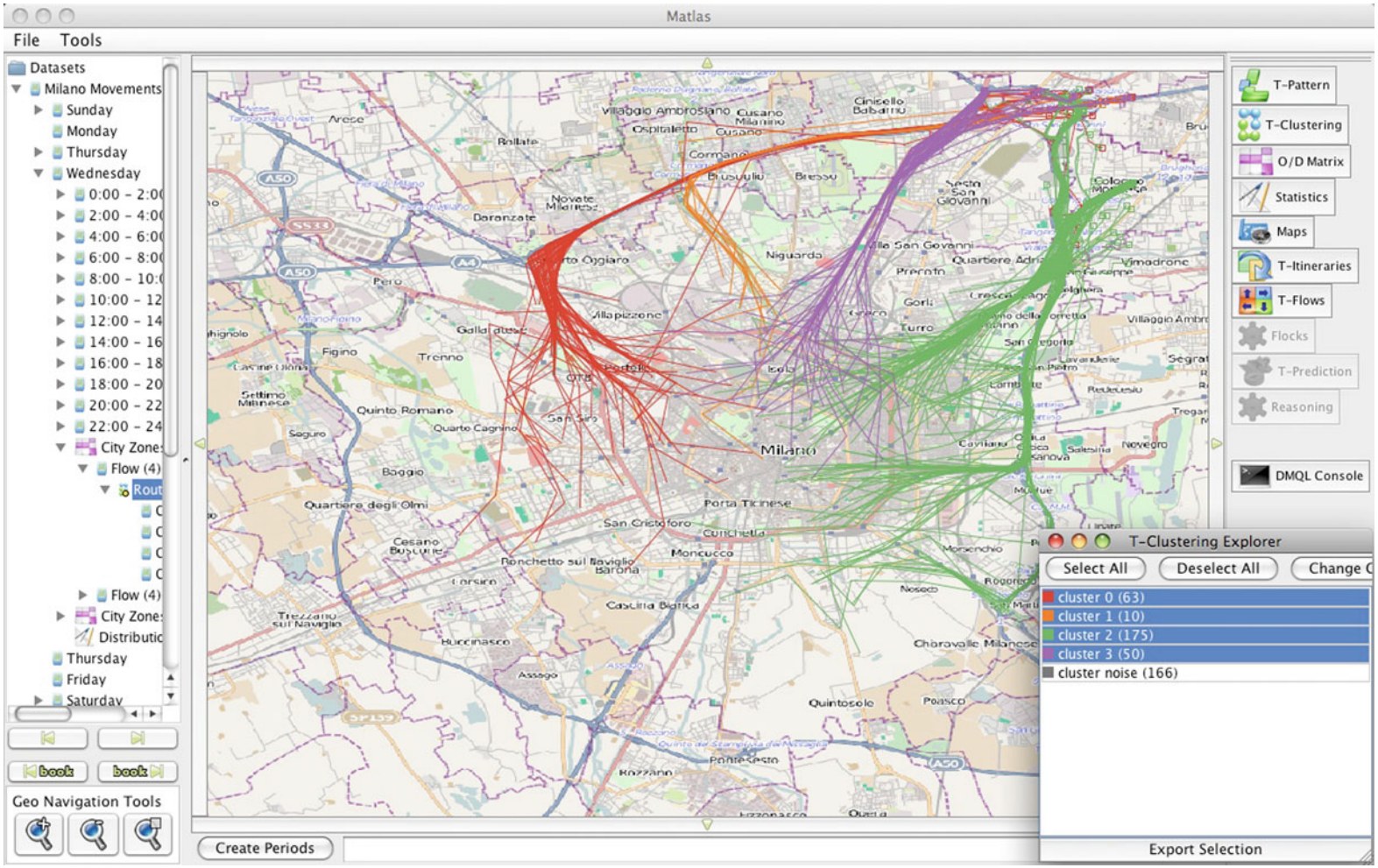
The M-Atlas system is equipped with a graphical user interface and a set of interactive tools allowing the user to navigate the data and model easily. Each interaction of the analyst with the interface is compiled into a sequence of M-Atlas queries which can be retrieved at any moment to describe or review the entire process. A screenshot of the M-Atlas interface is shown in Fig. 2.

However, getting useful information from a knowledge discovery process is not a straightforward process; in fact usually it is difficult even for an expert analyst user to choose the correct algorithm and set up all the parameters in order to extract meaningful patterns. This is especially true in decision support systems due to the complexity of the mobility data under analysis as well as the application requirements. For this reason, the inferred behavior can be only meaningful if a proper semantic-enriched mobility knowledge discovery process has been set up. This is detailed below following the steps of the KDD process: preprocessing, mining, post-processing.

Preprocessing

Data Reconstruction The mobility data mining algorithms apply to the concept of trajectory: but which is the definition of a trajectory? Is it simply the ordered sequence of observations of the user's history or a subsequence representing the movements between the stationary points? And how to define and compute a stop? Answering to these questions is crucial and affects deeply the reasoning tasks of a decision support system. There are several ways of reconstructing trajectories considering different constraints and thresholds thus leading to different sets of trajectories.

Data Manipulation Before the execution of a data mining algorithm, the analyst can manipulate (e.g., selecting the data in a particular area or period) or transform the data (e.g., anonymization).



Spatiotemporal Reasoning and Decision Support Tools, Fig. 2 The M-Atlas interface

For this reason, M-Atlas integrates a rich set of spatiotemporal primitives and transformations. For example, it is possible to compute the temporal distribution of the movements, and this is useful in the inductive reasoning task when the data to be mined have to be selected based on space and/or time as in the example above. An example of a transformation operation is the anonymization of trajectories, where the initial dataset is transformed in order to guarantee the anonymity of the users.

Data Mining

The induction step is the core of the process and consists in the proper execution the mining algorithms on trajectory data. M-Atlas realizes this step using a *mining* statement which creates a new model as the result of a mining task, specifying the inductive algorithm to execute on a selection of trajectories with the set of the parameters.

Mining a Data Sample Applying a data mining algorithm on a large trajectory dataset may be extremely time and memory consuming, making the direct application of the algorithm to the entire dataset not possible due the time or memory limitation. This problem can be solved using the data mining algorithms in combination with data sampling techniques. In general, data sampling is a technique to reduce the size of the data without altering the statistical properties.

The data can be sampled using semantic criteria such as dividing the data using the spatial or temporal characteristics of the trajectories. Whatever sampling technique is chosen by the analyst, the important issue is to maintain the consistency of the data or, at least, understand exactly the bias introduced since this may strongly affect the extracted patterns.

Model Manipulation Similarly to the trajectory data, the models resulting from the mining step can be stored and manipulated to produce a useful and meaningful representation of the trajectory behaviors. For this reason, the M-Atlas statements can be used also on models. In particular, M-Atlas provides a relation which is a bridge between data and models, called *entails*, which identifies the data which support a model. This relation is crucial in a decision support system since it allows the interaction between data and models where models are progressively mined and combined with data.

Progressive Reasoning As described above, a decision support system does not entail a straightforward sequence where a single run of data mining algorithm can perform the whole reasoning task. The iterative and interactive aspects are crucial to get a real understanding of the data and extracted patterns. The progressive inductive reasoning technique is the concatenation of a series of mining algorithms which restrict at each step their constraints removing the not-interesting data or noise. At each step the models are extracted and the data supporting them are reused in order to apply a more rigorous version of the mining algorithm.

Post-processing

Post-processing refers to techniques that can be applied once the inductive task have been performed and refers to the evaluation or interestingness of the extracted patterns. The result of this step may trigger a new iteration of the knowledge discovery process. The validation of the patterns aims at measuring how much the extracted patterns are valid and not just random results. The pattern reasoning task, instead, is more semantic in the sense that it aims at interpreting the patterns in the light of a domain knowledge.

Patterns Interpretation The intrinsic difficulty of pattern interpretation lies in the need of integrating into the discovery process the contextual dimension. We define *contextual dimension as any kind of information that is not only geometric and that has some relation with data domain knowledge*. Examples of contexts are the geographical environment where the entities move (e.g., hotels, roads, parks), any non-geometric moving entity feature (e.g., the age of the tracked person), or the application-specific concepts and behavior (e.g., purpose of the movement or predefined behavior like commuting, shopping, or leisure travelling).

The domain knowledge may be globally represented by formally encoding it into a

knowledge representation structure such as an ontology which can be used to represent the main concepts of the application.

An interesting feature of combining data mining with ontologies in a decision support system is the possibility of integrating deduction and induction aspects (Renso et al. 2013). The inductive power of the data mining, extracting patterns from data (bottom–up), is enriched with the possibility to deductively infer additional information based on some application domain knowledge (top–down). This combination allows us to classify the mobility patterns, as extracted from the mining step, into the domain knowledge concepts encoded in the ontology. An example of this induction-deduction combination is the framework Athena, an extension of M-Atlas that is an attempt to exploit ontologies in a knowledge discovery process. Athena represents domain knowledge into an ontology where axioms define the behavior we want to find in the mobility data. Therefore a classification of the extracted patterns into predefined behavior is performed directly by the ontology reasoning engine.

Illustrative Example

We present an example where park managers are interested in decision making to create and deliver recreation and fitness programs in a variety of settings, based on the understanding of the visitor behavior in the Dwingelderveld National Park, in the Netherlands. They are particularly interested in exploring and understanding the disturbing types of behavior of different types of visitors in this park (van Marwijk and Pitt 2008).

The data to be analyzed was obtained from three different information sources. The first source was a questionnaire containing records about visitor characteristics, preferences, and motivations for visiting the national park. The questionnaire was manually filled in by all visitors in the experiment. The raw trajectory data has been collected by the visitors carrying a GPS receiver during their visit in the park. This experiment was carried out during 7 days (weekend and weekdays) in spring and summer of 2006 for a total of 461 visitors.

Data Preprocessing During the data preprocessing step, the stop and move segments of the trajectories were computed. The spatial and temporal thresholds used to identify the stops were 10 min and 20 m, respectively. We extracted the interesting places (e.g., radio telescope, café) from the questionnaire filled in by the visitors.

Ontology The ontology, illustrated in Fig. 3, consists of nine concepts:

Interesting places: The places in the park where a visitor usually stops for recreational activity – such as eating and bird watching. Some examples include the café or the radio telescope located in the Dwingelderveld National Park.

Forbidden areas: The areas where a visitor should not stop at any time during his/her visit to the park to avoid disturbing the animals.

Intersection path: The path intersections where a visitor stops for orientation purposes.

Long: The period of time which identifies a stop with a long period of stay.

Flock pattern: Consists of a type of mobility mining pattern representing groups of entities moving closely at the same time.

Visitor behavior: Represents a trajectory where the stops occur at the predefined places such as interesting places, path intersections, and forbidden areas.

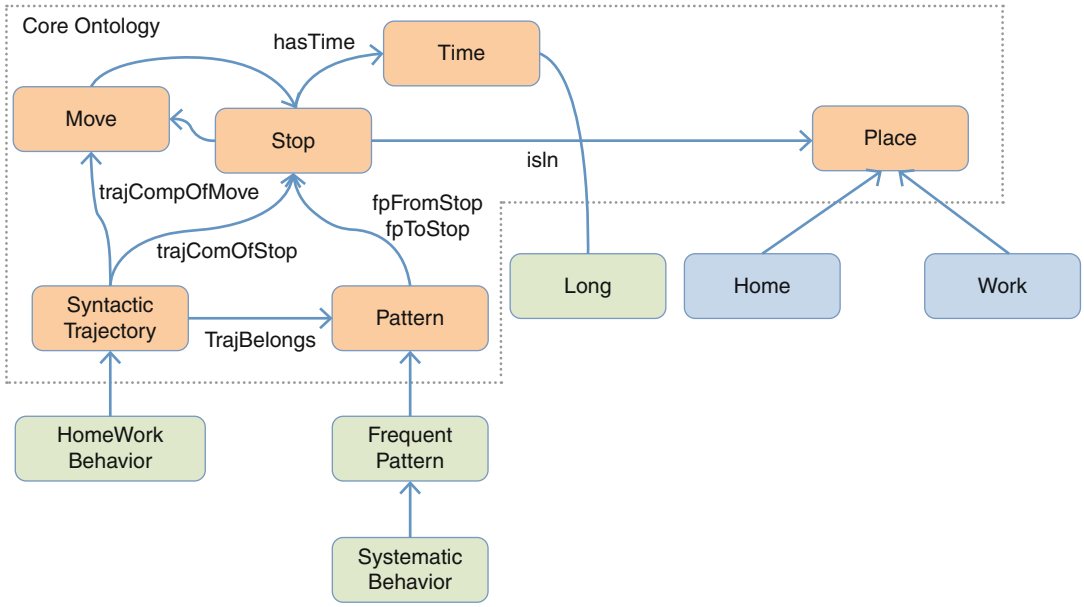
Exploring behavior: Is the movement of visitors in the park who are interested in exploring the features of the park.

Socializing behavior: Is defined as visitors encountering and staying together for a given period of time.

Disturbing behavior: A group of visitors who have stopped in a forbidden zone of the park for a given period of time.

Suspicious behavior: Any individual belonging to a disturbing group for a long period of time.

Inductive Reasoning During the mining step, we run the flock algorithm (more details on the algorithm is found in Wachowicz et al. (2011)) on



Spatiotemporal Reasoning and Decision Support Tools, Fig. 3 Behavior ontology

the trajectories to find clusters of people moving closely at the same time.

Figure 4 shows the visitors' trajectories (dark grey) belonging to the discovered flock patterns where the stops are depicted by one of the following:

- Interesting places: such as camp (green), radio telescope (purple), or cafe (yellow)
- Path intersections (blue)
- Forbidden areas (red)

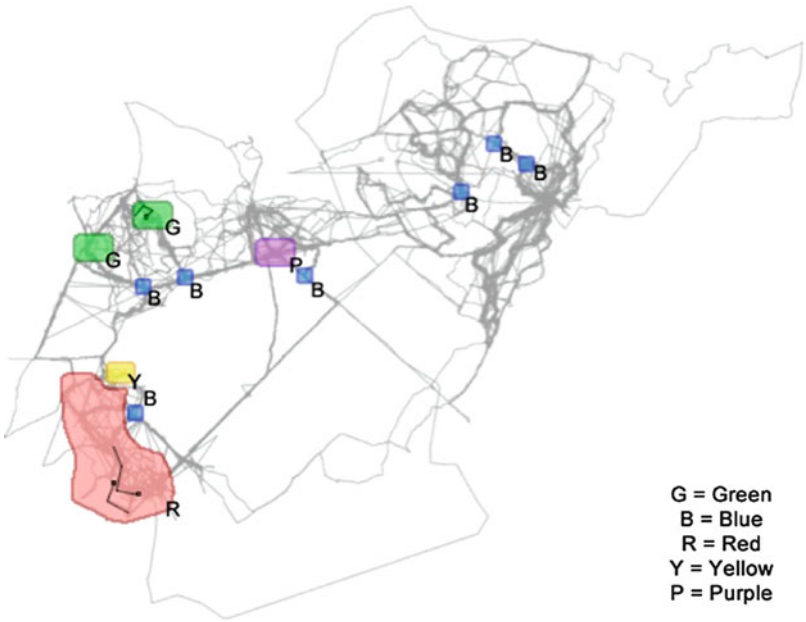
Deductive Reasoning All the patterns are imported into the mobility behavior ontology and then the reasoning engine is run to obtain the classification of visitors' behavior based on the ontology axioms. Once the reasoning step is completed, the graphical interface visualizes the different types of trajectories. For example, the user can visualize the visitors' trajectories belonging to disturbing and exploring behaviors as shown Fig. 5. The left part depicts the disturbing behavior whereas the right part shows the exploring behavior.

We can notice how the trajectories expressing the exploring behavior tend to be sparser and reach further areas of the park. To the contrary, the disturbing trajectories tend to stay around the forbidden area.

Key Applications

A variety of sources of information captured in real time by the pervasive digital media technology connecting us through smart indoor/outdoor sensors, remote sensing, mobile devices, and the "Internet of Things" are fuelling exponential growth in voluminous amount of mobility data. This exponential growing of mobility data is comparable to a similar growing rate of unstructured data generated by social networks such as Facebook, Google+, Twitter, and Foursquare. This big data revolution will continue over the next decade and beyond with opportunities that will include:

– *Smart governance*: enhancement of citizen participation in the decision-making process, creation of new public and social services, support for transparent governance, and advance political strategies and perspectives.
– *Transport and ICT*: improve local accessibility and develop sustainable, innovative, and safe transport systems.
– *Natural resources*: reduce pollution, support environmental protection, and achieve a sustainable resource management.



Spatiotemporal Reasoning and Decision Support Tools, Fig. 4 The discovered flock patterns

Spatiotemporal Reasoning and Decision Support Tools, Fig. 5 Disturbing behavior and exploring behavior

– *Quality of life*: creates new cultural, housing, and education facilities and improves health conditions and individual safety.

Future Directions

The integration of mobility mining with social mining is naturally a growing and promising future direction in this field. The challenge is to reach a deeper understanding of patterns regulating how people move and communicate. How does the social dimension affect the movements of an individual? What is the role of mobility in creation of new social ties? How do social communities change and "move" in time?

Although extremely challenging, it is obvious that this topic brings many issues. First of all the privacy of the individuals, already sensitive in the mobility case, becomes even more sensitive when, for example, combining the users' locations with their social contacts. New spatiotemporal reasoning techniques are needed for supporting privacy-preserving reasoning in which decision makers can securely pose queries against a decision support system using inferences drawn based on both hidden and visible part of a knowledge base, without revealing

the hidden knowledge. Second, collecting proper datasets for the analysis becomes much more cumbersome, and we can expect to rely more on location-enabled social networks like Flickr and Twitter. In these cases, new spatiotemporal reasoning techniques are required to infer from data having different spatiotemporal scales (e.g., from centimeters to kilometers; from seconds to years). Finally, it is clear that new ad hoc techniques have to be developed to take into account the combination of these two aspects (i.e., privacy and scale) in a proper way.

Acknowledgments

This work has been partially supported by EU projects MODAP GA N. 245410, DATASIM GA N. 270833, and SEEK GA N. 295179.

Cross-References

- ▸ Clustering Algorithms
- ▸ Data Mining
- ▸ Description Logics
- ▸ Reasoning

References

Chua S-L, Marsland S, Guesgen HW (2009) Behaviour recognition from sensory streams in smart environments. In: Nicholson AE, Li X (eds) Australasian conference on artificial intelligence, Melbourne. Volume 5866 of Lecture notes in computer science. Springer, pp 666–675

Dubba KSR, Cohn AG, Hogg DC (2010) Event model learning from complex videos using ILP. In: Proceedings of the ECAI, Lisbon. Volume 215 of Frontiers in artificial intelligence and applications. IOS Press, pp 93–98

Egenhofer MJ, Franzosa RD (1991) Point set topological relations. Int J Geogr Inf Syst 5(2):161–174

Freksa C (1991) Qualitative spatial reasoning. In: Mark D, Frank A (eds) Cognitive and linguistic aspects of geographic space. Kluwer, Dordrecht, pp 361–372

Galton A (2000) Qualitative spatial change. Oxford University Press, Oxford/New York. ISBN:0198233973

Geoffrion AM (1983) Can OR/MS evolve fast enough? Source for six essential characteristics of DSS. Interfaces 13:10

Giannotti F, Nanni M, Pedreschi D, Pinelli F, Renso C, Rinzivillo S, Trasarti R (2011) Unveiling the complexity of human mobility by querying and mining massive trajectory data. VLDB J 20(5):695–719

Imielinski T, Mannila H (1996) A data base perspective on knowledge discovery. Commun ACM 39:58–64. doi:10.1145/240455.240472

Jaffar J, Maher MJ (1994) Constraint logic programming: a survey. J Log Program 19/20:503–581

Memom N, Xu JJ, Hicks DL, Chen H (2010) Data mining for social network data. Annals of information systems, vol 12. Springer, New York/London

Renso C, Baglioni M, Macedo JA, Trasarti R, Wachowicz M (2013) How you move reveals who you are: understanding people's behavior by analyzing trajectory data. Knowl Inf Syst 37(2):331–362

van Marwijk R, Pitt DG (2008) Where Dutch recreationists walk: path design, physical features and walker usage. In: Raschi A, Tamperi S (eds) Proceedings 4th international conference on monitoring and management of visitor flows in recreational and protected areas, management for protection and sustainable development, Montecatini Terme, pp 428–432

Wachowicz M, Ong R, Renso C, Nanni M (2011) Discovering moving flock patterns among pedestrians through spatio-temporal coherence. Int J Geogr Inf Sci 25(11):1849–1864

Recommended Reading

Bhatt M, Guesgen H, Woelfl S, Hazarika S (2011) Qualitative spatial and temporal reasoning: emerging applications, trends, and directions. Spat Cogn Comput 11(1):1–14

Van Orshoven J, Kint V, Wijffels A, Estrella R, Bencsik G, Vanegas P, Muys B, Cattrysse D, Dondeyne S (2011) Upgrading geographic information systems to spatial decision support systems. J Math Comput For Nat Resour Sci 3(1):36–42

Spatio-Temporal-Thematic Analysis

▸ Twitris: A System for Collective Social Intelligence

Spatiotemporal Web

▸ Spatiotemporal Information for the Web

Spectral Analysis

Xiao-Dong Zhang
Department of Mathematics, MOE-LSC, Shanghai Jiao Tong University, Shanghai, P.R. China

Synonyms

Spectral graph analysis; Spectral network analysis; Spectral technique

Glossary

Network (graph) A network G is a triple consisting of a node set $V(G)$, a link set $E(G)$, and a relation that associates each link with two nodes

Adjacency Matrix Let $G = (V(G), E(G))$ be a network with $V(G) = \{v_1, \cdots, v_n\}$. The adjacency matrix $A(G) = (a_{ij})$ of G is $n \times n$ matrix with $a_{ij} = 1$ if v_i is adjacent to v_j, and 0 otherwise

Eigenvalues of a Graph All eigenvalues of the adjacency matrix $A(G)$ of a graph G are called eigenvalues of G and denoted by

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$

Degree Diagonal Matrix The degree diagonal matrix $D(G)$ of a network G is the diagonal matrix whose diagonal entries are degrees of the corresponding nodes

Laplacian Matrix The Laplacian matrix $L(G)$ is defined to be $L(G) = D(G) - A(G)$, where $D(G)$ is the degree diagonal matrix and $A(G)$ is the adjacency matrix

Laplacian Eigenvalues of a Graph All eigenvalues of the Laplacian matrix $L(G)$ of a graph G are called the Laplacian eigenvalues of G and denoted by

$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_n = 0$$

Normal Matrix The normal matrix $N(G)$ is defined to as the product of the inverse of degree diagonal matrix and the adjacency matrix

Normal Eigenvalues of a Graph All eigenvalues of the normal matrix $N(G)$ of a graph are called the normal eigenvalues of G and denoted by

$$1 = \nu_1 \geq \nu_2 \geq \cdots \geq \nu_n$$

Adjacency (Laplacian, Normal) Spectrum The set of all eigenvalues of the adjacency (Laplacian, normal) matrix

Walk A walk is a list $v_0, e_1, \ldots, e_k, v_k$ of nodes and links such that, for $1 \leq i \leq k$, the link e_i has endpoints v_{i-1} and v_i. The length of a walk is its number of links

Bipartite Network A network $G = (V, E)$ is called bipartite if V is decomposed into two disjoint sets such that each link has its ends in different sets

Definition

In this paper, we describe spectral analysis of networks. Generally speaking, the eigenvalues and eigenvectors of the different matrices associated with networks are intimately connected to important topological features, such as diameter, community structure, node centrality, etc.

Introduction

In the past decade, networks have attracted considerable attention in many disciplines such as statistical physics, social science, and applied mathematics. Generally speaking, the study of network representations of physical, biological, and social phenomena leading to predictive models of these phenomena is now called network science. Newman (2003) reviews some features of real-world networks and properties of several network models, including random graphs, the small-world model, models of network growth, and epidemiological processes. Boccaletti et al. (2006) survey the important concepts and results

in the network science, in particular the topological structure and synchronization and collective dynamics of complex networks. One important class of complex networks is the class of social networks, which are social structures consisting of individuals (or groups) called nodes and their relationships, such as friendship, common interest, and financial exchange, called links. The analysis of social networks is used in epidemiology, mass surveillance, diffusion of innovations, etc. There are many measures (metrics) in social network analysis, such as betweenness, centrality, and clustering coefficient. The community structure, or clusters, is one of the most important features in sociology. Recently, Fortunato (2010) gave a thorough exposition of community deletion, from the several main definitions of the community problem to the presentation of most methods developed.

In mathematics, spectral graph theory (or analysis) is the study of properties of a graph (network) in relationship to the characteristic polynomial, eigenvalues, and eigenvectors of matrices associated to the graph, such as its adjacency matrix or Laplacian matrix. There are two excellent books, i.e., "Spectral Graph Theory (Chung 1997)" and "Spectra of Graphs-Theory and Applications (Cvetkovi et al. 1995)," which focus on deducing the properties and structure of a graph from its graph spectrum and reveal increasingly rich connections with many areas of mathematics and other disciplines, such as quantum chemistry, statistical physics, and computer science. The canonical example is the use of eigenvalue techniques to prove that certain extremal graphs cannot exist. The eigenvalues of a network are intimately connected to important topological features such as diameter (average distance), clustering coefficient, connectivity, and how random the network is. The associated eigenvectors can be used to detect community structure or clustering. What is more, some important results purely on networks cannot be proved without resorting to algebraic methods, involving a consideration of eigenvalues of adjacency (Laplacian) matrices of graphs. For example, Gkantsidis et al. (2003) use the weights of the eigenvector corresponding to the largest eigenvalue of the adjacency matrix to obtain an alternative hierarchical ranking of the Autonomous System. Spectral analysis has been successfully applied to the detection of community structure of networks, being based on the adjacency matrix, the Laplacian matrix, the normalized Laplacian matrix, etc. Moreover, many real networks may be visualized by spectral methods (see Seary and Richards 2005). Social network analysis can be dated back in the early 1920s and has now become one of the most important methods in investigating the features and structures of social systems (see Scott 2000; Wasserman and Faust 1994).

In this entry we introduce results about spectral analysis of social networks and explain how to find the community structure and centrality of social networks by the means of spectral network theory.

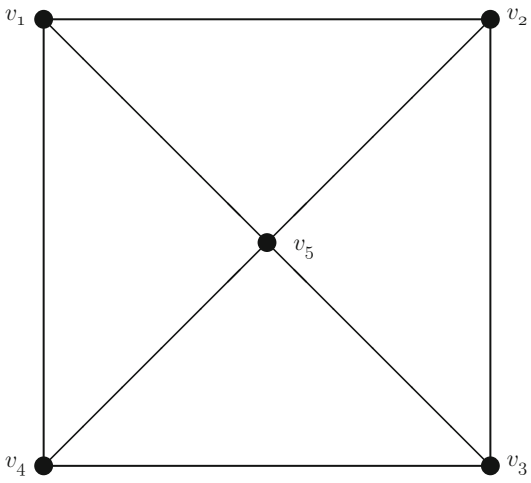
Adjacency Spectrum

There are several matrices associated with a network. For a network $G = (V, E)$ with $V = \{v_1, \cdots, v_n\}$, the most commonly used matrix may be the adjacency matrix $A(G) = (a_{ij})$ of order n. Clearly the adjacency matrix of a network is symmetric and the entries of the main diagonal are zeros. In this way, there is one to one correspondence between networks and $(0, 1)$-symmetric matrices with zeros on the main diagonal. So all information of networks can be presented and obtained by the properties of matrices. By the way, the adjacency matrix can also be generalized to represent weighted networks.

Example 1 For the Graph G in Fig. 1: G, five vertices and eight edges.

Then the adjacency matrix is

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix}.$$



Spectral Analysis, Fig. 1 Graph G

One of the important sets associated with a network is the set of all walks between any pair of nodes v_i and v_j.

Proposition 1 *Let $G = (V, E)$ be a network associated with the adjacency matrix A. Then the number of walks of length k starting at node v_i and ending at node v_j is the (i, j) entry of A^k.*

From this proposition, if the entry (i, j) of A^k is positive, then there exists at least one path from node v_i to node v_j. Hence, we can conclude that the diameter of a network is at most d if there exists an integer d such that all entries of A^d are positive.

Another property of networks is revealed by the entries of A.

Proposition 2 *Let $G = (V, E)$ be a network associated with the $n \times n$ adjacency matrix A. Then G is bipartite if and only if for some odd integer r, the diagonal entries of A^r are all zero.*

Since the adjacency matrix of G is symmetric, all eigenvalues of A are real and can be denoted by

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n.$$

Moreover, A is nonnegative matrix. By the Perron-Frobenius theorem, λ_1 must be positive and be greater than or equal to the absolute values of the other eigenvalue, i.e., $\lambda_1 \geq |\lambda_i|$ for $i = 2, \ldots, n$. Further, there exists, up to constant multiplication, only one eigenvector with all nonnegative entries, corresponding to the eigenvalue λ_1. For example, the eigenvalues of the star network of order n are

$$\sqrt{n-1},\ -\sqrt{n-1},\ 0,\ \ldots,\ 0.$$

The eigenvalues of the complete network of order n are

$$n-1,\ -1,\ \ldots,\ -1.$$

The eigenvalues of a path of order n are

$$2\cos\frac{\pi}{2(n+1)},\ 2\cos\frac{2\pi}{2(n+1)},\ \ldots,$$
$$2\cos\frac{k\pi}{2(n+1)},\ \ldots,\ 2\cos\frac{n\pi}{2(n+1)}. \quad (1)$$

The eigenvalues of cycle are

$$2,\ 2\cos\frac{2\pi}{n},\ 2\cos\frac{4\pi}{n},\ 2\cos\frac{2k\pi}{n},\ \ldots,$$
$$2\cos\frac{2(n-1)\pi}{n}. \quad (2)$$

The eigenvalues of G in Fig. 1 are

$$3.2361, 0,\ 0, -1.2361, -2,$$

and corresponding to eigenvectors,

$$(0.4253, 0.4253, 0.4253, 0.4253, 0.5257)^T,$$
$$(-0.6932, -0.1398, 0.6932, 0.1398, 0.0000)^T,$$
$$(-0.1398, 0.6932, 0.1398, -0.6932, -0.0000)^T,$$
$$(-0.2629, -0.2629, -0.2629, -0.2629, 0.8507)^T,$$
$$(0.5000, -0.5000, 0.5000, -0.5000, 0.0000)^T.$$

The following proposition reveals the relationship between the structure of the network and its spectrum.

Proposition 3 *Let G be a network with adjacency matrix $A(G)$. Then*

(i) G is bipartite if and only if $\lambda_1 = -\lambda_n$.

(ii) $\sum_{i=1}^{n} \lambda_i^k = trA^k$, *where* trM *which is the trace of a matrix* M *is equal to the sum of all entries of the main diagonal of* M.

(iii) If H *is subnetwork of a* G, *then*

$$\lambda_{\min}(G) \leq \lambda_{\min}(H) \leq \lambda_{\max}(H) \leq \lambda_{\max}(G).$$

In fact, (ii) of Proposition 3 is true for the Laplacian and normal matrices, while (iii) of Proposition 3 is not true for the normal matrix. For example, let G be a complete network of order 3 and H be a subnetwork of G by deleting an edge. Then

$$\begin{aligned}\lambda_1(G) = 2, \lambda_2(G) = \lambda_3(G) = -1, \lambda_1(H) \\ = \sqrt{2}, \lambda_2(H) = 0, \lambda_3(H) - \sqrt{2},\end{aligned}$$

$$\begin{aligned}\mu_1(G) = 3 = \mu_2(G) = 3, \mu_3(G) = 0, \mu_1(H) \\ = 3, \mu_2 = 1, \mu(H) = 0,\end{aligned}$$

$$\begin{aligned}\nu_1(G) = 1, \nu_2(G) = \nu_3 = -\frac{1}{2}, \nu_1(H) \\ = 1, \nu_2(H) = 0, \nu_3(H) = -1.\end{aligned}$$

Laplacian Spectrum

The Laplacian matrix of a network dates back to the Kirchhoff matrix-tree theorem. The discrete graph Laplacian shares many important properties with the well-known continuous Laplacian operator of mathematical physics. It is easy to see that the Laplacian matrix $L(G)$ of a network is symmetric positive semi-definite, that zero is its smallest eigenvalue, and that this eigenvalue corresponds to the eigenvector $x = (1, \ldots, 1)^T$. One of the most important results is the following:

Proposition 4 *Let* G *be a graph of order* n *with the Laplacian matrix* L. *Then the number of nonidentical spanning trees of* G *is equal to any cofactor of* L. *Moreover, the number of nonidentical spanning trees of* G *is equal to* $\frac{1}{n}\mu_1 \cdots \mu_{n-1}$.

From this proposition, we can see that the second smallest eigenvalue is positive if and only if G has a spanning tree. In other words, G is connected if and only if G has only one zero Laplacian eigenvalue. Further, there is a relationship between the number of zero eigenvalues and the number of connected components of a graph.

Proposition 5 *Let* G *be a simple network with the Laplacian matrix* L. *Then the number of zero eigenvalues of* L *is equal to the number of connected components of* G.

This proposition asserts that G is connected if and only if $\mu_{n-1} > 0$. Hence, Fiedler (1973) called μ_{n-1} the algebraic connectivity of G and denoted α or $\alpha(G)$. He also proved the following:

Proposition 6 *Let* G *be a simple graph other than the complete graph. Then the algebraic connectivity is no more than the vertex connectivity. In other words, the algebraic connectivity of* G *is bounded above by the vertex connectivity of* G.

This proposition suggests that the algebraic connectivity is a suitable measure for connectivity of a network.

Normal Spectrum

It is easy to see that the normal matrix of G $N = D(G)^{-1}A(G)$ is a stochastic matrix and serves as the probability transition matrix of a random walk, where $D(G)$ is degree diagonal matrix and $A(G)$ is the adjacency matrix of G. Consider a random walk on a network G, starting at a node v_i, at each step to each neighbor with probability $1/d(v_i)$, where $d(v_i)$ is the degree of vertex v_i. Random walks arise in many models in mathematics and physics and have important algorithmic applications. They can be used to reach "obscure" parts of large sets and also generate random elements in large and complicated sets. We observe that the random walk on a connected network is a Markov chain and the probability distribution is proportional to the degree distribution. But a major problem is how to determine the number of steps k required for the distributions of a random walk to become close to its stationary distribution, given an arbitrary initial distribution. Here the second modular eigenvalue ν of N plays an important role in the analysis of rapidly mixing Markov chains: if ν is far from 1, i.e., there is a large

eigenvalue gap, then the walk quickly forgets where it started. If ν closes to 1, then there must be parts of the network that are not easy to reach in a random walk, implying long paths or a nearly disconnected network. An important measure of the speed of convergence, called the relative pointwise distance, is given by

$$\Delta(k) = \max_{x,y} \frac{N^k(y,x) - \pi(x)}{\pi(x)},$$

where $N^k(y,x)$ is value of the (y,x) entry of N^k and π is the stationary distribution of the random walk. Chung (1997) showed that

$$\Delta(t) \leq e^{-t\mu} \frac{\mathrm{vol}(G)}{\min_x d_x},$$

where vol(G) is the sum of all degrees in G.

E et al. (2008) and Li et al. (2009) proposed several effective algorithms for network partition based on the framework of optimal predictions and probabilistic framework which is related to a discrete-time Markov chain with the normal matrix N of a network.

Finding Community Structure

A common feature of many networks in biological and social system is "community structure," which means that network nodes can be divided into groups, with dense connections within groups and sparse connections between them. For example, in the friendship school studied by Moody (2001), one of the principal divisions in the network is by individuals' race. The analysis of the community structure in large collaboration networks can be used to reveal the informal organization and the nature of information flows through the whole system. It will be of interest and practical importance if we are able to find community structure from the networks. Several methods to detect community structure have already been proposed. Girvan and Newman (2002) proposed a fast and effective algorithm, based on the link betweenness, which measures the fraction of all shortest paths passing through a given link. But it does not give an indication of the resolution of the clustering. An alternative way to deal with the communities is by spectral analysis. Further, Newman (2006a,b) proposed a number of possible algorithms for detecting community structure by means of the Laplacian eigenvectors. Recently, Bickel and Chen (2009) proposed the Random Graph Models which are aimed at unifying points of view and analyses of networks from social sciences. If a network has k clearly distinct communities, then the largest $k-1$ eigenvalues of the normal matrix is close to 1, the other eigenvalues being far from 1. Hence, there exists an eigenvector among $k-1$ eigenvectors corresponding to the largest $k-1$ eigenvalues, whose components are approximatively constant values on nodes belonging to the same community. Servedio et al. (2004) proposed an optimization problem based on the matrix of a network. The objective function is

$$z(x) = \frac{1}{2} \sum_{(i,j)\in E(G)} (x_i - x_j)^2 w_{ij},$$

where the sum is taken over all edges $(i,j) \in E(G)$ and x_i are values assigned to the nodes, with the constraint function

$$\sum_{i,j,=1}^{n} x_i x_j m_{ij} = 1,$$

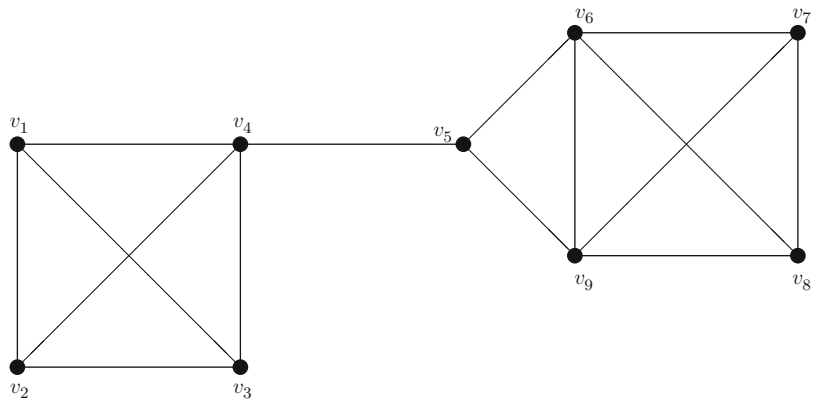
where m_{ij} are elements of a given symmetric matrix M. Then the stationary points of z are the solutions of

$$(D - A)x = \lambda M x,$$

where D is the diagonal matrix and λ is a Lagrange multiplier. If we choose $M = D$, then the solutions become the eigenvalue problem of $D^{-1}Wx = (1-2\lambda)x$. If we choose $M = I$, then the solutions becomes the Laplacian eigenvalue problem $(D - W)x = \lambda x$.

For example, the graph in Fig. 2 below,

H: 9 vertices and 15 edges



Spectral Analysis, Fig. 2 Graph H

Then the Laplacian matrix of H is

$$L(H) = \begin{pmatrix} 3 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 3 & -1 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & -1 & 4 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 3 & -1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & -1 & 4 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & -1 & -1 & 3 & -1 \\ 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 & 4 \end{pmatrix},$$

where vertex v_i is corresponding to the $i-$th rows in H. The eigenvector of H corresponding to the second smallest eigenvalue 0.2783 is (0.3911, 0.3911, 0.3911, 0.2822, −0.1228, −0.3082, −0.3581, −0.3581, −0.3082)^T. Clearly, the positive components of the eigenvector correspond with the nodes v_1, v_2, v_3, v_4 and the negative components of the eigenvector correspond with nodes v_5, v_6, v_7, v_8, v_9. Obviously the nodes of the network can be divided two groups, with one group v_1, v_2, v_3, v_4 and the other group v_5, v_6, v_7, v_8, v_9. In addition, Chauhan et al. (2009) investigate the properties of spectra of networks with community structure. Nascimento and de Carvalho (2011) presented a survey of graph clustering algorithms and different graph clustering formulations in literature, while Newman (2012) discussed the relations among communities, modules, and large-scale structure in networks. Wu et al. (2011) presented a graph partition algorithm adjcluster based on line orthogonality in adjacency eigenspace.

Eigenvector Centrality Structure

The concept of centrality in a network plays an important role in the analysis of its structure. However, there are many different features which have been used to create measures of a centrality. Ruhnau (2000) gave the following definition:

Definition 1 Let $G = (V, E)$ be a connected network with $|V| = n$ and nc be a function which assigns a real value to every node of G. $nc(v_i)$ is called a node centrality of node v_i if

(i) $nc(v_i) \in [0, 1]$ for every $v_i \in V$.

(ii) $nc(v_i) = 1$ if and only if G is the star $S_{1,n-1}$ and $i = 1$.

Bonacich (1972) defines the centrality $c(v_i)$ of a node v_i to be a positive multiple of the sum of adjacent centralities, i.e.,

$$\lambda c(v_i) = \sum_{j=1}^{n} a_{ij} c(v_j), \quad \forall\, i,$$

where $A = (a_{ij})$ is the adjacency matrix of a network. The equations are equivalent to the eigenvalue-eigenvector problem of A. By the Perron-Frobenius theorem, there exists, for connected graphs, an eigenvector corresponding to the largest eigenvalue, with all positive

entries. The entry $c(v_i)$ is called the eigenvector centrality of node v_i. Then the function

$$nc_e(v_i) \equiv \frac{\sqrt{2}c(v_i)}{\sqrt{\sum_{i=1}^{n} c(v_i)^2}}$$

is node centrality. Ruhnau analyzes the structure of networks by using several centrality concepts, including degree centrality, closeness, and betweenness eigenvector centrality. On the other hand, Van Mieghem et al. (2010) established some relationships among the spectrum, the maximum modularity, and assortativity. In particular, they showed that the maximum modularity increases as the number of clusters decreases, and the average hop count and the effective graph resistance increase with increasing assortativity.

Conclusions

In this entry, we have described some properties of three kinds of matrices associated with a network, which are used to analyze the topological structure, community structure, and centrality.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No:10971137 and 11271256)

Cross-References

- Clustering Algorithms
- Eigenvalues, Singular Value Decomposition
- Iterative Methods for Eigenvalues/Eigenvectors
- Matrix Algebra, Basics of
- Matrix Decomposition
- Probability Matrices
- Ranking Methods for Networks
- Spectral Evolution of Social Networks

References

Bickel PJ, Chen A (2009) A nonparametric view of network models and Newman-Girvan and other modularities. Proc Natl Acad Sci USA 106:21068–21073

Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U (2006) Complex networks: structure and dynamics. Phys Rep 424:275–308

Bonacich P (1972) Factoring and weighting approaches to status scores and clique identification. J Math Sociol 2113–120

Chauhan S, Girvan M, Ott E (2009) Spectral properties of networks with community structure. Phys Rev E 80:0561104

Chung FRK (1997) Spectral graph theory. AMS, Providence

Cvetković D, Doob M, Sachs H (1995) Spectra of graphs-theory and applications, 3rd edn. Academic, New Work

E W, Li T, Vanden-Eijnden E (2008) Optimal partition and effective dynamics of complex networks. Proc Natl Acad Sci U S A 105:7907–7912

Fortunato S (2010) Community detection in graphs. Phys Rep 48:75–174

Fiedler M (1973) Algebra connectivity of graphs. Czechoslovak Math J 23(98):298–305

Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci USA 99:7821

Gkantsidis C, Mihail M, Zegura E (2003) Spectral analysis of internet topologies. In: IEEE INFOCOM, San Franciso

Li T, Liu J, E W (2009) Probabilistic framework for network partition. Phys Rev E 80:026106

Moody J (2001) Race, school integration, and friendship segregation in America. Am J Sociol 107:679–716

Nascimento MCV, de Carvalho ACPF (2011) Spectral methods for graph clustering-a survey. Eur J Oper Res 211:221–231

Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45(2)167–245

Newman MEJ (2006a) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74:036104

Newman MEJ (2006b) Modularity and community structure in networks. Proc Natl Acad Sci USA 103:8577–8582

Newman MEJ (2012) Communities modules and large-scale structure in networks. Nat Phys 8:25–31

Ruhnau B (2000) Eigenvector-centrality – a node-centrality? Soc Netw 22:357–365

Seary AJ, Richards WD (2005) Spectral methods for analyzing and visualizing networks: an introduction. In: Breiger R, Carley KM, Pattison P (eds) Dynamic social network modeling and analysis. National Academies, Washington, DC, pp 209–228

Servedio VDP, Colaiori F, Capocci A, Caldarelli G (2004) Community structure from spectral properties in complex network. In: Mendes JFF, Dorogovtsev SN, Abreu

FV, Oliveira JG (eds) Science of complex networks: from biology to the internet and WWW; CNRT, Aveiro, pp 277–286
Scott J (2000) Social network analysis: a handbook. Sage, London
Van Mieghem P, Ge X, Schumm P, Trajanovski S, Wang H (2010) Spectral graph analysis of modularity and assortativity. Phys Rev E 82:056113
Wasserman S, Faust K (1994) Social network analysis. Cambridge University Press, Cambridge
Wu L, Ying X, Wu X, Zhou Z-H (2011) Line orthogonality in adjacency eigenspace with application to community partition. In: Proceedings of the 22nd international joint conference on artificial intelligence (IJCAI11), Barcelona, 16–22 July 2011

Spectral Evolution Model

▸ Spectral Evolution of Social Networks

Spectral Evolution of Social Networks

Jérôme Kunegis
Institute for Web Science and Technologies, University of Koblenz–Landau, Koblenz, Germany

Synonyms

Spectral evolution model

Glossary

Adjacency Matrix A characteristic matrix of a social network, typically denoted $\mathbf{A}$. If the social network contains n persons, the adjacency matrix is a 0/1 $n \times n$ that contains 1 in the entries $\mathbf{A}_{ij}$ that correspond to an edge $\{i, j\}$ and 0 otherwise

Eigenvalue Decomposition A decomposition of a square matrix giving $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}$, in which $\mathbf{U}$ contains the eigenvectors of $\mathbf{A}$ and $\mathbf{\Lambda}$ contains the eigenvalues

Singular Value Decomposition A decomposition of any matrix giving $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$, in which $\mathbf{\Sigma}$ contains the singular values of $\mathbf{A}$

Spectral Evolution Model The model that states that over time, eigenvectors stay constant and eigenvalues change

Spectrum The set of eigenvalues or singular values of a matrix

Definition

The term *spectral evolution* describes a model of the evolution of network based on matrix decompositions. When applied to social networks, this model can be used to predict friendships, recommend friends, and implement other learning problems.

Introduction

The analysis of the evolution of social networks is an important field of study in the areas of information retrieval, data mining, recommender systems, and network science. As an example, models of the evolution of social networks can be used to solve the problem of link prediction, i.e., to predict which edges will appear in a network in the future (Liben-Nowell and Kleinberg 2003). Another common problem associated to models of network evolution is the friend recommendation problem, in which users of social networking sites are recommended to other users.

The spectral evolution model describes the evolution of network using matrix decompositions, in particular the eigenvalue and singular value decompositions of matrices associated with a network, such as the adjacency matrix and the Laplacian matrix. In its most generic version, the spectral evolution is based on the eigenvalue decomposition of the symmetric adjacency matrix of an undirected social network and can be stated in terms of eigenvectors and eigenvalues. The spectral evolution model then asserts that over time, the eigenvectors stay constant and the eigenvalues grow. Other similar formulations exist for other matrix decompositions, other

characteristic graph matrices, and other types of social networks, such as directed networks.

Historical Background

In order to analyze graphs, algebraic graph theory is a common approach. In algebraic graph theory, a graph with n vertices is represented by an $n \times n$ matrix called the adjacency matrix, from which other matrices can be derived.

The edge set of an undirected graph $G = (V, E)$ can be represented by a matrix whose characteristics follow those of the graph. An unweighted undirected graph on n vertices can be represented by an $n \times n$ $0/1$ matrix $\mathbf{A}$ defined by

$$\mathbf{A}_{ij} = \begin{cases} 1 & \text{if } \{i, j\} \in E \\ 0 & \text{otherwise.} \end{cases}$$

The matrix $\mathbf{A}$ is called the adjacency matrix of G.

Spectral graph theory is a branch of algebraic graph theory that applies matrix decompositions to characteristic graph matrices in order to study a graph's properties (Chung 1997; Cvetković et al. 1997). The word *spectral* refers to the spectrum of networks, which is given by the eigenvalue decomposition of a graph's adjacency or Laplacian matrix. Spectral graph theory can be used to study graph properties such as connectivity, centrality, balance, and clustering.

A square symmetric matrix $\mathbf{A}$ can be written in the following way:

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}} \tag{1}$$

where $\mathbf{U}$ is an $n \times n$ orthogonal matrix and $\mathbf{\Lambda}$ is an $n \times n$ diagonal matrix. A matrix $\mathbf{U}$ is orthogonal when $\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}$ or equivalently when $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}$. Another characterization of an orthogonal matrix $\mathbf{U}$ is that its columns are pairwise orthogonal vectors and each has unit norm. The values $\mathbf{\Lambda}_{kk}$ are the eigenvalues of $\mathbf{A}$, and the columns of $\mathbf{U}$ are its eigenvectors. We will designate the eigenvalues by $\lambda_k = \mathbf{\Lambda}_{kk}$ and the eigenvectors by $\mathbf{u}_k = \mathbf{U}_{\cdot k}$ for $1 \leq k \leq n$.

A certain number of interesting graph properties can be described spectrally, such as connectivity (Mohar 1991), centrality (Brin and Page 1998), conflict and balance (Kunegis et al. 2010c), and clustering (Luxburg 2007). Spectral transformations were considered in 2009 in Kunegis and Lommatzsch. The spectral evolution model itself was introduced in 2010 (Kunegis et al. 2010b) and in detail in 2011 (Kunegis 2011).

The Spectral Evolution Model

We first describe the spectral evolution model for unweighted, undirected social networks based on the eigenvalue decomposition of the adjacency matrix, and will then review extensions of it to other types of networks and other characteristic graph matrices and decompositions.

Let $G_t = (V, E_t)$ be a social network that evolves over time, at time t. We assume that the set of vertices V is constant and will only consider evolving sets of edges E_t. Let $\mathbf{A}_t$ be the adjacency matrix of the social network as time t. We can now consider the eigenvalue decomposition

$$\mathbf{A}_t = \mathbf{U}_t \mathbf{\Lambda} \mathbf{U}_t^{\mathrm{T}}.$$

A priori, this eigenvalue decomposition will change from timepoint to timepoint. The spectral evolution model can now be stated as:

Definition 1 (Spectral evolution model) A network that changes over time is said to follow the spectral evolution model when its eigenvalues $\mathbf{\Lambda}_t$ evolves while its eigenvectors $\mathbf{U}_t$ stay approximately constant.

The spectral evolution model is a quantitative statement: The eigenvectors do not need to be exactly constant. In the general case, the spectral evolution model can be stated to hold when the eigenvectors change less than predicted by a random graph model and the eigenvalues change more than predicted by a random graph model.

Relationship to Link Prediction

The spectral evolution model can be compared to a number of link prediction models that are

special cases of it. A link prediction function is a function used to implement the link prediction problem in social networks (Liben-Nowell and Kleinberg 2003).

Let $\mathbf{A}_1$ be the current adjacency matrix of the social network. A link prediction function is a function

$$f : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n}$$

that maps the current adjacency matrix $\mathbf{A}_1$ to its predicted value in the future $\mathbf{A}_2$. We will call f a spectral transformation when it can be expressed using the eigenvalue decomposition $\mathbf{A}_1 = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}$ as

$$f(\mathbf{A}_1) = \mathbf{U} f(\mathbf{\Lambda}) \mathbf{U}^{\mathrm{T}}.$$

The spectral evolution model can then be stated as:

Definition 2 (**Spectral evolution model, alternative definition**) A network follows the spectral evolution model when a future value of its adjacency matrix can be predicted by application of a spectral evolution function.

Friend of a Friend Count

For instance, the friend of a friend count (or *common neighbor count*) is one such model: Given two users $i, j \in V$, the number of common friends of i and j can be used as a link prediction function. The higher the number of common neighbors, the more likely it is that an edge will appear between them in the social network. An example of that method is used on the social network Facebook (www.facebook.com) for recommending new friends. Mathematically, the common neighbor count can be expressed using the social network's adjacency matrix as the square $\mathbf{A}^2$. In fact, the entry $(\mathbf{A}^2)_{ij}$ equals the number of common friends of users i and j. Assuming that the probability that an edge will appear between i and j is proportional to $(\mathbf{A}^2)_{ij}$, there is a constant α such that the adjacency matrix in the future can be expressed as a spectral transformation of the original adjacency matrix:

$$\begin{aligned}\mathbf{A}_2 &= \mathbf{A} + \alpha \mathbf{A}^2 \\ &= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}} + \alpha(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}})^2 \\ &= \mathbf{U}(\mathbf{\Lambda} + \alpha\mathbf{\Lambda}^2)\mathbf{U}^{\mathrm{T}}.\end{aligned}$$

Thus, the friend of a friend model predicts a spectral transformation of $\mathbf{A}$ and thus justifies the spectral evolution model of social networks. This argument can be extended to the *friend-of-a-friend-of-a-friend* model and models based paths of any length.

Graph Kernels

A related class of link prediction functions are given by graph kernels. The exponential graph kernel is defined as the exponential function of the adjacency matrix (Kondor and Lafferty 2002):

$$e^{\alpha\mathbf{A}} = \mathbf{I} + \alpha\mathbf{A} + \frac{\alpha^2}{2}\mathbf{A}^2 + \frac{\alpha^3}{6}\mathbf{A}^3 + \ldots.$$

The Neumann graph kernel is defined using matrix inversion (Kandola et al. 2002):

$$(\mathbf{I} - \alpha\mathbf{A})^{-1} = \mathbf{I} + \alpha\mathbf{A} + \alpha^2\mathbf{A}^2 + \alpha^3\frac{1}{6}\mathbf{A}^3 + \ldots.$$

Both graph kernels can be expressed as a spectral transformation:

$$\begin{aligned}e^{\alpha\mathbf{A}} &= \mathbf{U}e^{\alpha\mathbf{\Lambda}}\mathbf{U}^{\mathrm{T}} \\ (\mathbf{I} - \alpha\mathbf{A})^{-1} &= \mathbf{U}(\mathbf{I} - \alpha\mathbf{\Lambda})^{-1}\mathbf{U}^{\mathrm{T}}\end{aligned}$$

Thus, the two graph kernels justify the spectral evolution model in the sense that if they produce accurate link predictions, the social network will grow according to the spectral evolution model.

Preferential Attachment

Preferential attachment is a simple link prediction model based on the idea that the probability of a new link being formed is proportional to the degrees of the nodes it connects. This idea can be extended to the decomposition of a graph's adjacency matrix, resulting in the latent preferential attachment model, first described in Kunegis (2011).

The eigenvalue decomposition of $\mathbf{A}$ can be written as a sum of rank-one matrices:

$$\mathbf{A} \approx \sum_{k=1}^{r} \lambda_k \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}}$$

where r is the rank of the decomposition. The usual interpretation of a matrix factorization is that each latent dimension k represents a *topic* in the network. Then $\mathbf{U}_{ik}$ represents the importance of vertex i in topic k, and λ_k represents the overall importance of topic k. Each rank-one matrix $\mathbf{A}^{(k)} = \lambda_k \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}}$ can be interpreted as the adjacency matrix of a weighted graph. Now, assume that preferential attachment is happening in the network, but restricted to the subgraph G_k. Then the probability of the edge $\{i, j\}$ appearing will be proportional to $d_k(i)d_k(j)$, where $d_k(i)$ is the degree of node i in the graph G_k. This degree can be written as the sum over edge weights in G_k:

$$\begin{aligned} d_k(i) &= \sum_l \mathbf{A}_{il}^{(k)} = \sum_l \lambda_k \mathbf{U}_{ik} \mathbf{U}_{lk} \\ &= \mathbf{U}_{ik} \lambda_k \sum_l \mathbf{U}_{lk} \sim \mathbf{U}_{ik}. \end{aligned}$$

In other words, $d_k(i)$ is proportional to $\mathbf{U}_{ik}$. Therefore, the preferential attachment value is proportional to the corresponding entry in $\mathbf{A}^{(k)}$:

$$d_k(i)d_k(j) \sim \mathbf{U}_{ik}\mathbf{U}_{jk}.$$

These values can be aggregated into a matrix $\mathbf{P}^{(k)}$ giving the preferential attachment values for all pairs (i, j):

$$\mathbf{P}^{(k)} \sim \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}}.$$

Assuming that a preferential attachment process is happening for each subgraph G_k separately, with a weight ε_k depending on the topic k, then the overall preferential attachment prediction can be written as $\mathbf{P} = \sum_k \varepsilon_k \mathbf{u}_k \mathbf{u}_k^{\mathrm{T}}$. Here, we replace proportionality by equality since the proportionally constants are absorbed by the constants ε_k. The matrix $\mathbf{P}$ can then be written in the following form, giving its eigenvalue decomposition $\mathbf{P} = \mathbf{U}\mathbf{E}\mathbf{U}^{\mathrm{T}}$, where $\mathbf{E}$ is the diagonal matrix containing the individual topic weights $\mathbf{E}_{kk} = \varepsilon_k$. This prediction matrix is a spectral transformation of the adjacency matrix $\mathbf{A}$. Under this model, network growth can be interpreted as the replacement of the eigenvalues $\mathbf{\Lambda}$ by $\mathbf{\Lambda} + \mathbf{E}$:

$$f(\mathbf{A}_1) = \mathbf{A} + \mathbf{P} = \mathbf{U}(\mathbf{\Lambda} + \mathbf{E})\mathbf{U}^{\mathrm{T}}.$$

Since the values $\mathbf{E}$ are not modeled by the latent preferential attachment model, every spectral transformation can be interpreted as latent preferential attachment, and thus, the latent preferential attachment model is equivalent to the spectral evolution model.

Learning Spectral Transformations

Under the assumption that a social network evolves according to the spectral evolution model, the best possible link prediction function can be learned using curve fitting (Kunegis and Lommatzsch 2009).

Given the current adjacency matrix $\mathbf{A}_1$ and the future adjacency matrix $\mathbf{A}_2$, the best possible link prediction function f that maps $\mathbf{A}_1$ to $\mathbf{A}_2$ is given by the following minimization problem:

$$\min_f \| f(\mathbf{A}_1) - \mathbf{A}_2 \|_{\mathrm{F}}.$$

Using the eigenvalue decomposition $\mathbf{A}_1 = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}$ of rank r, this problem is equivalent to

$$\min_f \| f(\mathbf{\Lambda}) - \mathbf{U}^{\mathrm{T}}\mathbf{A}_2\mathbf{U} \|_{\mathrm{F}}.$$

Since $\mathbf{\Lambda}$ is diagonal and $f(\mathbf{\Lambda})$ is diagonal too, only the diagonal elements of $\mathbf{U}^{\mathrm{T}}\mathbf{A}_2\mathbf{U}$ influence the minimization problem. Thus, the minimization problem is equivalent to

$$\min_f \sum_{i=1}^{r} (f(\mathbf{\Lambda}_{ii}) - (\mathbf{U}^{\mathrm{T}}\mathbf{A}_2\mathbf{U})_{ii})^2. \qquad (2)$$

This is a one-dimensional curve-fitting problem with r parameters and can be solved efficiently. For each spectral link prediction function, the

corresponding spectral transformation function can be fitted to solve the optimization problem in Eq. 2, learning its parameters in the process, for instance, the parameter α for the exponential and Neumann graph kernels.

An alternative way of learning spectral transformations is based on the extrapolation of the eigenvalues into the future (Kunegis et al. 2010b). This gives new values for the eigenvalues, which can be combined with the unchanging eigenvectors to give the predicted value of the adjacency matrix.

Tests of the Spectral Evolution Model

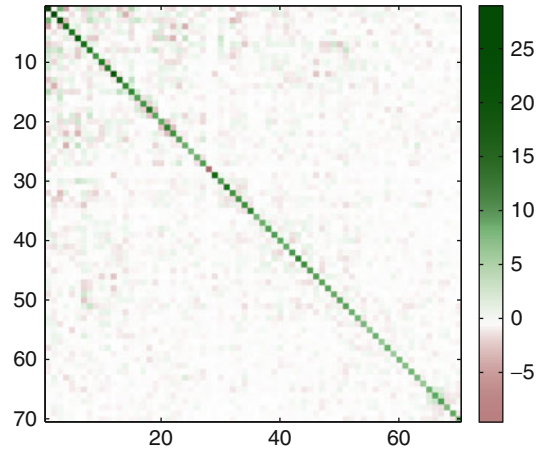
In addition to the fact that the spectral evolution model has known link prediction functions as special cases, it can be verified experimentally by measuring the change in the eigenvectors and eigenvalues of actual social networks, of which the temporal evolution is known. These observations can then be combined with the changes predicted by a random graph growth model in which edges are added randomly to a network.

When adding a small random perturbation $\mathbf{E}$ of size $\|\mathbf{E}\|_{\mathrm{F}} = \varepsilon$ to the adjacency matrix $\mathbf{A}$ to give $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{E}$, the expected change in the new eigenvalues $\tilde{\mathbf{\Lambda}}$ and the new eigenvectors $\tilde{\mathbf{U}}$ is given by

$$\left\| \mathbf{\Lambda} - \tilde{\mathbf{\Lambda}} \right\|_{\mathrm{F}} = O(\varepsilon^2)$$
$$\left| \mathbf{U}_{\cdot k} \cdot \tilde{\mathbf{U}}_{\cdot k} \right| = O(\varepsilon).$$

These results can be shown by a perturbation argument (Stewart 1990) and ultimately can be derived from theorems by Weyl (1912) and Wedin (1972). As a result, eigenvectors are expected to change faster than eigenvalues for random additions to the adjacency matrix, justifying the spectral evolution model for social networks.

When the growth of actual social network can be observed over time, the spectral evolution model can be verified directly. As an example for a method of achieving this, we describe the spectral diagonality test. The spectral diagonality test can be computed from the snapshot of a network at two different times 1 and 2, using the adjacency matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ (Kunegis et al. 2010b).



Spectral Evolution of Social Networks, Fig. 1 The spectral diagonality test matrix $\mathbf{\Delta}$ for a subset of the Facebook social network (Viswanath et al. 2009). Since the matrix is almost diagonal, the test shows that the evolution of that subset of Facebook follows the spectral evolution model

Using the eigenvalue decomposition $\mathbf{A}_1 = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}$, the spectral diagonality test consists in verifying the diagonality of the matrix $\mathbf{\Delta} = \mathbf{U}^{\mathrm{T}}\mathbf{A}_2\mathbf{U}$. If the matrix $\mathbf{\Delta}$ is diagonal, the evolution of the social network is perfectly spectral. In practice, $\mathbf{\Delta}$ is not perfectly spectral, but almost so. An example of the matrix $\mathbf{\Delta}$ is given Fig. 1 for a subset of the Facebook social network (Viswanath et al. 2009). In this instance of the spectral diagonality test, $\mathbf{\Delta}$ is indeed almost diagonal, and the evolution of that network can be concluded to follow the spectral evolution model.

Normalized Adjacency Matrix

The spectral evolution model can be extended to the normalized adjacency matrix, defined as $\mathbf{N} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$, in which $\mathbf{D}$ denotes the diagonal degree matrix with $\mathbf{D}_{ii}$ being the degree of node i. In this definition, we assume that the social network does not contain any isolated nodes, i.e., users without friends.

The theory of spectral network evolution can be extended to using the matrix $\mathbf{N}$ instead of the matrix $\mathbf{A}$ without much change. A key difference to the unnormalized case is in the evolution of the eigenvalues over time: While the eigenvalues of $\mathbf{A}$ grow in the general case, the eigenvalues of $\mathbf{N}$ cannot grow without bounds, as by construction,

they lie in the interval $[-1, +1]$. In fact, the eigenvalues of $\mathbf{N}$ will typically shrink over time (Kunegis et al. 2012).

Another difference in using $\mathbf{N}$ over $\mathbf{A}$ lies in the interpretation of corresponding link prediction functions. For instance, the exponential and Neumann graph kernels give the following link prediction functions in the normalized case:

$$e^{\alpha\mathbf{N}} = \mathbf{I} + \alpha\mathbf{N} + \frac{\alpha^2}{2}\mathbf{N}^2 + \frac{\alpha^3}{6}\mathbf{N}^3 + \dots$$

$$(\mathbf{I} - \alpha\mathbf{N})^{-1} = \mathbf{I} + \alpha\mathbf{N} + \alpha^2\mathbf{N}^2 + \alpha^3\frac{1}{6}\mathbf{N}^3 + \dots.$$

Laplacian Matrix

The Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{A}$, too, can be used as the basis for studying the spectral evolution of networks. This leads to a more complicated situation. Although the eigenvectors do stay constant in the general case, the eigenvalues will not change continuously, but grow in steps, which makes the diagonality test impracticable. However, link prediction function can still be used in that case. These include the regularized commute-time kernel $(\mathbf{I} + \alpha\mathbf{L})^{-1}$ and the heat diffusion kernel $e^{-\alpha\mathbf{L}}$.

Bipartite Networks

Bipartite networks are networks in which the set of nodes V can be partitioned into two sets $V = V_1 \cup V_2$ such that all edges connect a node in V_1 with a node in V_2. Social networks are not bipartite in the general case, since they contain triangles. Still, many bipartite networks can be found in social media, for instance, user–group inclusion networks or user–item rating networks. In such networks, the spectral evolution model can be applied as is with good results. However, a simplification of the expression is possible, due to the special structure of the networks (Kunegis et al. 2010a).

The adjacency matrix $\mathbf{A}$ of a bipartite network can always be written as

$$\mathbf{A} = \begin{bmatrix} \mathbf{0} & \mathbf{B} \\ \mathbf{B}^{\mathrm{T}} & \mathbf{0} \end{bmatrix}$$

for a matrix $\mathbf{B}$ of size $|V_1| \times |V_2|$. The matrix $\mathbf{B}$ is then called the biadjacency matrix of the network. This can be exploited to reduce the eigenvalue decomposition of $\mathbf{A}$ to the singular value decomposition of $\mathbf{B}$. Given the singular value decomposition $\mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathrm{T}}$, the eigenvalue decomposition of $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{bmatrix} \bar{\mathbf{U}} & \bar{\mathbf{U}} \\ \bar{\mathbf{V}} & -\bar{\mathbf{V}} \end{bmatrix} \begin{bmatrix} +\mathbf{\Sigma} & \mathbf{0} \\ \mathbf{0} & -\mathbf{\Sigma} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{U}} & \bar{\mathbf{U}} \\ \bar{\mathbf{V}} & -\bar{\mathbf{V}} \end{bmatrix}^{\mathrm{T}} \quad (3)$$

with $\bar{\mathbf{U}} = \mathbf{U}/\sqrt{2}$ and $\bar{\mathbf{V}} = \mathbf{V}/\sqrt{2}$. In this decomposition, each singular value σ corresponds to the eigenvalue pair $\{\pm\sigma\}$. Odd powers of $\mathbf{A}$ then have the form

$$\mathbf{A}^{2k+1} = \begin{bmatrix} \mathbf{0} & (\mathbf{B}\mathbf{B}^{\mathrm{T}})^k\mathbf{B} \\ (\mathbf{B}^{\mathrm{T}}\mathbf{B})^k\mathbf{B}^{\mathrm{T}} & \mathbf{0} \end{bmatrix},$$

where the alternating power $(\mathbf{B}\mathbf{B}^{\mathrm{T}})^k\mathbf{B}$ can be explained by the fact that in the bipartite network, a path will follow edges from one vertex set to the other in alternating directions, corresponding to the alternating transpositions of $\mathbf{B}$.

Thus, it is sufficient, in a bipartite network, to consider only odd functions of the biadjacency matrix $\mathbf{B}$. Here, an odd function is to be understood as a function f for which it holds that $f(-\mathbf{A}) = -f(\mathbf{A})$. Examples of resulting odd link prediction functions are the matrix hyperbolic sine $\sinh(\alpha\mathbf{A})$ and the Neumann pseudokernel $\alpha\mathbf{A}(\mathbf{I} - \alpha^2\mathbf{A})^{-1}$. These functions are pseudokernels and not kernels, as they are not positive definite.

Directed Networks

The case of directed networks is more complicated than the other cases, since the eigenvectors of the adjacency matrix are not orthogonal anymore in that general case. Four methods can be used for directed networks:

Ignoring Edge Directions By ignoring edge directions, the problem is reduced to the undirected case. This is sensible in social networks that tend to be symmetric, such as communication networks, but does not give good results in networks that are inherently directed, such as trust networks.

Working on the Bipartite Double Cover By considering the bipartite double cover of a directed network, the problem reduces to the bipartite case. The bipartite cover of a directed graph is constructed by replacing each node by two nodes, one that keeps all in-edges and one that keeps all out-edges. The resulting link prediction methods work well when the primary mechanism of graph growth follows paths of alternating signs. An example of such networks are citation networks, in which co-citation can be interpreted using paths of alternating directions.

Non-orthogonal Decomposition The non-orthogonal eigenvalue decomposition of a directed network can be used with difficulty. Since the matrix $\mathbf{A}$ is asymmetric, the eigenvalue decomposition must be written as $\mathbf{A} = \mathbf{U}\Lambda\mathbf{U}^{-1}$ and will contain complex eigenvalues. The link prediction methods described in the previous sections do not perform well in that case. In the extreme case, if a directed network is acyclic, for instance, a scientific citation network, then all eigenvalues are zero, and all graph kernels and other link prediction methods return only the value zero.

DEDICOM The last variant uses matrix decompositions of the form $\mathbf{A} = \mathbf{U}\mathbf{X}\mathbf{U}^{\mathrm{T}}$ in which $\mathbf{U}$ is orthogonal and $\mathbf{X}$ is not diagonal. Such decompositions are called DEDICOM (decomposition into directed components) (Harshman 1978). This decomposition is not unique, and thus, there are multiple variants of DEDICOMs. In general, the choice of a variant will involve the trade-off between a fast computation and an accurate decomposition. This method is best suited to networks in which directed triangle closing is the main mechanism by which new edges are formed, for instance, in trust networks.

Key Applications

The spectral evolution model can be used to implement link prediction functions which themselves can be used to solve several different kinds of problems in social networks:

- Applying the link prediction problem to an ordinary social network leads to the recommendation of new friends. In this case, edges are unweighted, and the links to be predicted describe the similarity between nodes.
- Trust prediction in a social network consists of predicting trust edges in a directed social network consisting of trust edges. In some cases, distrust edges are additionally known.
- Rating prediction is a special case of link prediction, where edges are weighted. An important application of rating prediction is collaborative filtering, in which the network is either unipartite when users are rated as in dating sites or items are rated as in movie rating sites.
- In a signed network, the prediction of an edge's sign, knowing that the edge is part of the network, is known as the link sign prediction problem.
- To predict future interactions in social networks, for instance, emails or scientific coauthorship, link prediction can be performed in a network with multiple edges.

The spectral evolution model applies equally to all these variants of the link prediction problem, with appropriate choice of matrix and decomposition type.

Future Directions

As of 2012, the link prediction problem in all its variants is not fully covered by research, and new applications are still being published. In particular, the application of social network analysis methods such as the spectral evolution model is increasingly applied to other kinds of networks, such as content networks or hyperlink networks. Another area of research lies in the exploration of more complex matrix decompositions, such as nonnegative decompositions and tensor decompositions.

Acknowledgments

We thank our collaborators on previous work: Christian Bauckhage, Damien Fay, and Andreas Lommatzsch. The author of this

work has received funding from the European Community's Seventh Frame Programme under grant agreement nº 257859, ROBUST.

Cross-References

- ▸ Community Evolution
- ▸ Data Mining
- ▸ Eigenvalues, Singular Value Decomposition
- ▸ Link Prediction: A Primer
- ▸ Matrix Decomposition
- ▸ Network Models
- ▸ Recommender Systems: Models and Techniques
- ▸ Spectral Analysis
- ▸ Temporal Networks

References

Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. Comput Netw ISDN Syst 30(1–7):107–117

Chung F (1997) Spectral graph theory. Am Math Soc

Cvetković D, Rowlinson P, Simić S (1997) Eigenspaces of graphs. Cambridge University Press, Cambridge

Harshman RA (1978) Models for analysis of asymmetrical relationships among *n* objects or stimuli. In: Proceedings of the first meeting of the psychometric society and the society for mathematical psychology, Hamilton, Ontario, Canada

Kandola J, Shawe-Taylor J, Cristianini N (2002) Learning semantic similarity. In: Advances in neural information processing systems, Whistler, British Columbia, Canada, pp 657–664

Kondor R, Lafferty J (2002) Diffusion kernels on graphs and other discrete structures. In: Proceedings of the international conference on machine learning, Sidney, Australia, pp 315–322

Kunegis J (2011) On the spectral evolution of large networks. PhD thesis, University of Koblenz–Landau. http://userpages.uni-koblenz.de/~kunegis/paper/kunegis-phd-thesis-on-the-spectral-evolution-of-large-networks.pdf

Kunegis J, Lommatzsch A (2009) Learning spectral graph transformations for link prediction. In: Proceedings of the international conference on machine learning, Montréal, Québec, Canada, pp 561–568. http://uni-koblenz.de/~kunegis/paper/kunegis-spectral-transformation.pdf

Kunegis J, De Luca EW, Albayrak S (2010a) The link prediction problem in bipartite networks. In: Proceeding of the international conference in information processing and management of uncertainty in knowledge-based systems, Dortmund, Germany, pp 380–389. http://uni-koblenz.de/~kunegis/paper/kunegis-hyperbolic-sine.pdf

Kunegis J, Fay D, Bauckhage C (2010b) Network growth and the spectral evolution model. In: Proceeding of the international conference on information and knowledge management, Toronto, Ontario, Canada, pp 739–748. http://uni-koblenz.de/~kunegis/paper/kunegis-spectral-network-evolution.pdf

Kunegis J, Schmidt S, Lommatzsch A, Lerner J (2010c) Spectral analysis of signed graphs for clustering, prediction and visualization. In: Proceedings SIAM international conferences on data mining, Columbus, Ohio, USA, pp 559–570. http://uni-koblenz.de/~kunegis/paper/kunegis-spectral-analysis-of-signed-graphs.pdf

Kunegis J, Sizov S, Schwagereit F, Fay D (2012) Diversity dynamics in online networks. In: Proceedings of the conference on hypertext and social media, Milwaukee, Wisconsin, USA, pp 255–264. http://userpages.uni-koblenz.de/~kunegis/paper/kunegis-diversity-dynamics-in-online-networks.pdf

Liben-Nowell D, Kleinberg J (2003) The link prediction problem for social networks. In: Proceedings international conference on information and knowledge management, New Orleans, Louisiana, USA, pp 556–559

Luxburg Uv (2007) A tutorial on spectral clustering. Stat Comput 17(4):395–416

Mohar B (1991) The Laplacian spectrum of graphs. Graph Theory Combin Appl 2:871–898

Stewart GW (1990) Perturbation theory for the singular value decomposition. Tech. rep., University of Maryland, College Park

Viswanath B, Mislove A, Cha M, Gummadi KP (2009) On the evolution of user interaction in Facebook. In: Proceeding of the workshop on online social networks, Barcelona, Spain, pp 37–42

Wedin PÅ (1972) Perturbation bounds in connection with singular value decomposition. BIT Numer Math 12(1):99–111

Weyl H (1912) Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differenzialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). Math Ann 71(4):441–479

Spectral Graph Analysis

- ▸ Spectral Analysis

Spectral Network Analysis

▸ Spectral Analysis

Spectral Technique

▸ Spectral Analysis

Stability and Evolution of Scientific Networks

Eugenia Galeota[1], Susanna Liberti[1], Frederic Amblard[2], and Walter Quattrociocchi[3,4]
[1]Department of Biology, University of Rome Tor Vergata, Rome, Italy
[2]IRIT – Universit Toulouse 1 Capitole, Toulouse Cedex 9, France
[3]London Institute of Mathematical Sciences, London, UK
[4]Natwork Department, IMT Alti Studi Lucca, Lucca, Italy

Synonyms

Emergence; Evolving structures; Scientific communities; Social network analysis; Social selection; Temporal metrics; Time-varying graphs

Glossary

APS American Physical Society
DOI Digital Object unique Identifiers associated to papers
Egalitarian Growth The growth benefiting on average equally to each node
Evolution of Social Networks The change in time of the structure of a social network due to changing interactions between the components
TVG Time-Varying Graphs

Introduction

Nowadays one of the most pressing as well as interesting scientific challenges deals with the analysis and the understanding of social systems' dynamics and how these evolve according to the interactions among their components. The efforts in this area strive to understand what are the driving forces behind the evolution of social networks and how they are articulated together with social dynamics – e.g., opinion dynamics, the epidemic or innovation diffusion, the teams formation, and so forth (Deffuant et al. 2001; Moore and Newman 2000; Lelarge 2008; Butts and Carley 2007; Powell et al. 2005; Guimera et al. 2005; Quattrociocchi et al. 2009, 2010). In this paper we approach the challenge of depicting the evolution of social systems from a network science's perspective. As an example, we chose the case of scientific communities by analyzing a portion of the American Physical Society dataset (APS). The analysis addresses the coexistence of coauthorship and citation behaviors of scientists. On the one hand, the studies on scientific network dynamics deal with the understanding of the factors that play a significant role in their evolution, not all of them being neither objective nor rational, e.g., the existence of a star system (Wagner and Leydesdorff 2005; Newman 2001a, 2004a; Jeong et al. 2002), the blind imitation concerning the citations (MacRoberts and MacRoberts 1996), and the reputation and community affiliation bias (Gilbert 1977). On the other hand, having some elements to understand such dynamics could enable for a better detection of the hot topics and of the vivid subfields and how the scientific production is advanced with respect to the selection process inside the community itself. Among the available data to analyze such a system, a subset of the publications in a given field is the most frequently used such as in De Solla Price (1965), Newman (2001b); **?**, Quattrociocchi et al. (2012), Amblard et al. (2011), Santoro et al. (2011), and Radicchi et al. (2009). The scientific publications correspond to the production of such a system and clearly identify who are the producers (the authors), which institution they belong to (the

affiliation), which funded project they are working on (the acknowledgement), and what are the related publications (the citations), having most of the time a public access to these data explain also a part of its frequent use in the analyses of the scientific field. Classical analyses concern either the coauthorships network (Jeong et al. 2002; Newman 2001a) or the citation network (Hummon and Dereian 1989; Redner 2005), more rarely the institutional network (Powell et al. 2005). Moreover, such networks are often considered as static and their structure is rarely analyzed over time (an exception is the one performed by Radicchi et al. (2009), and Leskovec et al. (2005a). The illustrative analysis presented in the paper passes through different data transformations aimed at providing different perspectives on the APS network and its evolution. In Newman (2001a) the network of scientific collaborations, explored upon several databases, shows a clustered and small-world structure Watts (1999) and Tang et al. (2009). Moreover, several differences between the collaborations' patterns of the different fields studied are captured. Such differences have been deepened in Newman (2004a) with respect to the number of papers produced by a given group of authors, the number of collaborations, and the topological distances between scientists. Peltomaki and Alava in Peltomaki and Alava (2006) propose a new emulative model aimed at approximating the growth of scientific networks by incorporating bipartition and sublinear preferential attachment. A model for the self-assembly of creative teams based on three parameters (e.g., team size, the rate of newcomers in the scientific production, and the tendency of authors to collaborate with the same group) has been outlined in Guimera et al. (2005). Connectivity patterns in a citations network have been studied with respect to the development of the DNA theory (Hummon and Dereian 1989). The work of Klemm and Eguiluz (2002) observed that real network (e.g., movie actors, coauthorship in science, and word synonyms) growing patterns are characterized by a clustering trend that reaches an asymptotic value larger than regular lattices of the same average connectivity. In the field of social network analysis, several works have approached the problem of temporal metrics (Holme 2005; Kostakos 2009; Kossinets et al. 2008). The focus is on the definition of instruments able to capture the intrinsic properties of complex systems' evolution, that is, characterizing the interdependencies and the coexistence between local behaviors (interactions) and their global effects (emergence) (Davidsen et al. 2002; Mataric 1992; Woolley 1994; Deffuant et al. 2001; Quattrociocchi et al. 2010). The research approach to characterize the evolution patterns of social networks at the very beginning was mainly based upon simulations, while in the past few years, due to the large availability of real datasets, either the methodology of analysis or the object of research has changed (Taramasco et al. 2010; Leskovec et al. 2007; Kossinets et al. 2008).

Analysis of Scientific Network Dynamics

In this work we present a very basic analysis aiming at understanding the social aspects of the scientific systems by coupling the collaborations between scientists and their effect on the scientific community itself through the citation network. The data to build up the networks analyzed in this work has been extracted from the APS (American Physical Society) dataset, made available upon request by the APS for research purposes. The database contains information about 463,343 articles published on 11 journals of the APS in a time span ranging from 1892 to 2009. For the citations network we used a list of 2,944,144 DOI pairs in which the first DOI identifies an article containing a reference to the article identified by the second DOI. A date flag corresponding to the issue date of the citing article has been associated to each couple of DOIs in the list to represent the citation date. Such information has been obtained from the "Article metadata" part of the database which is divided by journal and provides for each paper the following fields: DOI, journal, volume, issue, first page, and last page OR article ID and number of pages, title, authors, affiliations, publication history, PACS codes, table of contents

heading, article type, and copyright information. The list of authors provided for each DOI has been used to generate the collaboration network where authors of the same paper form small coauthorship cliques. Starting from the metadata a list of 17,069,841 total coauthorships has been generated for 119,172 unique authors' surnames. In order to assign a date to the collaboration, the submission date of the coauthored article has been associated to each couple of authors. The data transformation is performed through the *time-varying graphs* formalism. The *time-varying graph* (TVG) formalism, recently introduced in Casteigts et al. (2010), is a graph formalism based on an *interaction-centric* point of view and offers concise and elegant formulation of temporal concepts and properties (Santoro et al. 2010). Let us consider a set of entities V (or *nodes*), a set of relations E among entities (*edges*), and an alphabet L labeling any property of a relation (*label*), that is, $E \subseteq V \times V \times L$. The set E enables multiple relations between any given pair of entities, as long as these relations have different properties, that is, for any $e_1 = (x_1, y_1, \lambda_1) \in E$, $e_2 = (x_2, y_2, \lambda_2) \in E$, $(x_1 = x_2 \wedge y_1 = y_2 \wedge \lambda_1 = \lambda_2) \implies e_1 = e_2$. Relationships between entities are assumed to occur over a time span $\mathbb{T} \subseteq \mathbb{T}$, namely, the *lifetime* of the system. The temporal domain $\mathbb{T}$ is assumed to be $\mathbb{N}$ for discrete-time systems or $\mathbb{R}$ for continuous-time systems. The time-varying graph structure is denoted by the set $\mathcal{G} = (V, E, \mathbb{T}, \rho, \zeta)$, where $\rho : E \times \mathbb{T} \to \{0, 1\}$, called *presence function*, indicates whether a given edge is present at a given time and $\zeta : E \times \mathbb{T} \to \mathbb{T}$, called *latency function*, indicates the time it takes to cross a given edge if starting at a given date. As in this paper the focus is on the temporal and structural analysis of a social network, we will deliberately omit the latency function and consider TVGs described as $\mathcal{G} = (V, E, \mathbb{T}, \rho)$. Given a TVG $\mathcal{G} = (V, E, \mathbb{T}, \rho)$, one can define the *footprint* of this graph from t_1 to t_2 as the static graph $G^{[t_1,t_2)} = (V, E^{[t_1,t_2)})$ such that $\forall e \in E$, $e \in E^{[t_1,t_2)} \iff \exists t \in [t_1, t_2), \rho(e, t) = 1$. In other words, the footprint aggregates interactions over a given time window into static graphs. Let the lifetime $\mathbb{T}$ of the time-varying graph be partitioned in consecutive subintervals $\tau = [t_0, t_1), [t_1, t_2) \ldots [t_i, t_{i+1}), \ldots$, where each $[t_k, t_{k+1})$ can be noted τ_k. We call *sequence of footprints* of $\mathcal{G}$ according to τ the sequence $SF(\tau) = G^{\tau_0}, G^{\tau_1}, \ldots$.

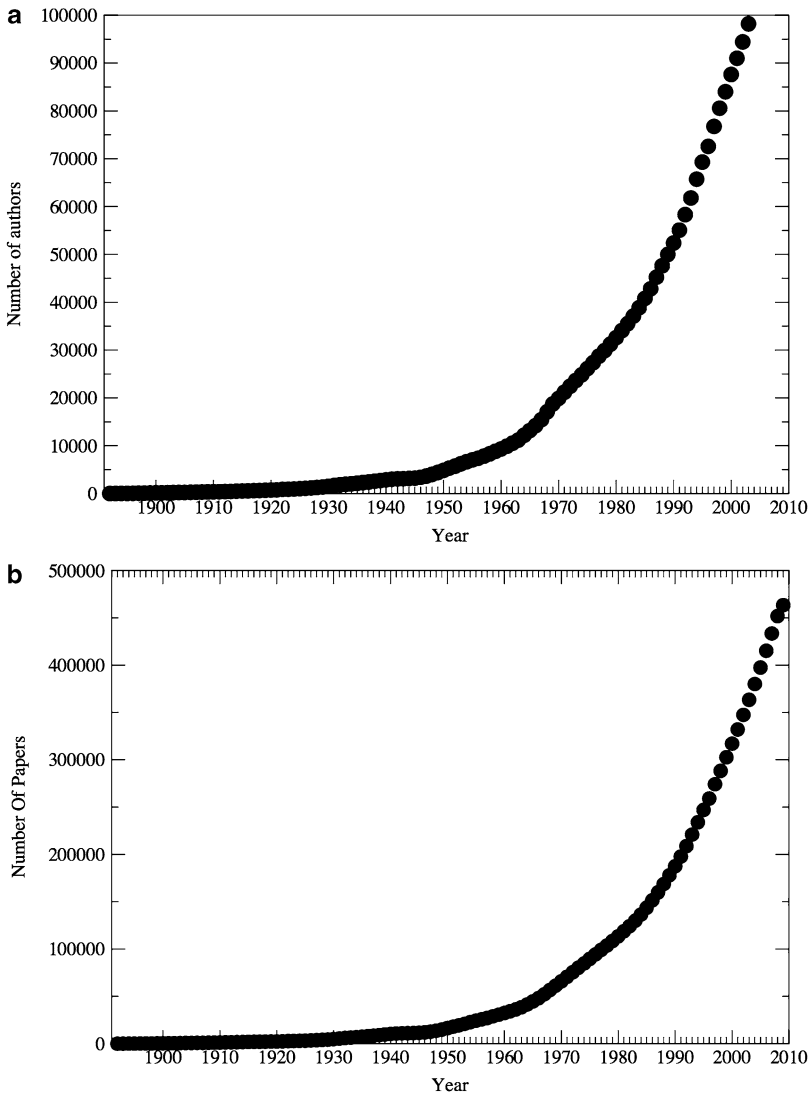
Hence, we derive two time-varying graphs: the *temporal coauthorships network*, with undirected edges and authors as nodes where a link stands for the relations of coauthoring a paper and the *temporal citations network* having papers as nodes and the links (directed) representing the citations from a paper to another one. The temporal dimension of both networks is derived by the paper's submission date. The temporal coauthorship network has edges labeled with the date of submission, while the temporal citations network has the nodes labeled with the publication date of papers citing other papers.

More formally, we can define this system as two networks:

- The **temporal coauthorships network** as a quadruplet $G_a^t : (V, E, \mathbb{T}, \rho)$, where the nodes in $v \in V$ are the authors and links $e \in E$ connect a couple of scientists coauthoring a paper. The temporal domain $\mathbb{T} = [t_a, t_b)$ of the function ρ is the *lifetime* of each node v that in this context is assumed as t_a to be the submission date of the paper and $t_b = \infty$.
- The **temporal citations network** as a quadruplet $G_c^t : (V, E, \mathbb{T}, \rho)$, where the nodes in the set V are the papers and each edge $e \in E$ corresponds to a citation to another paper. As for the coauthorships network, the temporal dimension $\mathbb{T} = [t_a, t_b)$ of the presence function ρ of G_c^t is defined within the submission date of papers and ∞.

Networks Evolution

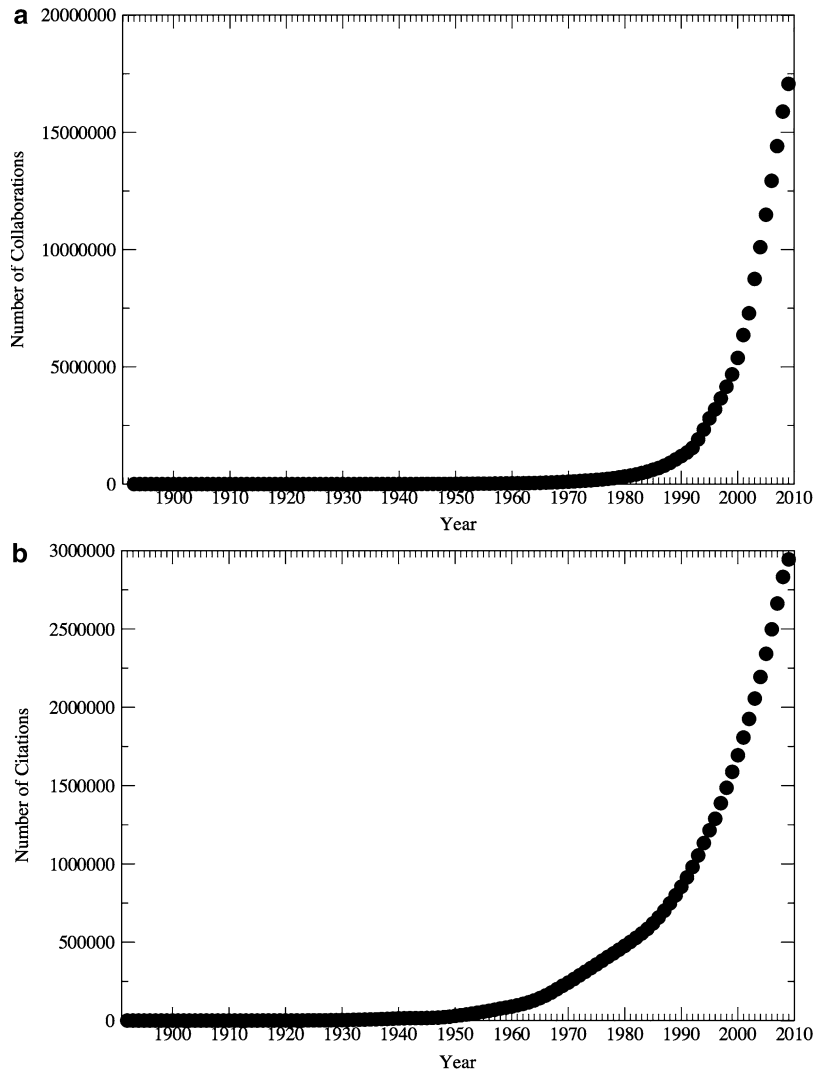
In Fig. 1 we show the number of authors and the number of papers for each year. One can observe from such figures an exponential growth of both the number of authors and of papers along time. Such results are not surprising and have been highlighted by several former works (for instance in Radicchi et al. (2009)). The exponential growth



Stability and Evolution of Scientific Networks, Fig. 1 (**a**) Number of authors and (**b**) number of papers

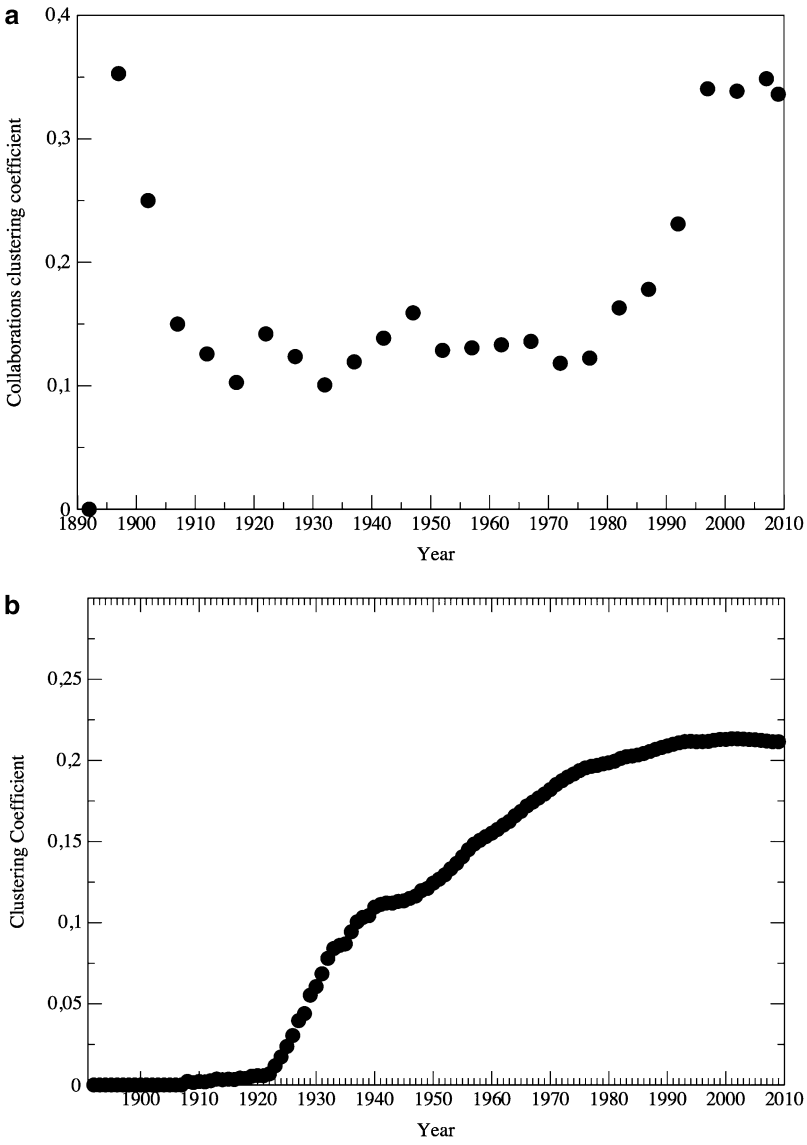
in the number of publications is more or less directly attributed to a change in the behavior of scientists induced by the pressure to publish all along their career (it has been popularized through the proverb publish or perish). The exponential growth of the number of authors is more surprising at a first attempt, as it does not translate an exponential increase of the positions in research that does not exist. It is much more seriously explained by an indirect effect of the exponential growth of publications. We have to remind that this dataset concerns the APS publications, and such publications do not render the effective number of physicists. As the popularity of the APS journals increases, they probably attract more and more physicists worldwide, and we can expect a stabilization of such tendency once as the APS will tend to reference nearly the whole population of physicists worldwide.

In Fig. 2 we show the number of collaborations within authors and the number of citations within papers. Those two measures correspond basically to the number of edges in each of the two networks. The first important element concerns the increase of the number of collaborations that scales as a power law rather than an exponential. This feature results clearly of a double effect over the past few years. The first one is



Stability and Evolution of Scientific Networks, Fig. 2 (**a**) Number of collaborations and (**b**) number of citations

directly linked to the increase in the number of papers that increases the potential of collaboration among authors. The other effect comes from the progressive increase of the number of authors per paper. Translated into network terms, it means that each paper coauthored by N scientist creates N*(N−1)/2 links in the collaboration network. As a consequence, if you follow the current tendency to increase the number of authors, you increase the power coefficient, the number of links among them. Concerning the other figure, the exponential growth is probably less essential but again it results from two combined effects. On the one hand, the number of papers published increases in the same way as the total number of citations. On the one end, the slight tendency to progressively increase the number of papers cited in each paper straightened again the slope. Considering the two graphs mentioned in Fig. 1, the basic feature that we can observe is a global tendency of the increase of the number of nodes in the corresponding networks. The point that the number of links on each graph increases more rapidly than the number of nodes leads to the conclusion that the coauthorship and the citation graphs tend to grow and densify as well. However, we don't have any clue concerning the properties of such a density growth, mainly, is it an egalitarian growth or is it an elitist system with some few nodes benefiting from this



Stability and Evolution of Scientific Networks, Fig. 3 The evolution of the clustering coefficient. (**a**) Collaboration network. (**b**) Citation network

increase in density and the majority of nodes being left behind without many links? The measure of the evolution of the clustering coefficient on such networks can bring arguments for this distinction.

In Fig. 3 we show the clustering coefficient – i.e., the transitivity among nodes – for the collaboration and citation network. Qualitatively, the curves are totally different on the two networks. On the coauthorship network, the evolution follows first an important decrease and then stabilizes before increasing again. On the citation network instead, we can observe an increase that tends to stabilize in the last 20 years. The elements of interpretation behind those two figures are the following. From the coauthorship network, the first global decrease can be explained mainly because it starts from an important number of non-connected components in the network. Therefore, the creation of new links among those components (or communities) that corresponds to a porosity of the different communities in physics results in a global decrease of the clustering coefficient as it tends to dissolve locally the density of each component. Once a global giant component is created (corresponding to the observe plate on

the figure), then there is a stabilization of the clustering coefficient. The final increase is maybe the most interesting feature of this analysis as it corresponds to the case where, in a global single component, the clustering is increasing. This is the case where communities tend to emerge from a global network. Therefore this last increase could be interpreted as the formation and the radicalization of scientific communities on the global network. Such network communities correspond to the effective work in the scientific communities, i.e., coauthorship. Concerning the evolution of the clustering coefficient in the citation network, the first observation we have to make is that the global big component appears very soon on this network (this is much more probable to cite works from outside the field than to collaborate with people from outside the field). Therefore this global and progressive increase of the clustering coefficient corresponds solely to the progressive formation of scientific communities on the network. However, the final stabilization of the index results from a consolidation of the communities that have reached a relative equilibrium. We have to notice that in the case of the creation of new communities or emerging fields, we could see the global clustering coefficient increase again. Such an observation can be made on the figure where around 1940, we can observe a global stabilization of the index and therefore of the corresponding communities before than to increase again, such a new burst being the result of the inclusion of new communities in the network. However, in order to relativize such an effect, we have to remind that the dataset we analyze corresponds to the publications of the APS, and such an inclusion of new communities can result simply from an editorial choice corresponding to the launch of new journals on new thematics for the APS, but not necessarily for the scientific domain of physics.

Conclusions

In this paper we characterize the evolution of a scientific community extracted by the APS dataset. The temporal dimension and the metrics used for the analysis were formalized using time-varying graphs (TVG), a mathematical framework designed to represent the interactions and their evolution in dynamically changing environments.

Since we are interested in the relationships between collaborations and citation behaviors of scientists, we focus on the network of most cited authors and on its structural evolution where several interesting aspects emerge. Through our approach, we capture the role played by famous authors on coauthorship behaviors. They act as attractors on the community. The driving force is a sort of preferential attachment driven by the number of citations received by a given group that in terms of the goal of any scientific community indicates a strategy oriented to the community belonging.

Furthermore, the evolution of the network from a sparse and modular structure to a denser and homogeneous one can be interpreted as a threefold process reflecting the natural selection. The first phase is the exploration of ideas by means of separated works, once some ideas start to be cited (selected) more than others, then authors tend to join groups that have produced highly cited works. The selection is performed by individuals in a goal-oriented environment, and such a (social) selection produces self-organization because it is played by a group of individuals which act, compete, and collaborate in order to advance science. In fact, the driving force is an emergent effect of the interdependencies between citations and the goal of the scientific production since the social selection determines the emergence of a topic and of the scientists working on it by determining the so called preferential attachment toward groups and topics having high potential of citations.

Acknowledgments

Thanks to the American Physical Society for allowing us to use the APS dataset.

Cross-References

▸ Analysis and Visualization of Dynamic Networks
▸ Community Evolution
▸ Dynamic Community Detection
▸ Modeling and Analysis of Spatiotemporal Social Networks

References

Amblard F, Casteigts A, Flocchini P, Quattrociocchi W, Santoro N (2011) On the temporal analysis of scientific network evolution. In: CASoN, Salamanca, pp 169–174

Butts CT, Carley KM (2007) Structural change and homeostasis in organizations: a decision-theoretic approach. The Journal of Mathematical Sociology 31(4): 295–321

Casteigts A, Flocchini P, Quattrociocchi W, Santoro N (2010) Time-varying graphs and dynamic networks. Technical report, University of Carleton, Ottawa

Davidsen J, Ebel H, Bornholdt S (2002) Emergence of a small world from local interactions: modeling acquaintance networks. Phys Rev Lett 88(12):128701

Deffuant G, Neau D, Amblard F, Weisbuch G (2001) Mixing beliefs among interacting agents. Adv Complex Syst 3:87–98

De Solla Price DJ (1965) Networks of scientific papers. Science 149(3683):510–515

Gilbert N (1977) Referencing as persuasion. Soc Stud Sci 7:113–122

Guimera R, Uzzi B, Spiro J, Amaral LA (2005) Team assembly mechanisms determine collaboration network structure and team performance. Science 308(5722):697–702

Holme P (2005) Network reachability of real-world contact sequences. Phys Rev E 71(4):46119

Hummon NP, Dereian P (1989) Connectivity in a citation network: the development of DNA theory. Soc Netw 11(1):39–63

Jeong H, Neda Z, Ravasz E, Schubert A, Barabasi AL, Vicsek T (2002) Evolution of the social network of scientific collaborations. Physica A 311:590–614

Klemm K, Eguíluz VM (2002) Highly clustered scale-free networks. Phys Rev E 65(3):036123+

Kossinets G, Kleinberg J, Watts D (2008) The structure of information pathways in a social communication network. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2008), Las Vegas, pp 435–443

Kostakos V (2009) Temporal graphs. Phys A Stat Mech Appl 388(6):1007–1023

Lelarge M (2008) Diffusion of innovations on random networks: understanding the chasm. In: Papadimitriou CH, Zhang S (eds) Internet and network economics. Springer, Berlin, Vol. 5385, pp 178–185

Leskovec J, Kleinberg J, Faloutsos C (2005a) Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining, Chicago. ACM, pp 177–187

Leskovec J, Kleinberg JM, Faloutsos C (2007) Graph evolution: densification and shrinking diameters. TKDD 1(1)

MacRoberts MH, MacRoberts BR (1996) Problems of citation analysis. Scientometrics 36(3):435–444

Mataric M (1992) Designing emergent behaviors: from local interactions to collective intelligence. In: Proceedings of the international conference on simulation of adaptive behavior: from animals to animats, Honolulu, Hawai, USA. vol 2, pp 432–441

Moore C, Newman MEJ (2000) Epidemics and percolation in small-world networks. Phys Rev E 61:5678–5682

Newman MEJ (2001a) Proceedings of the National Academy of Sciences of the United States of America 98(2):404–409

Newman MEJ (2001b) Clustering and preferential attachment in growing networks. Phys Rev E 64:025102

Newman MEJ (2004a) Coauthorship networks and patterns of scientific collaboration. Proc Natl Acad Sci 101:5200–5205

Newman MEJ (2004b) Who is the best connected scientist? A study of scientific coauthorship networks. In: Ben-Naim E, Frauenfelder H, Toroczkai Z (eds) Complex Networks Lecture Notes in Physics Springer. Berlin/New York. Vol. 650, pp 337–370
Complex networks. Lecture notes in physics.

Peltomaki M, Alava M (2006) Correlations in bipartite collaboration networks. J Stat Mech 2006:P01010

Powell WW, White DR, Koput KW (2005) Network dynamics and field evolution: the growth of interorganizational collaboration in the life sciences. Am J Sociol 110(4):1132–1205

Quattrociocchi W, Paolucci M, Conte R (2009) On the effects of informational cheating on social evaluations: image and reputation through gossip. Int J Knowl Learn 5(5/6):457–471

Quattrociocchi W, Conte R, Lodi E (2010) Simulating opinion dynamics in heterogeneous communication systems. In: ECCS 2010, Lisbon

Quattrociocchi W, Amblard F, Galeota E (2012) Selection in scientific networks. Soc Netw Anal Min 2(3):229–237

Radicchi F, Fortunato S, Markiness B, Vespignani A (2009) Diffusion of scientific credits and the ranking of scientists. Phys Rev E 80:056103

Redner S (2005) Citation statistics from 110 years of physical review. Phys Rev Phys Today 58:49–54

Santoro N, Quattrociocchi W, Flocchini P, Casteigts A, Amblard F (2010) Time varying graphs and social network analysis: temporal indicators and metrics. Technical report, University of Carleton, Ottawa

Santoro N, Quattrociocchi W, Flocchini P, Casteigts A, Amblard F (2011) Time-varying graphs and social network analysis: temporal indicators and metrics. In: 3rd AISB social networks and multiagent systems symposium (SNAMAS), York, Apr 2011, pp 32–38
Tang J, Scellato S, Musolesi M, Mascolo C, Latora V (2009) Small-world behavior in time-varying graphs. Arxiv preprint arXiv:0909.1712
Taramasco C, Cointet J-P, Roth C (2010) Academic team formation as evolving hypergraphs. Scientometrics 85:721–740
Wagner CS, Leydesdorff K (2005) Network structure, self-organization, and the growth of international collaboration in science. Res Policy 34(10):1608–1618
Watts DJ (1999) Networks, dynamics and the small world phenomenon. AJS 105:493–527
Woolley DR (1994) Plato: the emergence of online community. Comput Mediat Commun Mag 1(3):5

Statistical Analysis

▸ Extracting Individual and Group Behavior from Mobility Data

Statistical Inference

▸ Theory of Statistics, Basics, and Fundamentals

Statistical Modeling

▸ Siena: Statistical Modeling of Longitudinal Network Data

Statistical Models

▸ Theory of Statistics, Basics, and Fundamentals

Statistical Relational Learning

▸ Probabilistic Logic and Relational Models
▸ Relational Models

Statistical Relational Models

▸ Relational Models

Statistical Research in Networks – Looking Forward

Eric D. Kolaczyk
Department of Mathematics and Statistics, Boston University, Boston, MA, USA

Synonyms

Propagation of uncertainty; Research challenges

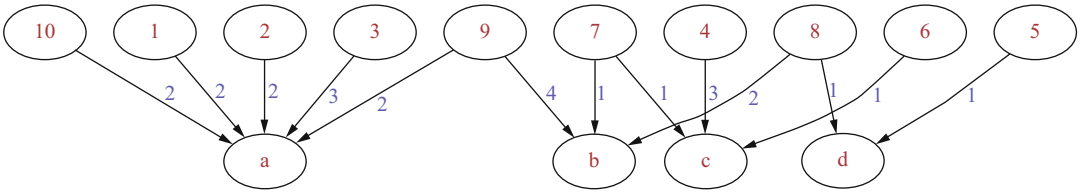
Glossary

Network Summary Statistic A statistic summarizing a network graph
Propagation of Uncertainty Understanding the effect of uncertainty in an initial set of measurements on functions thereof

Introduction

The emerging field of network analysis, through its roots in social network analysis, has had a nontrivial statistical component from the start. In the ensuing years, problems in network analysis have motivated – and continue to motivate – new research in the field of statistics. Conversely, new developments in statistics are routinely integrated into network research. It is therefore rather surprising that, despite the many interesting and important statistical challenges in network analysis to which researchers have already been able to respond, there nevertheless are a number of challenges of an entirely fundamental nature that remain almost untouched!

We will support this central claim through two examples. Additional examples will be mentioned in passing at the end. All of these examples



Statistical Research in Networks – Looking Forward, Fig. 1 A bipartite representation of Internet traffic flow measurements, from ten sources (i.e., *1* through *10*) to four destinations (i.e., *a* through *d*)

relate to the basic problem of *propagation of uncertainty* – understanding the effect of uncertainty in an initial set of measurements on functions thereof.

Network construction is often a sophisticated (and frequently complicated) process. Importantly, the network graph $G = (V, E)$ that is considered to be "observed" typically is the result of taking some sort of basic measurements and then going through a series of steps (whether formal or informal) before arriving at a set of vertices V and edges E. Accordingly, to the extent that there is uncertainty in the basic underlying measurements, there will be uncertainty in the corresponding edges (and perhaps vertices as well, depending on context). In turn, therefore, this uncertainty will impact any further processing of G.

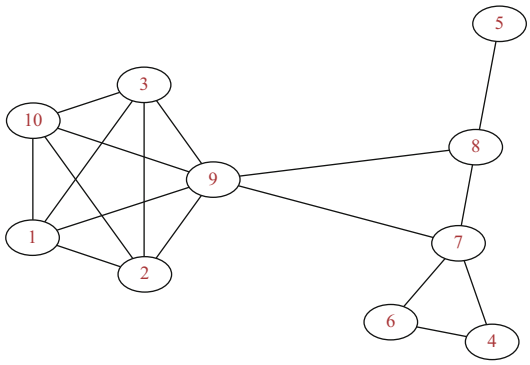
The examples of such "processing" that we will consider are (i) network summary statistics and (ii) network modeling. Before doing so, however, the following motivating illustration from the context of Internet traffic data analysis will be useful.

Illustration: Network Traffic Graphs

Consider the recent trend towards using social network principles and analysis in the study of Internet traffic flow data. Each time an actor, working from an Internet-capable device (e.g., a smartphone), uses an Internet application (e.g., an iPhone or Android app), traffic is generated in the form of packets of information that are exchanged between the device and the relevant Internet destination(s). The collection of such packets relevant to a given basic task (e.g., opening a webpage in a browser or downloading a new song) is termed a *flow*. It is possible, using measurement technology deployed within the physical layer of the Internet (typically at Internet routing devices), to capture information on such flows. Researchers then use bipartite graphs to represent these flows, with vertices corresponding to origins (e.g., the IP address of an actor's iPhone) and destinations (e.g., the IP address of a server hosting a webpage), and edges indicating that a flow was declared to have passed between the two. A small graph of this sort is shown in Fig. 1.

Social network tools are then used to study these typically massive networks, or variations thereof. For example, beginning with the bipartite representation just described, Ding et al. (2012) construct networks $G = (V, E)$ of actors $v \in V$ (formally, IP addresses understood to correspond to users) that have an edge $e = (u, v) \in E$ between them when there is least one web server with which both u and v exchanged flows, indicating some level of common behavior. (Formally, these authors construct a one-mode projection of the bipartite network traffic graph). An example may be found in Fig. 2. Exploiting data on IP addresses known to have exhibited malicious behavior during the same time period, they find that the corresponding vertices, say $v^* \in V$, in the graph G tend to be overrepresented in regions of G falling *between* natural communities (e.g., as bridge nodes between communities). Thus, these nodes demonstrate a curious (anti)social behavior.

Ding et al. (2012) use this observation to create an anomaly detection strategy for finding IP addresses participating in malicious behavior. Similarly, other authors have used networks



Statistical Research in Networks – Looking Forward, Fig. 2 One-mode projection of the bipartite network in Fig. 1

of this type (termed "traffic activity graphs" or "traffic dispersion graphs") for a variety of purposes, ranging from characterization to detection. See, for example, Jiang et al. (2010), Jin et al. (2009), Iliofotou et al. (2009), and Iliofotou et al. (2007).

Notwithstanding such successes, it is important to note that there are various sources of uncertainty in the basic measurements underlying these networks, and hence in the networks themselves. Most fundamentally, Internet routers typically do not keep record of every packet to pass through them. Rather, they record only some fraction of those packets (e.g., 1 in 1,000), through either random or deterministic sampling. Hence, some flows will not be observed at all, corresponding to missing edges in the initial bipartite graph representation. In addition, again due to sampling, a single flow initially may be recorded as multiple flows. As a result, what should be a single edge in the bipartite graph ends up being represented as multiple edges. Post-processing typically is done to ameliorate this latter effect of sampling, but cannot be expected to succeed completely.

The extent to which such low-level sources of uncertainty in the initial underlying flow measurements affect high-level Internet traffic analysis tasks, like characterization and anomaly detection, appears not to have been studied systematically to date. However, there has been promising work understanding and statistically correcting for sampling artifacts at the level of flow summary statistics (e.g., distributions of flow counts and lengths) and queries thereof. See, for example, Duffield et al. (2005a,b) and Cohen et al. (2008). For a general overview of Internet traffic packet sampling and related issues, see Duffield (2004).

Propagation of Uncertainty in Network Analysis

As noted earlier, the issues just described pertain to the problem of propagation of uncertainty – from the initial measurements to the network graph G to any further processing of G. Of course, the statistical analysis of network data – and, in particular, attempting to account for uncertainty inherent to the data – is by no means new. See Kolaczyk (2009), for example, for a recent overview of statistical methods and models in network analysis. However, there remains much to be done and, surprisingly, some of it of a particularly fundamental nature, that is, "fundamental" in the sense that any student of a one-semester elementary statistics course can be expected to have tools for doing analogous tasks with classical data (i.e., independent and identically distributed observations). Yet we lack these same tools in the context of network analysis. We describe two examples in detail below.

Uncertainty in Network Summary Statistics

A standard paradigm in network analysis goes as follows. Data are obtained from a complex system of interest, a network graph G is constructed, and summary statistics of G, say $\eta(G)$, are reported (e.g., degree distribution, clustering coefficient, and various measures of centrality). When G itself is the primary object of interest, this is a sensible paradigm. However, when in reality we have only a "noisy" version of G, say G^*, then the statistic we calculate, i.e., $\eta(G^*)$, is only a noisy version of $\eta(G)$. In that case, interpreting $\eta(G^*)$ as a point estimate of $\eta(G)$, it is natural to wish to equip this estimate with some quantification of its inherent uncertainty, such as

a standard error or, more ambitiously, confidence intervals.

To date there is little in the way of general statistical methodology for this problem. Some progress has been had for (a) certain sources of uncertainty and (b) certain summary statistics $\eta(\cdot)$. Frank and colleagues have, for example, established standard errors for statistics in the form of dyad and triad sums, when G^* is obtained from G through specific sampling plans, such as induced subgraph sampling. See Frank (2004) for an overview of such results.

More generally, Viles (2013) has recently established the limiting distribution of statistics $\eta(\cdot)$ that take the form of sums of configurations of G^*, such as dyads and triads, under a general measurement error model, in the limit as the number of vertices tends to infinity. Intriguingly, rather than the seemingly ubiquitous standard normal distribution, this limiting distribution is a so-called Skellam distribution – the difference of two independent Poisson random variables.

To see intuitively why this distribution might arise, consider the case where $\eta(\cdot)$ simply counts the number of edges in its argument. The error in $\eta(G^*)$ in estimating the true $\eta(G)$ will be (proportional to) the difference of (i) the total number of edges in G^* that are false (i.e., not in G) and (ii) the total number of non-edges in G^* that are false (i.e., in G). These totals are each sums of binary random variables and hence might be expected to possess characteristics of a Poisson distribution, under appropriate conditions. That this is the case, however, is nontrivial to demonstrate, since the binary variables are dependent. The arguments in Viles (2013) make use of Stein's method, typically part of the toolset learned in advanced graduate statistics courses.

These preliminary results suggest that the problem of propagating uncertainty to network summary statistics, although fundamental, has a complexity associated with it that goes beyond that of, say, a simple sample mean as encountered in "Statistics 101". More general results will likely depend on the smoothness of $\eta(\cdot)$, with respect to changes in G^* and the measurement error involved in obtaining G^*, as well as characteristics of the true underlying G itself. Formal statistical arguments establishing such results will require techniques beyond those employed in the classical setting and quite possibly the development of new techniques altogether.

Uncertainty in Network Modeling

While the above paradigm, in which we conceptualize ourselves as having observed a "noisy" version G^* of a "true" graph G, is appropriate for many contexts, another useful perspective is that in which we think of G^* as having derived from some network distribution, $\mathbb{P}(G)$. This perspective underlies, for example, the large body of work in social network analysis using exponential random graph models (ERGMs).

With a history going back roughly 30 years, ERGMs have become a mainstay of social network analysis. See the edited volume by Luscher et al. (2012), for example, for a recent overview. This class of models specifies that the distribution of the adjacency matrix, say Y, for a random graph G, follows an exponential family form, i.e., $p_\theta(Y = y) \propto \exp\left(\theta^T g(y)\right)$, for vectors θ of parameters and $g(\cdot)$ of sufficient statistics. However, despite this seemingly appealing feature, work in the last 5 years has shown that exponential random graph models must be handled with some care, as both their theoretical properties and computational tractability can be rather sensitive to model specification. See Robins et al. (2007), for example, and Chatterjee and Diaconis (2011), for a more theoretical treatment. Benefiting from these findings, software is now available for dependably fitting well-posed ERGM models, and typically estimates of model parameters θ are accompanied by standard errors, where the latter are based on standard arguments for exponential families.

Unfortunately, while such standard errors are perhaps useful in summarizing relative levels of uncertainty associated with estimates of the parameters in θ, there is to date no general theory supporting their use in creating confidence intervals or performing tests. Even more unfortunate is that the importance of this fact does not appear to be universally appreciated, since it is not

unheard of to find applied papers in social network analysis in which ERGM parameter estimates are cited with what are purported to be confidence intervals or results of significance tests! Contrast this situation with that of, say, linear regression analysis, for which a "Statistics 101" student would, by the end of a single semester, typically have been exposed to methods for confidence intervals and tests of all sorts in the classical setting.

In recent work, Kolaczyk and Krivitsky (2013) have demonstrated that the asymptotic analysis necessary to establish a parallel theory for ERGMs likely will be rather more subtle. These authors concentrate on the simplest of ERGM models, in which dyads (y_{ij}, y_{ji}) are independent, and focus on a comparison of the cases of sparse versus non-sparse networks. (We consider a network graph G to be sparse if the number of edges is of the order of the number of vertices, i.e., $N_e = O(N_v)$, rather than the square of that number, i.e., $N_e = O(N_v^2)$.) In that setting they demonstrate that the very order of the asymptotics will depend critically on the sparseness of G. More specifically, they show that the maximum likelihood estimates of the ERGM parameters for attraction and mutuality will converge asymptotically to a bivariate normal distribution in both sparse and non-sparse cases but at rates $N_v^{1/2}$ and N_v, respectively.

At an intuitive level, these results say that the nature of the dependency in the relational measurements y_{ij} leads to variations in the effective sample size. For non-sparse networks, we have effectively $O(N_v^2)$ measurements – in fact, the same number as entries in the adjacency matrix Y. But for sparse networks, we have effectively only N_v measurements! Since the effective sample size drives the relative magnitude of the standard error, as a function of network order, it is a critical factor in establishing asymptotic results justifying confidence intervals and tests based on the latter.

While the results of Kolaczyk and Krivitsky (2013) use tools from, say, the latter part of a first-year course in theoretical statistics (i.e., stochastic convergence of estimating equations, coupled with a double-array central limit theorem), it is likely that additional traction on this problem for ERGM models with more complex forms of dependency (e.g., stars, triadic structure) will require the development of new tools.

Future Directions

In looking forward at the challenges facing us for statistical research in relation to social network analysis, there is a curious feeling of looking back as well, that is, "looking back" in the sense that we realize there is much to be done in this context when we note what has already been done previously in more established contexts. Much that is now considered foundational, in that it is part of a now-standard toolset.

Certainly the classical case of independent and identically distributed observations forms the gold standard. As noted earlier, much of the foundational material in the classical setting – such as confidence intervals for summary statistics and regression model parameters – forms part of the core of what is presented to students at the very earliest stage of statistics education. Social network analysis currently lacks a number of such foundational aspects. Yet it can arguably take hope from the example of time series analysis and spatial data analysis. In both cases, the data deviate from the classical case in that they are dependent. And, moreover, in both cases, over time, the analogous foundations were laid. See, for example, the books Brockwell and Davis (2009) and Cressie (1993) for time series and spatial analysis, respectively.

It can be expected that new tools and techniques will be required from statisticians to fill the gaps in the foundations for network analysis. Networks share dependency with time series and spatial data but lack the temporal and geometric aspects of the latter. (More formally, they lack the properties of Euclidean space that can be exploited in the context of time series and spatial data.) These aspects were critical for successfully extending classical results to time series and spatial data, due to the fact that they facilitate a notion of "local" dependence (e.g., local in time or in space) that emerges naturally as a "loosening" of

the stricter assumption of independent and identically distributed. Developing and working with analogous notions of "localness" in the network context is a key hurdle to be faced.

In terms of specific areas requiring work, from the descriptions above, it should be clear that progress is only just starting to be made on the two examples cited. For confidence intervals for summary statistics, there is much to be explored in understanding the interaction between (a) characteristics of the sources of uncertainty (e.g., based on sampling, measurement, missingness), on the one hand, and (b) the nature of the summary statistics to be computed (e.g., smooth or unsmooth, in an appropriate sense), on the other hand. Furthermore, establishing limiting parametric distributions (such as the Skellam distribution in Viles (2013)) is one key way to facilitate the construction of confidence intervals; it would be useful to see a version of bootstrapping or related resampling approaches justified in the context of networks.

Similarly, asymptotic theory supporting methods for the construction of confidence intervals for network parameters is only beginning to emerge. The most traction appears to have been gained in the context of stochastic block models (e.g., Bickel and Chen 2009; Choi et al. 2010; Celisse et al. 2011; Rohe et al. 2011), although progress is beginning to be had with exponential random graph models as well (e.g., Chatterjee et al. 2011; Chatterjee and Diaconis 2011; Rinaldo et al. 2013). Most of these works present consistency results for maximum likelihood and related estimators, with the exception of Bickel and Chen (2009), which also includes results on asymptotic normality of estimators. See Haberman (1981) for another contribution in this direction, proposed as part of the discussion of the original paper of Holland and Leinhardt (1981). Finally, for some initial (non-asymptotic) results in the context of more mathematical models (e.g., preferential attachment, copying), see Wiuf et al. (2006).

While there are various other similarly important statistical topics that remain to be explored in network analysis, arguably one of the most pressing of those is that of missing data. It is known, again in the classical setting first and foremost, that depending on the mechanism of missingness, the impact of missing data on statistical inference can range anywhere from mild to devastating. See Kolaczyk (2009, Chap. 3) for some general discussion, including comparisons to the importance of missingness and related notions in the context of spatial data analysis. A general framework for thinking about the impact of missingness on network modeling has recently been initiated in Handcock and Gile (2010) and Jiang and Kolaczyk (2012) have recently demonstrated that accounting for observation errors that include missingness (using a hierarchical modeling formulation) can lead to marked improvement in accuracy of link prediction. But, as with the other areas cited above, much remains to be done to explore and develop the necessary statistical infrastructure for understanding and dealing with missingness generally in network contexts.

Cross-References

- Exponential Random Graph Models
- Learning Networks
- Network Representations of Complex Data
- Network Models
- Probabilistic Analysis
- Sampling Effects in Social Network Analysis
- Sources of Network Data
- Theory of Statistics, Basics, and Fundamentals

References

Bickel PJ, Chen A (2009) A nonparametric view of network models and Newman–Girvan and other modularities. Proc Natl Acad Sci 106(50):21068

Brockwell PJ, Davis RA (2009) Time series: theory and methods. Springer, New Dehli

Celisse A, Daudin JJ, Pierre L (2011) Consistency of maximum-likelihood and variational estimators in the stochastic block model. Arxiv preprint arXiv:1105.3288

Chatterjee S, Diaconis P (2011) Estimating and understanding exponential random graph models. Arxiv preprint arXiv:1102.2650

Chatterjee S, Diaconis P, Sly A (2011) Random graphs with a given degree sequence. Ann Appl Probab 21(4):1400–1435

Choi DS, Wolfe PJ, Airoldi EM (2010) Stochastic blockmodels with growing number of classes. Arxiv preprint arXiv:1011.4644

Cohen E, Duffield N, Lund C, Thorup M (2008) Confident estimation for multistage measurement sampling and aggregation. ACM SIGMETRICS Perform Eval Rev 36:109–120

Cressie NAC (1993) Statistics for spatial data (revised edn.). Wiley, New York

Ding Q, Katenka N, Barford P, Kolaczyk ED, Crovella M (2012) Intrusion as (anti)social communication: characterization and detection. In: Proceedings of the 18th ACM SIGKDD conference on knowledge discovery and data mining, Beijing. ACM, pp 886–894

Duffield N (2004) Sampling for passive internet measurement: a review. Stat Sci 19(3):472–498

Duffield N, Lund C, Thorup M (2005a) Estimating flow distributions from sampled flow statistics. IEEE/ACM Trans Netw 13(5):933–946

Duffield N, Lund C, Thorup M (2005b) Optimal combination of sampled network measurements. In: Proceedings of the 5th ACM SIGCOMM conference on internet measurement, Berkeley. USENIX Association, pp 8–8

Frank O (2004) Network sampling and model fitting. In: Carrington PJ, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, New York

Haberman SJ (1981) An exponential family of probability distributions for directed graphs: comment. J Am Stat Assoc 76(373):60–61

Handcock MS, Gile KJ (2010) Modeling social networks from sampled data. Ann Appl Stat 4(1):5–25

Holland PW, Leinhardt S (1981) An exponential family of probability distributions for directed graphs. J Am Stat Assoc 76:33–50

Iliofotou M, Pappu P, Faloutsos M, Mitzenmacher M, Singh S, Varghese G (2007) Network monitoring using traffic dispersion graphs (tdgs). In: Proceedings of the 7th ACM SIGCOMM conference on internet measurement, San Diego. ACM, pp 315–320

Iliofotou M, Faloutsos M, Mitzenmacher M (2009) Exploiting dynamicity in graph-based traffic analysis: techniques and applications. In: Proceedings of the 5th international conference on emerging networking experiments and technologies, Rome. ACM, pp 241–252

Jiang X, Kolaczyk ED (2012) A latent eigenprobit model with link uncertainty for prediction of protein–protein interactions. Stat Biosci 4(1):84–104

Jiang N, Cao J, Jin Y, Li LE, Zhang ZL (2010) Identifying suspicious activities through DNS failure graph analysis. In: 18th IEEE international conference on network protocols (ICNP) 2010, Kyoto. IEEE, pp 144–153

Jin Y, Sharafuddin E, Zhang ZL (2009) Unveiling core network-wide communication patterns through application traffic activity graph decomposition. In: Proceedings of the 11th international joint conference on measurement and modeling of computer systems, Seattle. ACM, pp 49–60

Kolaczyk ED (2009) Statistical analysis of network data: methods and models. Springer, New York/London

Kolaczyk ED, Krivitsky PN (2013) On the question of effective sample size in network modeling. Stat Sci (under invited revision)

Luscher D, Koskinens J, Robins G (2012) Exponential random graph models for social networks: theory, methods, and applications. Cambridge University Press, Cambridge

Rinaldo A, Petrovic S, Fienberg SE (2013) Maximum likelihood estimation in the beta model. Ann Stat (to appear)

Robins G, Snijders T, Wang P, Handcock M, Pattison P (2007) Recent developments in exponential random graph (p*) models for social networks. Soc Netw 29(2):192–215

Rohe K, Chatterjee S, Yu B (2011) Spectral clustering and the high-dimensional stochastic blockmodel. Ann Stat 39(4):1878–1915

Viles WE (2013) Uncertainty propagation from network inference to network characterization. PhD thesis, Department of Mathematics & Statistics, Boston University

Wiuf C, Brameier M, Hagberg O, Stumpf MPH (2006) A likelihood approach to analysis of network data. Proc Natl Acad Sci 103(20):7566–7570

Statistical Simulation

▸ Simulated Datasets

Status

▸ Time- and Event-Driven Modeling of Blogger Influence

Status Update

▸ Microtext Processing

Stochastic Actor-Based Models

▸ Actor-Based Models for Longitudinal Networks

Stochastic Models

▸ Probabilistic Analysis

Storage, Infusion, Detection, Fault Diagnostics, Prevention

▸ Network Management and Governance

Strain Model

▸ Visualization of Large Networks

Strategic Allocation of Resources

▸ Network Management and Governance

Strategic Decision-Making

▸ Top Management Team Networks

Stream Querying and Reasoning on Social Data

Jayanta Mondal and Amol Deshpande
Department of Computer Science, University of Maryland, College Park, MD, USA

Synonyms

Continuous query processing; Dynamic social networks; Incremental computation; Temporal analytics

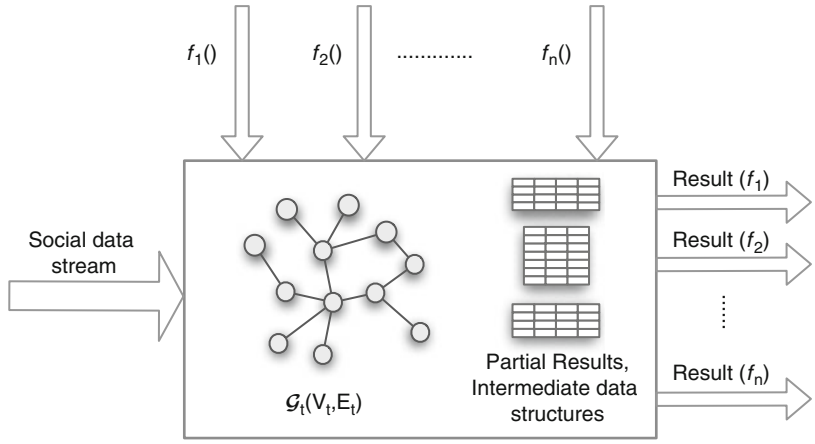
Glossary

Social Data Stream A time-stamped sequence of updates to a social network
SNA Social network analysis
CQP Continuous query processing
CEP Complex event processing

Introduction

Since the inception of online social networks, the amount of social data that is being published on a daily basis has been increasing at an unprecedented rate. Smart, GPS-enabled, always-connected personal devices have taken the data generation to a new level by making it tremendously easy to generate and share social content like *check-in* information, *likes*, *microblogs* (e.g., Twitter), multimedia data, and so on. There is an enormous value in reasoning about such streaming data and deriving meaningful insights from it in real time. Examples of potential applications include advertising, sentiment analysis, detecting natural disasters, social recommendations, personalized trends, spam detection, to name a few. There is thus an increasing need to build scalable systems to support such applications. Complex nature of social networks and their rapid evolution, coupled with the huge volume of streaming social data and the need for real-time processing, raise many computational challenges that have not been addressed in prior work.

Social network data comprises two major components. First, there is a network (*linkage*) component that captures the underlying interconnection structure among the entities in the social network. Second, there is *content data* that is typically associated with the nodes and the edges in the social network. The social network data *stream* contains updates to both these components. The structure of the network may itself change rapidly in many cases, especially when things like webpages and user tags (e.g., Twitter *hashtags*) are treated as nodes of the network. However, most of the social network data stream consists of updates to the data



Stream Querying and Reasoning on Social Data, Fig. 1 High-level overview of a stream querying system

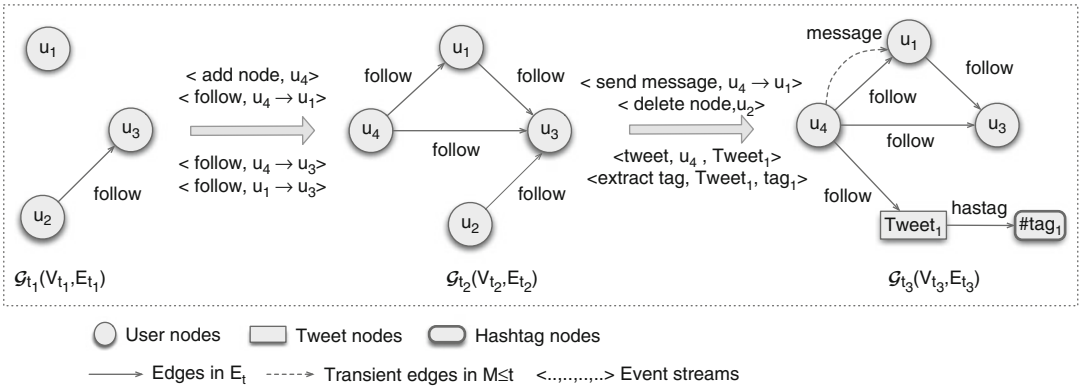
associated with the nodes and the edges, e.g., status updates and other content uploaded by the users, communication among the users, and so on. There is interest in performing a wide variety of queries and analytics over such data streams in real time. The queries can range from simple publish-subscribe queries, where a user is interested in being notified when something happens in his or her friend circle, to complex anomaly detection queries, where the goal is to identify anomalous behavior as early as possible.

In this paper, we present an introduction to this new research area of *stream querying and reasoning* over social data. This area combines aspects from several well-studied research areas, chief among them, social network analysis, graph databases, and data streams. We provide a formal definition of the problem, survey the related prior work, and discuss some of the key research challenges that need to be addressed (and some of the solutions that have been proposed). We note that we use the term *stream reasoning* in this paper to encompass a broad range of tasks including various types of analytics, probabilistic reasoning, statistical inference, and logical reasoning. We contrast our use of this term with the recent work by Valle et al. (2008, 2009) who define this term more specifically to refer to integration of logical reasoning systems with data streams in the context of the Semantic Web. Given the vast amount of work on this and related topics, it is not our intention to be comprehensive in this brief overview. Rather we aim to cover some of the key ideas and representative work.

Problem Definition

An online social network is defined to be a community of people (called *users*) connected via a variety of social relations, that use online technologies to communicate with each other and share information. Social data is defined to be the data arising in the context of a social network that includes both the embedded structural information as well as the data generated by the users. Online social networks continuously generate a huge volume of such social data that includes both structural changes to the network and updates that are associated with the nodes or the edges of the network. The task of "stream querying and reasoning" refers to ingesting and managing such continuously generated data and querying and/or reasoning over it in real time as the data arrives (Fig. 1).

To make the discussion more concrete and formal, let $\mathcal{G}_t(V_t, E_t)$ denote the underlying social graph at time t, with V_t and E_t denoting the sets of nodes and edges at time t, respectively. In general, $\mathcal{G}_t$ is a heterogeneous, multi-relational graph that may contain many different types of nodes and may contain both directed and undirected edges (Fig. 2 shows an example graph). Along with nodes representing the users



Stream Querying and Reasoning on Social Data, Fig. 2 Example of a multi-relational, heterogenous dynamic graph

of the network, V_t may include other types of nodes, e.g., nodes representing communities or groups, user tags, webpages, and so on. Similarly, E_t includes not only symmetric *friendship* (or analogous) edges, but may include asymmetric *follows* edges, *membership* edges, and other types of semipermanent edges that are usually in existence from the time they are formed till the time they are deleted (or till the current time). We distinguish such edges from *transient* edges that can be used to capture specific interaction between two nodes in V_t (e.g., a message being sent from one node to another). A transient edge is typically time-stamped and is only valid for the specific time instance. To allow us to clearly distinguish between these two types of edges, we do not include such transient edges in E_t; instead, we use $M_{\leq t}$ to denote all such transient edges that were generated from the beginning (i.e., from time 0) till time t. This distinction is not necessary, but affords clearer distinctions between different types of stream reasoning tasks in many cases.

The information associated with the nodes and edges can be captured through a set of *key-value* pairs (also called *attribute-value* pairs) associated with them. We once again can make a distinction between semipermanent information associated with the nodes or the edges (e.g., user *names*, *interests*, or *locations*) and transient information associated with them (e.g., *status updates*). The former type of information can be seen as being valid for a given time period, whereas the latter is typically associated with a single time instance. Given this, we define a stream reasoning or querying task to be a declaratively-specified query or an analysis or reasoning task that is posed (submitted) once by the user, but is executed continuously (or periodically with a user-specified frequency) as updates arrive into the system (Fig. 1). Along with a task, denoted $f()$, the user must specify what forms the *input* to the task, when to compute the *output*, and when to *return* the output to the user.

In many cases, the input is the *current* graph, i.e., the input is $\mathcal{G}_t(V_t, E_t)$ (that is continuously changing). An example of such a task is *dense subgraph maintenance* (Angel et al. 2012) where the goal is to compute and maintain the dense subgraphs in a dynamically changing graph. In other cases, the input to $f()$ may be defined using a *sliding window*, i.e., it may be defined as the set of all updates that arrived in recent past. An example of such a task is continuously identifying dense subgraphs in the graph formed by all message edges over say the last 24 h (i.e., the input to the task is $M_{\leq t} - M_{\leq (t-24\ hours)}$). As time progresses, the window slides and new message edges will be added to the graph, and old message edges (that fall out of the window) will be deleted (Table 1).

The second key issue is when to compute the output and when to return it to the user. In some cases, the user may desire continuous execution of the query, i.e., for every relevant change in the input, $f()$ needs to be recomputed (from either scratch or incrementally). Anomaly detection queries typically need to be

Stream Querying and Reasoning on Social Data, Table 1 Notation

Notation	Description
$\mathcal{G}_t(V_t, E_t)$	Current state of the network
$M_{\leq t}$	Transient edges generated till time t
$f_1(), f_2(), \ldots$	Stream querying or reasoning tasks
$\mathcal{N}^k(v)$	k-hop ego network of node v

executed in this fashion since anomalies must be detected as soon as they are formed. But in other cases, the user may specify a frequency with which to execute the query or the task (e.g., every hour or every day). Finally, for simplicity, we will assume that the user should be notified anytime the output of $f()$ is computed and is different from the prior output. However, in many cases, the output may need to be returned to the user only when he asks for it. In those cases, partial pre-computation of the query results (with the rest of the processing performed at query time) becomes a possibility.

Historical Background

Stream querying and reasoning over social networks combines aspects from several different research areas that have been very well studied over the last few decades. Here we will provide very brief background on three of the most closely related research areas: social network analysis, data streams, and graph databases. A more detailed background, including references to related work, can be found in an extended version of this article (Mondal and Deshpande 2013).

Social Network Analysis (SNA) Social network analysis, sometimes called *network science*, has been a very active area of research over the last decade, with much work on network evolution and information diffusion models, community detection, centrality computation, and so on. We refer the reader to well-known surveys and textbooks on that topic (see, e.g., Newman 2003; Scott 2012; Boccaletti et al. 2006). There has been an increasing interest in dynamic or temporal network analysis in recent years, fueled by the increasing availability of large volumes of temporally annotated network data and the real-time requirements of various popular online services. Such analysis has the potential to lend much better insights into various phenomena, especially those relating to the temporal or evolutionary aspects of the network. Many works have focused on designing analytical models that capture how a network evolves, with a primary focus on social networks and the Web. There is also much work on understanding how communities evolve, identifying key individuals, locating hidden groups, identifying changes, and visualizing the temporal evolution in dynamic networks. Most of that prior work, however, focuses on off-line analysis of static datasets.

Data Streams Data stream management is another research area that has seen tremendous amount of work over the last decade (see Aggarwal 2007; Muthukrishnan 2005; Garofalakis et al. 2011 for comprehensive surveys), resulting in several data management systems being built. Several SQL extensions have also been proposed to express continuous queries over data streams. Similarly, languages have also been designed for specifying event patterns to be matched against data streams. Continuous query processing (CQP) also bears strong resemblance to materialized view maintenance, an area that has also seen much work (Gupta and Mumick 1999). The key difference between the two research areas has been that CQP systems are designed to simultaneously support large numbers of relatively simple queries over highly dynamic data, whereas view maintenance techniques usually focus on a small number (usually just one) of more complex queries. The former also tend to build intermediate data structures like *predicate indexes* to efficiently identify the queries whose results are affected by new updates. Another line of work has focused on development of *one-pass* algorithms that can incrementally compute some quantities of interest over very large volumes of data (e.g., statistics or aggregates) while using very small amounts of memory (see, e.g., Muthukrishnan 2005).

Graph Databases Since social networks are naturally represented as graphs, specialized graph data management systems are a natural option to store social network data. There has been much work on single-site graph databases and, in recent years, on distributed graph databases and programing frameworks for specifying batch analysis tasks over graphs. There is also much work on executing specific types of queries efficiently over graphs (both in centralized or distributed settings) through strategic traversal of the underlying graph, e.g., reachability, keyword search queries, subgraph pattern matching, and shortest path queries. However, distributed management of dynamic graph data is not as well studied, especially in the data management research community.

Proposed Solution and Methodology

The area of stream querying and reasoning over social networks is still in its infancy, and as a result, the research in this area is somewhat fragmented with several ongoing attempts at unifying the different research themes. Here we begin with a broad classification of the different types of stream querying and reasoning tasks and give examples of different types of tasks that have been studied in prior literature. We then discuss some of the key research challenges in effective stream querying and reasoning that need to be addressed.

Classifying Tasks by Scope

Here we attempt to classify stream reasoning and querying tasks by their input scope, i.e., what data forms the input to the task at anytime. Broadly speaking, there are two crucial dimensions along which the tasks may differ.

Temporal Scope

The first key dimension captures the temporal scope of the task and has a direct impact on the amount of state that must be stored and reasoned about.

Entire Stream At one extreme, the temporal scope of a stream reasoning task may stretch from the beginning of the stream to the current time. Note that not all the data generated so far may be of interest – e.g., the task may only see a subset of the data by choosing to focus only on certain attributes of the nodes or edges. However, the data of interest may have arrived into the system at any point in the past. For example, in a social network with location data, a stream reasoning task may wish to process all the location updates ever produced by a user for predicting future user movements. We expect such types of stream reasoning tasks to be somewhat uncommon given the large volumes of data generated in most online social networks.

Current State of the Network Many stream reasoning tasks will take the current state of the network (i.e., $G_t(V_t, E_t)$) as the input. An example of this task is online dense subgraph maintenance (Angel et al. 2012) where the goal is to maintain the dense subgraphs of the current social network at all times.

Sliding Window The third alternative that falls in between the two extremes above is that the reasoning task defines a sliding window on the data stream and the input consists of all updates that arrive during that window. For instance, one may be interested in analyzing all messages that were exchanged during the last 24 h among the users of a network to identify anomalous behavior in real time. Another example of such a task is detection of personalized trends where the goal is to find the most commonly seen words or phrases in the recent status updates or blog posts by the friends of a user.

Network Traversal Scope

The second key dimension is what we call *network traversal scope* of a query, which refers to the portion of the network that provides the input to a stream reasoning query or task.

Global Scope Many stream reasoning tasks require reasoning over the entire network. An example of such a task is computation of *PageRank* (or other centrality measures like *betweenness centrality*, *eigenvector centrality*, etc.). Dense

subgraph maintenance task discussed above is also an example of a task with global scope.

Egocentric Scope On the other hand, in many cases, a reasoning task or a query may only focus on a local neighborhood in the network, often termed *ego networks*. For example, if the goal is to identify *social circles* for a user (McAuley and Leskovec 2012), then only a 1- or 2-hop neighborhood around the user may be of interest (Fig. 3). Personalized trend detection task, discussed above, is another example of such a task. Note that, in many cases, we may want to execute the same task for every node in the network (e.g., we may wish to do continuous trend detection for every user of the social network), and in total, updates in the entire network may need to examined. However, those should be treated as separate tasks, each of which is egocentric in scope. The most common example of an ego network is the network over the immediate set of neighbors of a node. However, in general, an ego network of node could be defined as *k-hop neighborhood* containing all nodes reachable within k hops from the node (and all the incident edges among those).

Types of Stream Reasoning Tasks

Next we attempt to provide a categorization of different stream reasoning and querying tasks by type. Given the wide variety in the stream reasoning tasks of interest, unlike the categorization by scope, the categorization that follows is less precise and not fully disjoint. Our intention here is not to be comprehensive, but rather to discuss some representative stream reasoning tasks.

Publish-Subscribe Queries

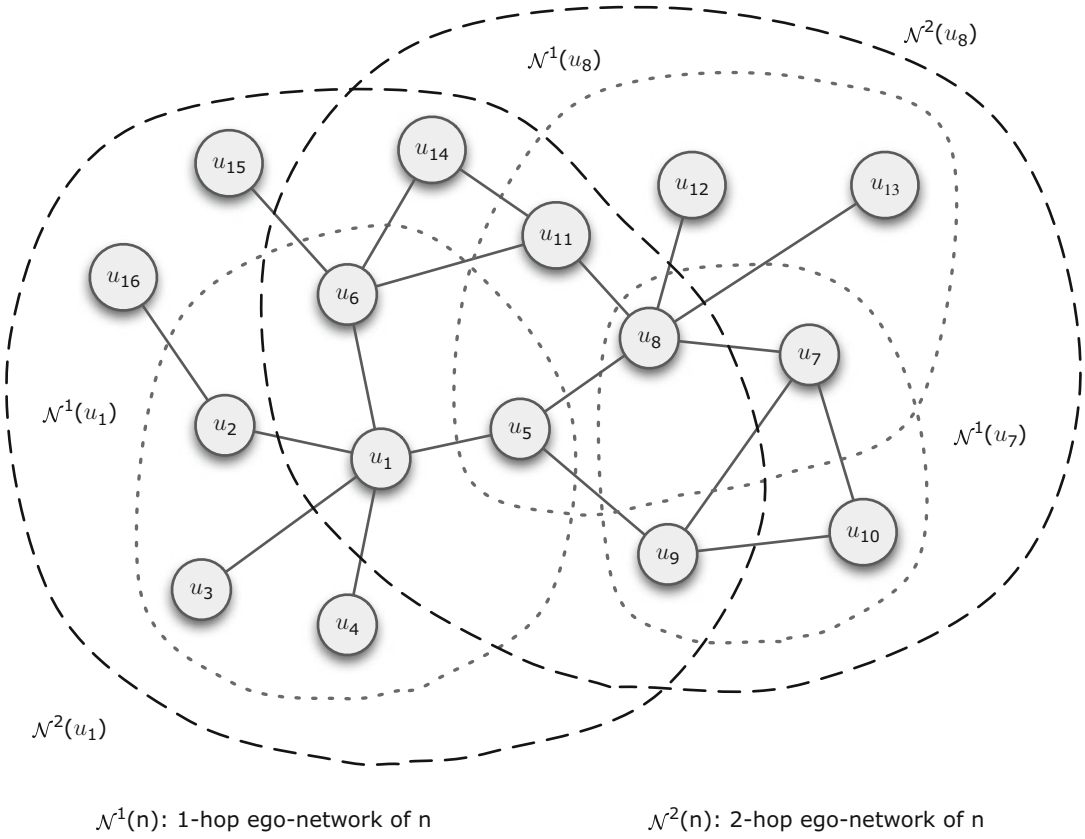
Perhaps the simplest kind of queries over streaming data is what are commonly referred to as *publish-subscribe* queries. These queries form a subclass of the more general class of *event monitoring* queries, where the users specify events or updates of interest and they should be notified as soon as a matching event is detected in the data stream. We make a loose distinction between simple event monitoring queries (what we call publish-subscribe queries) and more complex event monitoring or anomaly detection queries (discussed subsequently). For publish-subscribe queries, the events are typically defined over one or a few data stream updates (i.e., they have very limited temporal and traversal scopes). For example, a user may be interested in tweets that contain a particular key word, or a user may want to know as soon as a friend is online. In a location-enabled social network, a user may be interested in getting notified when one of his friends checks-in in a nearby restaurant or cafe. The key challenge with executing simple event monitoring queries is not so much the complexity of detecting the events, but rather dealing with the very large update rates as well as a very large number of queries.

Complex Event Processing (CEP)

On the other hand, in complex event processing, the events (often called *patterns*) to be detected often have larger temporal or network traversal scopes or both. Hence, unlike simpler publish-subscribe queries, efficiently detecting the events can be a major challenge in CEP. An example of such a query is a continuous subgraph pattern matching query, where the goal is to detect matches to a given query graph in real time. Choudhury et al. (2012) use such queries for continuous detection of accidents from incoming traffic information. CEP systems often support specification of the events using a high-level declarative language. For example, in recent work, Anicic et al. (2011) proposed a language called EP-SPARQL that extends the SPARQL query language with support for specifying complex event processing queries over RDF data streams. Similarly, Mozafari et al. (2012) present a language for detecting hierarchical patterns over hierarchical data (e.g., XML data), that may be generalizable to graph-structured data as well.

Anomaly Detection

Anomaly detection queries can be seen as a form of complex event processing; however, due to their importance, we discuss them separately.



Stream Querying and Reasoning on Social Data, Fig. 3 Stream queries often have ego-centric scope: figure shows 1-hop ego-networks of u_1, u_7 and u_8 and 2-hop ego-networks of u_1 and u_8

The goal with real-time anomaly detection is to identify anomalous behavior in a dynamic network as quickly as possible. Two issues need to be addressed: (1) how to define what constitutes an "anomaly?" and (2) how to efficiently detect anomalies in presence of very high data rates? Generally speaking, anomalous behavior can be defined as behavior that deviates significantly from *normal* behavior. However, in highly dynamic and rapidly changing environments like an online social network, there is often no clear definition of normal behavior, making it a challenge to identify anomalous behavior. There have been many proposals for defining anomalous behavior in social networks over the years. For example, Akoglu et al. (2010) present an approach called Oddball that is based on analyzing the ego-networks of the nodes in the network. Aggarwal et al. (2011) propose a probabilistic algorithm that maintains summary structure models about graph streams to detect outliers. We refer the reader to the tutorial by Akoglu and Faloutsos (2013) for a more comprehensive discussion of different anomaly detection algorithms.

Perhaps because of a lack of a clear definition of an anomaly, there is much less work on efficient techniques for real-time anomaly detection. From the efficiency perspective, an important issue is the scope (both temporal and network traversal) of an anomaly detection task. For example, if the goal is to identify users with anomalous behavior, then the network traversal scope could be limited to ego networks of the users. However, in many cases, detecting anomalous behavior may require global reasoning over the entire network.

Continuous Aggregates/Statistics Computation

In these types of queries, the goal is to incrementally maintain or compute an aggregate or a statistic over the network (Mondal and Deshpande 2013, Supporting ego-centric aggregate queries over large dynamic graphs, unpublished manuscript). An example of such a task is maintaining the top-*k* trending *hashtags* in Twitter, i.e., hashtags with the highest activity over a recent window in past. Another well-studied task is the computation of *global clustering coefficient* in presence of streaming updates to the network structure (Jowhari and Ghodsi 2005; Becchetti et al. 2008). A simpler aggregate query might be to continuously maintain, for all users, their friends that are (physically) closest to them (the aggregate function here is MIN). There are two key properties of aggregate functions that have significant impact on the computational complexity of the computation task: *duplicate sensitivity* and *decomposability*. A duplicate-insensitive aggregate function will return the same value even if some of its inputs are repeated. Examples include MAX, MIN, and UNIQUE. Duplicate-insensitive aggregates are amenable to additional optimizations during computation (Madden et al. 2002a). On the other hand, whether the aggregate function is *holistic* or decomposable has a significant impact on the optimizations that we can perform (Madden et al. 2002a). A holistic aggregate function (e.g., MEDIAN) requires all the input values to compute the final result, whereas decomposable aggregate functions are amenable to optimizations centered around partial aggregate computation and can be computed with much less memory. Clustering coefficient is an example of the latter type of aggregate function since the number of triangles can be counted (mostly) independently for each node.

Maintenance of Views or Other Derived Information

In this type of a task, the goal is to incrementally maintain the result of running an algorithm or performing a computation on the social network in presence of updates. Such tasks can be seen as a generalization of *materialized view maintenance* in traditional relational databases. In traditional view maintenance, the goal is to incrementally maintain the result of a declaratively specified query; however, in social networks, the focus is often on more complex reasoning tasks. Examples of such tasks include incremental maintenance of *PageRank*, *dense subgraphs*, *spanning trees*, *shortest paths*, and *communities*. In general, for any graph algorithm that is of interest in SNA, the question of incremental maintenance of the result in a dynamic setting may need to be addressed. For example, Bahmani et al. (2010) address the problem of incrementally maintaining PageRank over a social network. Several works have considered the problem of incremental maintenance of dense subgraphs (e.g., Angel et al. 2012). The key challenge here is to avoid re-computation from scratch, and so far, most of the proposed techniques are heavily focused on a specific task.

Research Challenges and Future Directions

In this section, we look at some of the key research challenges in supporting stream reasoning and querying tasks over social networks and briefly review the prior work on addressing those challenges. We stress that the area of stream reasoning over social network is still in its infancy, and the solutions discussed here should be considered as the starting point for future research on this topic.

Query Language

One of the major challenges in building general-purpose data management techniques or systems for stream reasoning over social networks is the lack of a high-level declarative query language for specifying the tasks. This issue arises in the context of graph data management in static settings as well. Well-established relational or XML query languages are not appropriate for graph-structured data because they lack support for specifying graph traversals. Although there have been proposals for graph query languages, none has gained wide acceptance; perhaps the

only exception is the SPARQL query language, but the use of that query language has been largely limited to RDF datasets. This lack of a declarative language has led to a significant repetition of work by researchers that are developing tools for stream reasoning and querying over social networks. Clearly it is impossible to specify all of the wide range of tasks that we discussed in the previous section using a high-level, declarative language. However, we believe that it is possible to develop a declarative query language that will serve the needs of many stream reasoning and querying tasks; further, those tasks that cannot be fully expressed in the language can use the language to do part of the computation, with the remaining part done using a program written in a procedural language that ingests the result (analogous to how *user-defined functions (UDFs)* are often used in conjunction with SQL in relational databases).

There are several starting points for designing such a query language. Several languages have been proposed in recent years that build upon SPARQL, e.g., streaming SPARQL (Bolles et al. 2008), continuous SPARQL (C-SPARQL) (Barbieri et al. 2009), and EP-SPARQL (Anicic et al. 2011). Although these languages focus on RDF data streams, they could be adapted to use in social networks by treating social network data as RDF data. Example 1 shows a C-SPARQL query that, given a stream of tweets along with the identified *hashtags* in it, returns all the hashtags with their cumulative frequencies within the last hour. Some of the key extensions to SPARQL include the use of "REGISTER QUERY" keyword to specify a continuous query that should be evaluated continuously and a way to specify a window over the stream (using keyword "RANGE").

Another option is to generalize XPath. For example, Mozafari et al. (2012) propose XSeq, an extension to XPath to express both sequential and Kleene-closure expressions for XML streams. Example 2 shows an XSeq query that reports Twitter users who have been active for over a month. A key challenge here is that X-Path is designed to operate on tree-structured data, not graph-structured data. However, recent theoretical work suggests that it may be possible to use XPath for specifying graph queries (Libkin et al. 2013).

Example 1: C-SPARQL Example (Barbieri et al. 2009). Given the static user information and a stream of tweets, compute the total number of tweets per hashtag in last hour.

```
1: REGISTER QUERY
   NumberOfTweetsPerHashTag COMPUTE
   EVERY 10m AS
2: PREFIX ex: <http://example/>
3: SELECT DISTINCT ?hashtag ?total
4: FROM STREAM <http://twitter.com/alltweets>
   [RANGE 1h STEP 10m]
5: WHERE
6: ?user ex:from ?country .
7: ?user ex:tweets ?tweet .
8: ?tweet ex:has ?hashtag FILTER
   (?country="USA")
9: AGGREGATE { (?total. COUNT(?tweet).
   ?hashtag }
```

Example 2: XSeq Example: In a stream of tweets, report users who have been active over a month. A user is active if he posts at least a tweet every 2 days.

```
1: return first(T)@userid
2: from /twitter/ Z* ($T)*
3: where tag(Z) = 'tweet' and tag(T) = 'tweet'
4: and T@date-prev(T)@date < 2
5: and last(T)@date-first(T)@date > 30
6: partition by /twitter/tweet@userid
```

Finally, the option that we have taken in our work (Moustafa et al. 2011; Mondal and Deshpande 2013 Supporting ego-centric aggregate queries over large dynamic graphs. Unpublished manuscript) is to extend Datalog (Ramakrishnan and Ullman 1995) for this purpose. In recent years, Datalog has been shown to be an effective centerpiece in enabling declarative specification in a range of domains including networking, data cleaning, machine learning, and SNA. Compared to the above two languages, Datalog seems more amenable to be extended to support a large class of complex aggregate queries (e.g., global queries like *PageRank* computation and *shortest paths* can be specified using *recursion*).

Example 3: Datalog Example (Moustafa et al. 2011): Compute the clustering coefficient of each node.

```
1: NeighborCluster(X, COUNT<Y, Z>) :=
2:                                  Edge(X,Y),
   Edge(X,Z), Edge(Y,Z)
3: Degree(X, COUNT<Y>) := Edge(X, Y)
4: ClusteringCoeff(X, C) :=
5:            NeighborCluster(X,N), Degree(X,D),
   C=2*N/D*(D-1)
```

Datalog snippet in Example 3 specifies computation of *local clustering coefficient*, a measure of connectedness of a node's neighborhood. With some extensions, Datalog can also be used to specify social network transformation tasks as we showed in our prior work (Moustafa et al. 2011, 2013). Such flexibility may make a Datalog-based language, a superior option in the end to specify a wide variety of stream reasoning tasks over social data.

Efficient Execution Strategies

Irrespective of how the stream reasoning tasks are specified, we must devise efficient execution strategies that can handle the very high update rates expected in online social networks. Below we briefly survey the key ideas that have been used successfully in past research on data streams for low-latency execution.

Incremental Computation The naive option of re-executing a query or a reasoning task when a new update arrives is likely to be infeasible except for very low-rate data streams. Instead the goal of incremental computation is to maintain sufficient intermediate state in memory so that the new answer can be computed in an incremental fashion with minimal work. Such incremental techniques are unfortunately often specific to the task at hand. Eppstein et al. (1999) did an early survey on the related topic of dynamic graph algorithms. In a recent work, Angel et al. (2012) and Agarwal et al. (2012) devise techniques for maintaining dense subgraphs; Bahmani et al. (2010) present an approach to incremental computation of PageRank; Kutzkov and Pagh (2013) present an incremental algorithm for computing clustering coefficient; and so on. A key research challenge here is to identify incremental techniques that are applicable to a wide variety of tasks (one way to do that is to focus on a high-level query language as we discussed in the previous section, e.g., C-SPARQL (Barbieri et al. 2010)). There is also often a natural trade-off between the amount of intermediate state that is maintained and the amount of work that needs to be done when a new update arrives. Better understanding of this trade-off also presents a rich area for future work.

Sharing Across Multiple Queries Unlike traditional data management systems, in stream query processing systems, we may have thousands to millions of continuous queries running simultaneously. For instance, a personalized trend detection query where the goal is to monitor trends in every user's ego network can be seen as a collection of a large number of independent queries, one for each user. Sharing of computation across these queries is crucial in order to limit the computational cost. Such sharing has been shown to be an effective way to deal with high-rate data streams in past work on data streaming systems (Madden et al. 2002b; Diao et al. 2002). However, these types of techniques have not been well studied in social network setting. In a recent work, we designed novel techniques based on graph compression to exploit such sharing for continuous aggregate computation in social networks (Mondal and Deshpande (2013) Supporting ego-centric aggregate queries over large dynamic graphs. Unpublished manuscript).

Approximate Computation One way to mitigate the execution complexity is to consider computing approximate answers instead of exact answers. This is especially attractive in scenarios where exact computation can be shown to be prohibitively expensive. For example, Becchetti et al. (2008) show how one can incrementally compute local clustering coefficient with small error bounds where the exact algorithm (Alon et al. 1997) can require $O(n^{2.3727})$ time. Although there is much work on this topic in the data streams community, only recently have re-

searchers started investigating similar problems for network algorithms. Zhao et al. (2011) present a graph-sketching technique, called gSketch, and show it can be used to answer several primitive frequency estimation techniques. Similarly, Ahn et al. (2012) present graph-sketching techniques for approximating *cut* values and for approximating the number of matches to a subgraph pattern query.

Sampling Another general technique to deal with the high update rates is use random sampling to reduce the size of the data that needs to be processed. We may sample at two levels in a social network: first, we can try to sample from the network structure itself to reduce the size of the graph that needs to be processed; second, we can sample from the updates to the content. The latter is generally well understood, and the theory developed in the data streams literature could be extended for some types of queries. However, sampling the network structure is tricky since a naive random sample is likely to yield a network with very different properties than the original network. We refer the reader to Ahmed et al. (2012) for a detailed discussion of network sampling, both in static and streaming settings.

Parallel Computation The increasing scale of most online social networks necessitates use of parallel and distributed solutions. Unfortunately computations on social networks are not easily distributable because of their highly interconnected nature. In fact, partitioning a social graph, which is key to distributed graph processing, is a hard problem to tackle because of overlapping community structure and existence of highly connected dense components (cores) in most social networks. One of the simplest examples of a stream query on social data is a publish-subscribe query that asks to *fetch all updates from all friends* (this is also called *feed following*). Answering such queries with very low latencies is challenging if the data is distributed across a set of machines – for most users, their friends' data is likely to be located across multiple machines necessitating expensive distributed traversals. One extreme option is to replicate the data sufficiently so that, for each user, the required data (i.e., status updates of all their friends) is located on some machine (Pujol et al. 2010). However, both the memory overhead and the replica maintenance overhead can be very high for that solution (Mondal and Deshpande 2012). More intelligent and sophisticated techniques for partitioning and replica maintenance must be developed to address these issues for more general stream reasoning and querying tasks. Another key challenge is designing appropriate distributed programing frameworks to support specifying general-purpose stream querying and reasoning tasks. Although there has been some progress on addressing this challenge in recent years (e.g., Kineograph (Cheng et al. 2012), GraphInc (Cai et al. 2012)), much more needs to be done to scalably support a variety of complex stream querying and reasoning tasks.

Conclusions

Stream querying and reasoning over social data is an emerging research area that combines aspects from social network analysis, graph databases, and data streams and is motivated by an increasing need for real-time processing of continuously generated social data. In this paper we presented a brief overview of this field and discussed some of the key research challenges therein. There has been much work on specific problems in this field over the last few years (e.g., detecting specific types of events or anomalies, incremental maintenance of derived structures like dense subgraphs, approximating different types of summary statistics). However, designing general-purpose data management systems that enable declarative specification of stream querying and reasoning tasks and that can efficiently execute such tasks over high-rate data streams remains a fruitful direction for future research.

Cross-References

▸ Analysis and Visualization of Dynamic Networks
▸ Modeling and Analysis of Spatiotemporal Social Networks

- ▸ Querying Volatile and Dynamic Networks
- ▸ SPARQL
- ▸ Temporal Networks

References

Agarwal MK, Ramamritham K, Bhide M (2012) Real time discovery of dense clusters in highly dynamic graphs: identifying real world events in highly dynamic environments. PVLDB 5(10):980–991

Aggarwal C (ed) (2007) Data streams: models and algorithms. Springer, New York

Aggarwal C, Zhao Y, Yu P (2011) Outlier detection in graph streams. In: 27th international conference on data engineering (ICDE), Hannover, pp 399–409

Ahmed NK, Neville J, Kompella RR (2012) Network sampling: from static to streaming graphs. CoRR abs/1211.3412

Ahn KJ, Guha S, McGregor A (2012) Graph sketches: sparsification, spanners, and subgraphs. In: PODS, Scottsdale

Akoglu L, Faloutsos C (2013) Anomaly, event, and fraud detection in large network datasets. In: WSDM, Rome

Akoglu L, McGlohon M, Faloutsos C (2010) Oddball: spotting anomalies in weighted graphs. In: Proceedings of the 14th Pacific-Asia conference on advances in knowledge discovery and data mining (PAKDD), Hyderabad, pp 410–421

Alon N, Yuster R, Zwick U (1997) Finding and counting given length cycles. Algorithmica 17:209–223

Angel A, Sarkas N, Koudas N, Srivastava D (2012) Dense subgraph maintenance under streaming edge weight updates for real-time story identification. VLDB 5:574–585

Anicic D, Fodor P, Rudolph S, Stojanovic N (2011) EP-SPARQL: a unified language for event processing and stream reasoning. In: WWW, Hyderabad

Bahmani B, Chowdhury A, Goel A (2010) Fast incremental and personalized pagerank. Proc VLDB Endow 4:173–184

Barbieri DF, Braga D, Ceri S, Grossniklaus M (2010) An execution environment for C-SPARQL queries. In: Proceedings of the 13th international conference on extending database technology, EDBT '10, Lausanne, pp 441–452

Barbieri DF, Braga D, Ceri S, Della Valle E, Grossniklaus M (2009) C-SPARQL: SPARQL for continuous querying. In: WWW, Madrid

Becchetti L, Boldi P, Castillo C, Gionis A (2008) Efficient semi-streaming algorithms for local triangle counting in massive graphs. In: KDD, Las Vegas

Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U (2006) Complex networks: structure and dynamics. Phys Rep 424(4):175–308

Bolles A, Grawunder M, Jacobi J (2008) Streaming SPARQL: extending SPARQL to process data streams. In: The semantic web: research and applications, Springer, New York, pp 448–462

Cai Z, Logothetis D, Siganos G (2012) Facilitating real-time graph mining. In: Proceedings of the fourth international workshop on cloud data management, CloudDB'12, Sheraton, Maui, pp 1–8

Cheng R, Hong J, Kyrola A, Miao Y, Weng X, Wu M, Yang F, Zhou L, Zhao F, Chen E (2012) Kineograph: taking the pulse of a fast-changing and connected world. In: Proceedings of the 7th ACM European conference on computer systems, EuroSys '12, Bern, pp 85–98

Choudhury S, Holder LB, Ray A, Chin G Jr, Feo J (2012) Continuous queries for multi-relational graphs. CoRR abs/1209.2178

Diao Y, Fischer P, Franklin MJ, To R (2002) Yfilter: efficient and scalable filtering of XML documents. In: Proceedings of the 18th international conference on data engineering, San Jose. IEEE, pp 341–342

Eppstein D, Galil Z, Italiano GF (1999) Dynamic graph algorithms. In: Atallah MJ (ed) Algorithms and theory of computation handbook, chapter 8. CRC, Boca Raton

Garofalakis M, Gehrke J, Rastogi R (eds) (2011) Data-Stream management – processing high-speed data streams. Data-Centric systems and applications series. Springer, New York

Gupta A, Mumick IS (1999) Materialized views: techniques, implementations, and applications. MIT, Cambridge

Jowhari H, Ghodsi M (2005) New streaming algorithms for counting triangles in graphs. In: Wang L (ed) Computing and combinatorics. Lecture notes in computer science, vol 3595. Springer, Berlin/Heidelberg, pp 710–716

Kutzkov K, Pagh R (2013) On the streaming complexity of computing local clustering coefficients. In: WSDM, Rome

Libkin L, Martens W, Vrgoc D (2013) Querying graph databases with XPath. In: ICDT, Genoa

Madden S, Franklin MJ, Hellerstein JM, Hong W (2002a) TAG: a tiny aggregation service for Ad-Hoc sensor networks. In: OSDI, Boston

Madden S, Shah MA, Hellerstein JM, Raman V (2002b) Continuously adaptive continuous queries over streams. In: SIGMOD, Madison

McAuley JJ, Leskovec J (2012) Discovering social circles in ego networks. CoRR abs/1210.8182

Mondal J, Deshpande A (2012) Managing large dynamic graphs efficiently. In: SIGMOD, Scottsdale

Mondal J, Deshpande A (2013) Stream querying and reasoning on social data. http://www.cs.umd.edu/~jayanta/papers/SRQ-ESNAM.pdf

Moustafa WE, Miao H, Deshpande A, Getoor L (2013) GrDB: a system for declarative and interactive anal-

ysis of noisy information networks: demo, SIGMOD, New York
Moustafa WE, Namata G, Deshpande A, Getoor L (2011) Declarative Analysis of noisy information networks. In: ICDE GDM workshop, Hannover
Mozafari B, Zeng K, Zaniolo C (2012) High-performance complex event processing over xml streams. In: SIGMOD, Scottsdale
Muthukrishnan S (2005) Data streams: algorithms and applications. Now Publishers, Boston/Hanover
Newman MEJ (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256
Pujol J, Erramilli V, Siganos G, Yang X, Laoutaris N, Chhabra P, Rodriguez P (2010) The little engine (s) that could: scaling online social networks. In: SIGCOMM, New Delhi
Ramakrishnan R, Ullman JD (1995) A survey of deductive database systems. J Log Program 23(2):125–149
Scott J (2012) Social network analysis. Sage, London
Valle ED, Ceri S, Barbieri DF, Braga D, Campi A (2008) A first step towards stream reasoning. In: FIS, Vienna, pp 72–81
Valle ED, Ceri S, van Harmelen F, Fensel D (2009) It's a streaming world! Reasoning upon rapidly changing information. IEEE Intell Syst 24(6):83–89
Zhao P, Aggarwal CC, Wang M (2011) gSketch: on query estimation in graph streams. VLDB 5:193–204

Stress Model

▸ Visualization of Large Networks

Structural and Locational Properties

▸ Path-Based and Whole-Network Measures

Structural Attribute

▸ Collective Classification, Structural Features

Structural Autonomy

▸ Structural Holes

Structural Holes

Alona Labun[1] and Rafael Wittek[2]
[1]Jeugdhulp Friesland, Leeuwarden, The Netherlands
[2]Theoretical Sociology – Department of Sociology, University of Groningen, Groningen, The Netherlands

Synonyms

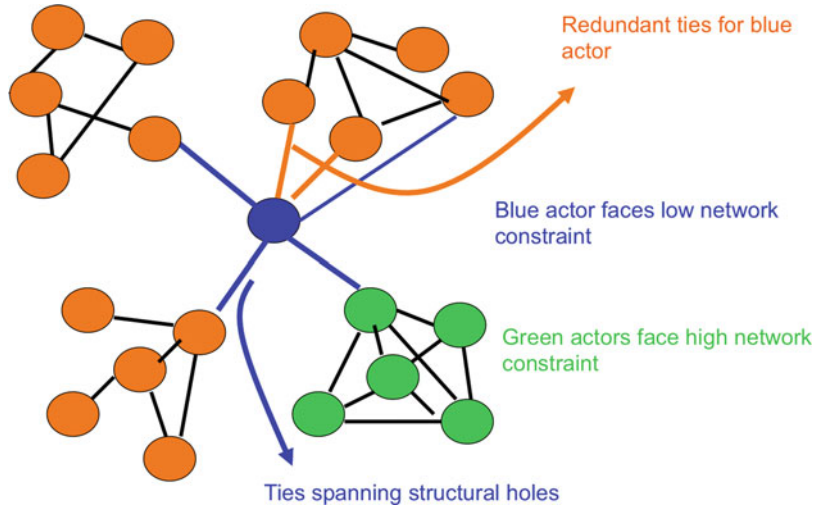
Brokerage; Middlemen; Network entrepreneurs; Social capital; Structural autonomy

Glossary

Secondary Hole Gaps in the networks of a focal actor's primary contacts
Dyadic Constraint Degree to which a focal actor's primary contact can constrain exchange opportunities with third parties
Aggregate Constraint The sum of dyadic constraints imposed on a focal actor by all his contacts
Redundant Tie A tie to a cluster of contacts to which a focal actor already has ties with other actors
Effective Size The number of non-redundant contacts in a focal actor's personal network

Definition

A structural hole refers to an "empty space" between contacts in a person's network. It means that these contacts do not interact closely (though they may be aware of one another). Actors on either side of the structural hole have access to different flows of information (see Fig. 1). Structural holes therefore reflect "an opportunity to *broker* the flow of *information* between people, and *control* the projects that bring together people from opposite sides of the hole" (Burt 2000).



Structural Holes, Fig. 1

Several measures are used to capture structural-hole networks.

Effective network size is an elementary building block in all structural-hole measures. It is composed of three elements: First, the *proportion* of an actor i's time and energy invested in a relation with q:

$$p_{iq}=(z_{iq}+z_{qi})/\left[\sum_{j}(z_{ij}+z_{ji})\right], \quad (1)$$

z_{iq}, z_{qi},z_{ij}, z_{ji} represent time or energy actor i invests in q, i in j, and j in i, respectively.

Second, the *marginal strength* of j's relation with q:

$$m_{jq}=(z_{jq}+z_{qj})/\max(z_{jk}+z_{kj}) \quad j\neq k. \quad (2)$$

m_{jq} is the marginal strength of contact j's relation with actor q. Z_{jq} is the network variable measuring the strength of the relation from j to q and $\max(z_{jq})$ is the largest of j's relations with anyone (Burt 1992:51).

Third, the *redundant portion* (RP) of i's network. The portion of i's relation with j that is redundant to i's relations with other primary contacts is defined as the following:

$$RP=\sum_{q}p_{iq}m_{jq}. \quad (3)$$

Effective size (ES) is obtained by aggregating across all of i's primary contacts j:

$$ES=\sum_{i}\left[1-\sum_{q}p_{iq}m_{jq}\right]. \quad (4)$$

The effective size of i's network ranges from 1 (network provides one single contact) to N (all contacts are non-redundant), with N being the number of all contacts in i's network. The *efficiency* of an actor's network is computed as the effective size divided by the number of actors in the network.

Dyadic constraint C_{ij} measures the degree to which an actor j imposes structural constraint on the focal actor i. Dyadic constraint is highest in a situation where the focal actor's network is inefficient (i.e., he or she invests time and energy in the relation to someone whose network lacks structural holes and is also tied to other contacts in the focal person's network). A low dyadic constraint originates from actors who do not have many ties to a focal person's contacts. Dyadic constraint is a function of effective size:

$$C_{ij} = \left(p_{ij} + \sum_{q} p_{iq} p_{qi} \right)^2 . \qquad (5)$$

C_{ij} = level of constraint that contact j poses on focal actor i; p_{ij}, p_{iq}, p_{qj} see Eq. (1).

Aggregate constraint indicates the extent to which an actor is constrained by the structure of the network involving other members of his or her group. High constraint values indicate low autonomy: the actor has few structural holes, i.e., little entrepreneurial opportunities. Technically, aggregate constraint is the sum of all contact-specific dyadic constraints in an actor's network. This indicator is also the most frequently used one in structural-hole research.

Hierarchy (H) indicates the extent to which aggregate constraint on ego is concentrated in a single alter. If the total constraint on the person is concentrated in a single other actor, the hierarchy measure will have a higher value. If the constraint results more equally from multiple actors in a person's network, hierarchy will be less. The hierarchy measure, in itself, does not assess the degree of constraint. Independently of the constraint on a focal actor, it measures inequality in the distribution of constraints on a focal person across the other actors in its neighborhood.

$$H = \left(\frac{C_{ij}}{C/N} \right). \qquad (6)$$

C_{ij} = level of constraint that j poses on i; C = sum of constraint (from an actor's network) across all N relationships of an actor; N = number of contacts in the actor's network; C/N = mean level of constraint per contact; and the ratio is 1 for contact j posing an average level of constraint.

Oligopoly Primary structural holes were defined as the aggregate of all dyadic constraint on a focal actor. Contact j's constraint on a focal actor i was defined as the product of two terms (Burt 1992:62): (1) the network time and energy i invested to reach j multiplied by (2) the lack of structural holes around j. The second term, in turn, is the product of two conditions: (a) the lack of primary structural holes between the contact j and others in the player's network and (b) the lack of secondary structural holes between the contact and others outside the network who could replace the contact. Burt refers to this second term as the *oligopoly*: "a measure of the organization of players within the cluster around contact j such that it would be difficult to replace j, or threaten him with being replaced, by some other player in the cluster" (Burt 1992:62).

Hole signatures of a focal actor's network describe "the distribution of opportunity and constraint across the individual relationships in a player's network" (Burt 1992:62). Hole signatures can be graphically represented, with the time and energy devoted by a focal actor i to a specific alter j (p_{ij}) delimiting the upper boundary and the dyadic constraint (c_{ij}) defining the lower boundary. Hole signatures allow to assess structural features of a focal actor's network (clique, center-periphery, leader hierarchy, and leaderless hierarchy).

Hole depth The depth of a structural hole reflects "the ease with which it can be developed for control and information benefits" (Burt 1992:42–44). The depth of a hole between two actors is a function of both the degree of cohesion between two players and the degree of structural equivalence of their ties to others: in the ideal-typical structural hole, both actors are neither connected nor do they have equivalent relations to others. A deep structural hole characterizes two unrelated actors with equivalent ties to third parties: they are "competitors in the same market." In a shallow structural hole, two actors have a tie, but do not share equivalent relations to third parties.

Historical Background

Structural-hole theory and the related measures can be seen as the confluence of three streams of work. First, during the late 1960s and early 1970s, Harrison White and his group (now often referred to as the Harvard School),

formalized ideas focusing on the absence of ties between individuals ("gaps"). This resulted in the development of *blockmodeling* algorithms, which grouped structurally equivalent nodes into blocks, and identified "zero blocks" – nodes that did not share similar relations with third parties. These "zero blocks" have qualities similar to structural holes.

Second, the article "*The Strength of Weak Ties*" by one of White's graduate students (Granovetter 1973) produced the counterintuitive empirical finding that in some situations – like job search, the topic of Granovetter's study – individuals benefit more from weak ties (like acquaintances) rather than strong ties (like friends or relatives), because one's weak ties can provide access to circles of information we usually are not familiar with. The "strength" of an interpersonal tie is a linear combination of the amount of time, the emotional intensity, the intimacy (or mutual confiding), and the reciprocal services which characterize each tie. Strong ties represent closer friendship and greater frequency of interaction, whereas weak ties correspond to acquaintances (Granovetter 1973). Members of closely knit groups connected through strong ties tend to be exposed to similar sources of information. Truly novel, valuable information is often likely to come from more distant acquaintances who may serve as a conduit to hard-to-reach parts of the network. A key proposition in Granovetter's argument is that "all bridges are weak ties," which rules out that strong ties can be bridges (also known as the "forbidden triad" assumption). However, Burt (1992:27) argues that the main source of benefits in a network is not the weakness of the tie, but the hole it spans. From this perspective, the focus on the weakness or strength of a tie even obscures the importance of control benefits. "Bridge strength is an aside in the structural hole argument, since information benefits are expected to travel over all bridges. Benefits vary between redundant and non-redundant ties" (Burt 1992:30).

But Granovetter's article by now is among the most frequently cited papers in the social sciences. In addition to stimulating much substantive research, e.g., on job search, it also sparked the interest for social network indicators reflecting an individual's *centrality* in the network (Freeman 1979). *Degree* centrality captures communication *activity* and has been defined by the number of ties an actor has with others in the network or the number of others who choose a focal actor. *Betweenness* centrality reflects the potential for *control* of communication and has been defined as the extent to which an actor has control over other actors' access to various regions of the network. *Closeness* centrality captures either *independence* or *efficiency* and has been conceptualized as an actor's ability to access independently all other members of the network. *Eigenvector centrality* (Bonacich 1987:1172) measures centrality as the summed connection to others, weighted by their centralities. This measure allows to distinguish situations in which being connected to others with many contacts (powerful others) is advantageous for a focal actor (as is the case in communication networks), from situations in which being connected to powerful others is a liability (as is the case in bargaining situations). These centrality measures only partly capture the essence of structural holes, mainly because they are less sensitive to the gaps in the networks of a focal actor's primary contacts.

Third, Burt was among the first who did a serious effort to ground structuralist reasoning on a behavioral micro-foundation. Many of the ideas presented in his 1992 book on structural holes – including the core argument on structural autonomy – had actually been elaborated in detail about a decade earlier in his *Toward a Structural Theory of Action. Network Models of Social Structure, Perception and Action* (Burt 1982). Here, he exposes the rational choice framework underlying structural-hole theory. A key assumption is that individuals are purposeful actors, who strive for improving their well-being by evaluating the costs and benefits of different action alternatives, taking into consideration structural constraints. Individuals in similar network positions face similar constraints. As a result, the network is simultaneously an indicator

of entrepreneurial opportunity and of motivation (Burt 1992:35).

By combining an innovative structural approach with a theory of action, Burt's structural-hole framework significantly advanced previous network research, which clearly lacked a behavioral micro-foundation.

Structural-Hole Theory

In social networks, access to advantageous structural positions is not equally distributed across all actors: some group members may be positioned at the interface between multiple groups with access to boundary-spanning links, while others are positioned in the middle of a single tightly knit group. Structural holes offer two main benefits.

Information benefits come in three forms: access, timing, and referrals. A network rich in structural holes provides one with *access* to non-redundant sources of information originating in multiple, noninteracting parts of the network. It also increases the likelihood of receiving information earlier than individuals in less advantageous network positions (*timing*) and that others talk positively about the focal actor in their own networks (*referrals*).

Control benefits of structural holes result from the opportunity to either play two unrelated parties out against each other (*tertius gaudens*) or to bring them together (*tertius iungens*). In both cases, the third party can reap benefits.

Structural-hole theory further assumes actors to strategically and proactively creating and manufacturing their social network. This means that actors will actively develop the information and control benefits of existing structural holes and manage the constraint of absent structural holes (Burt 1992:230). They have three strategies to achieve this: they can *withdraw* from a contact, they can *expand* their network by adding a contact's competitor to their network, or they can "leave the constraint-generating network in place but to manage the offending constraint by *embedding* it in a second relationship over which you have more control" (Burt 1992:233).

Key Applications

Structural-hole theory has stimulated considerable empirical research on networks, mostly in and between organizations, as well as on entrepreneurship. It was used to explain a wide range of outcomes at the level of individuals and organizations.

Performance With information being a critical resource in organizational settings (McCall 1979; Mechanic 1962; Pettigrew 1972; Pfeffer 1981), individuals rich in structural holes have a better opportunity to manipulate information for their purpose. According to a meta-analysis (Balkundi et al. 2009), and a recent review (Brass 2011), spanning structural holes increased performance or innovation for the focal actor (Ahuja 2000; Burt 1992, 2004; Mehra et al. 2001; Seibert et al. 2001). Disconnected networks help brokers realize value by offering them the opportunity to transfer ideas from one isolated group to another, a process that involves recognizing when solutions current in one part of the network are likely to have applications elsewhere in the network (Hargadon and Sutton 1997).

Promotions Knowing whom to consult for information and aid becomes of crucial importance at times of competition for career opportunities within organizations. In his work "*Structural Holes*" (1992), Burt has systematically explored the network effects on career advancement within the firm. According to his analysis, a configuration of network ties that creates opportunities for brokering and entrepreneurialism (i.e., a network full of structural holes) enhances career opportunities for actors competing for promotions within organizations (Burt 1992, 2005). The findings of another study on social networks and mobility at the workplace further substantiated Burt's claims that the network structures most conducive to maximizing access to information, resources, and brokerage opportunities (i.e., large, sparse networks) are a meaningful determinant of intraorganizational advancement (Podolny and Baron 1997).

Creativity A network "rich in structural holes" has also been found to facilitate the development of novel valuable ideas by increasing the actor's

ability to merge the distinct sources of information in new ways, thus boosting individual creativity. The empirical findings suggest that between-group brokers are more likely to have a vision advantage, express ideas evaluated as valuable, and are less likely to have ideas dismissed (Burt 2004). Moreover, brokerage appears to provide the opportunity for social "gatekeeping" – regulating the access of others to the tightly knit group one belongs to, while at the same time controlling the ways in which one's own group members learn about information coming from other groups (Burt 2004).

Power Occupying a strong or weak structural position in the network has recently been found to affect the inferences organizational actors draw about one another (Labun 2012). In particular, the empirical evidence suggests that the more an individual is constrained by the structure of his network, the more likely he is to attribute power to others. Embeddedness in networks "poor in structural holes" implies a condition of dependence and limited autonomy (Burt 1992), potentially triggering feelings of helplessness and apprehension, and thereby contributing to increased number of power attributions to other group members (Labun 2012).

Trust and gossip Trustworthy and confidential collegial environment may be advantageous when establishing informal cooperation and forming alliances against powerful third parties. According to Burt's study on trust and gossip in social networks (2001), gossip can act as a strategic tool in this process, allowing the group members to control their fellow members' actions and to weaken the reputation of competitors. The manipulation of information flow to one's own advantage becomes easier when employees occupy brokerage positions in the organizational network – connect to colleagues who are not connected with one another. The more trust exists in an employee network, the further negative gossip echoes, so that single incidents of negative gossip can have far-reaching impacts (Burt 2001). Thus, people may ensure norms of cooperation and punish the uncooperative actors (i.e., the untrustworthy group members) through gossiping – by spreading reputation-harming information about them in the broader informal network (Burt 2005).

The gender contingency effects The synthesis of the informal social network theories with research related to career advancement of women has generated interesting insights. Burt (1998) argued that women often lack sufficient legitimacy in their organizations and therefore need to "borrow" social capital (i.e., structural holes) from a strategic partner (sponsor) in order to get promoted. Whereas senior male managers indeed benefit more from a personal network rich in structural holes, women (as well as junior and non-White managers) fare better with a hierarchical network, in which a tie to an influential "sponsor" provides access to this person's entrepreneurial network (Burt 1998).

The hierarchy contingency effects Actor's position in an organizational hierarchy may serve as one of the conditions under which either structural-hole networks or cohesive networks are likely to provide the focal actor with advantages. Burt (1997) showed that the benefits of structural holes flow mainly to members of senior management. Other research has shown that the benefits of cohesion flow mainly to people occupying lower hierarchical levels in organizations, for whom issues of organizational identity and belonging remain salient for potential career advancement (Podolny and Baron 1997).

The cultural contingency effects Another contingent factor that has been found to moderate the effect of structural holes includes the specific cultural and organizational context in which the mechanisms of social capital operate. In stark contrast to the results of studies using Western samples, the empirical findings of Xiao and Tsui (2007) show that in a collectivistic Chinese culture, structural holes in an employee's career network tend to be detrimental to the employee's career development. Moreover, it has been suggested that the network consequences of social capital may differ across organizations: whereas in a market-like, low commitment organizational culture, structural holes bring positive returns to individual actors, it is network closure that appears to bring advantages to the actors by

facilitating trust, reciprocity, and reputation in a clan-like, cohesive, high-commitment organization with a strong cooperative culture (Lazega 2001; Xiao and Tsui 2007).

Future Directions

The existing work utilizing the insights from Burt's structural-hole theory has recently been extended in a number of interesting directions, namely, explicit inclusion of actor characteristics, agency, and cognitions, as well as increasing use of longitudinal (dynamic) research designs. Drawing inspiration from the leading ideas of social network research, new theory and innovative hypotheses are being proposed, providing additional valuable insights.

Actor characteristics Researchers have increasingly started to incorporate personality variables in their study designs (e.g., self-monitoring) as potential predictors of variance in network outcomes (Kilduff and Krackhardt 2008; Mehra et al. 2001). People with different self-monitoring orientations have been suggested to occupy different structural positions. High self-monitors, relative to low self-monitors, tend to ingratiate themselves into distinctly different social circles of acquaintances with few links between these clusters and thereby occupy structural holes. Burt's (2005:34) "structural entrepreneur personality index" quantifies the individual inclination to exploit social resources. Structural entrepreneurs recognize the opportunities offered by structurally advantageous positions and place themselves in the "hole" by initiating ties with actors from opposite sides of the hole who can subsequently be played off against each other. This recent work challenges the ideological refusal of the traditional social network research to acknowledge ways in which individual actors differ in their attributes and actively explore the possibility of complementary synergies between actors and network structure (Kilduff and Brass 2010). Future research on personality and social networks is likely to be generative of compelling insights on the link between individual attributes and structural outcomes.

Agency Social network research also moves forward by explicitly assuming that actors differ in their abilities, skills, and motivation to take advantage of advantageous network positions. The earlier research has shown that some individuals can choose not to reap the profits derived from their network (Burt 1992). Drawing on these earlier findings, the more recent studies suggest that the more strategically skilled group members enjoy greater access to network resources and appear to be more competent at utilizing and leveraging these resources to advance their career and performance (Ferris et al. 2007; Labun 2012; Wei et al. 2012). This work uncovers the comprehensive role that individual strategic skills may play in the process of network resource building. Following this line of analysis, the incorporation of additional types of personal or social influence skills that may affect network resource development would be an interesting and fruitful avenue for future research. Moreover, future work might consider more closely the question of how much control actors have over the networks that constrain and enable their behaviors (Kilduff and Brass 2010).

Cognition Another research area drawing from the core concepts of social network program puts a special emphasis on subjective meanings (i.e., cognitions) inherent in networks rather than on "concrete" relations such as exchanges between actors (Kilduff and Brass 2010). The cognitive social network research line has led to the conceptualization of networks as "prisms" through which others' reputations and potentials are perceived, as well as "pipes" through which resources flow (Podolny 2001). Perceived status of one's exchange partners may indeed act as a distorting prism filtering attributions concerning the focal individual (Labun 2012): having a trust relationship with a superior had a significant positive effect on other's perceptions of one's power. The role of cognitions inherent in networks was further accentuated in a study demonstrating that individuals tend to bias perceptions to highlight small world features of clustering and connectivity (Kilduff et al. 2008): across four different organizational friendship networks, people have been found to perceive more "small worldedness"

than was actually the case, including the perception of more network clustering than actually existed and the attribution of more popularity and brokerage to the perceived popular than to the actually popular.

Network dynamics Finally, longitudinal research designs that allow considering and effectively addressing the dynamic nature of networks is likely to drive the social network research program forward. The very recent analytical developments (Snijders et al. 2010) allow unraveling and tackling the intriguing novel phenomena concerning interpersonal network change, coevolution of networks, and individual behavior (e.g., friendship, music preferences, and alcohol consumption (Steglich et al. 2006); friendship and smoking behavior (Mercken et al. 2010)), as well as different types of networks (e.g., friendship and gossip (Ellwardt et al. 2012); friendship and power (Labun 2012)). For example, the friendship and power study showed that power perceptions breed friendship (Labun 2012). Through a power attribution to a colleague, an individual may signal his or her trust in the colleague's competence, thereby triggering a friendship nomination from/facilitating friendship with him or her. The multiplex effect showed up also when analyzing the conditions that influence the formation of social ties (i.e., friendship) to the high-power organizational actors. However, in this case, the relationship between two networks appeared to depend on individual's strategic orientations (Labun 2012). This emergent research contributes to a better understanding of the coevolution of multiplex networks as well as networks and individual behavior, thereby allowing us to fully grasp the antecedents, dynamics, and consequences of the "informal organization."

Using a game theoretic model of network formation, Buskens and Van de Rijt (2008) confirm Burt's own speculation that when the monopoly on structural entrepreneurship is lifted, structural advantages most likely disappear (Burt 2005): when everyone strives for structural holes, no one will be able to maintain a structural advantage in the long run (Buskens and Van de Rijt 2008).

It would be interesting to perform further empirical studies in different types of organizational settings to help elucidate the dynamics of structural holes. The ongoing methodological advancements and the theoretical insights gained from the above-mentioned recent work are certainly beneficial for the future development and possible extension of existing structural-hole research.

Cross-References

- ▸ Actor-Based Models for Longitudinal Networks
- ▸ Centrality Measures
- ▸ Exchange Networks
- ▸ Exponential Random Graph Models
- ▸ Futures of Social Networks: Where Are Trends Heading?
- ▸ Game Theory and Social Networks
- ▸ Personal Networks: The Intertwining of Ties, Internet and Geography
- ▸ Social Capital
- ▸ Social Influence Analysis
- ▸ Trust in Social Networks

References

Ahuja G (2000) Collaboration networks, structural holes, and innovation: a longitudinal study. Adm Sci Q 45:425–455

Balkundi P, Wang L, Harrison DA (2009) Bridging the gap: consequences of structural hole spanning at multiple levels. Working paper, SUNY, Buffalo

Bonacich P (1987) Power and centrality: a family of measures. Am J Sociol 92:1170–1182

Brass DJ (2011) A social network perspective on organizational psychology. In: Kozlowski SWJ (ed) The oxford handbook of organizational psychology. Oxford University Press, New York

Burt RS (1982) Toward a structural theory of action. Academic, New York

Burt RS (1992) Structural holes: the social structure of competition. Harvard University Press, Cambridge

Burt RS (1997) The contingent value of social capital. Adm Sci Q 42:339–365

Burt RS (1998) The gender of social capital. Ration Soc 10:5–46

Burt RS (2000) The network structure of social capital. Res Organ Behav 22:345–423

Burt RS (2001) Bandwidth and echo: trust, information, and gossip in social networks. In: Casella A, Rauch JE (eds) Networks and markets: contributions from economics and sociology. Russell Sage, New York, pp 30–74
Burt RS (2004) Structural holes and good ideas. Am J Sociol 110:349–399
Burt RS (2005) Brokerage and closure: an introduction to social capital. Oxford University Press, Oxford
Ferris GR, Treadway DC, Perrewe PL, Brouer RL, Douglas C, Lux S (2007) Political skill in organizations. J Manage 33:290–320
Freeman LC (1979) Centrality in social networks: conceptual clarification. Soc Netw 1:215–239
Granovetter M (1973) The strength of weak ties. Am J Sociol 6:1360–1380
Hargadon AB, Sutton RI (1997) Technology brokering and innovation in a product development firm. Adm Sci Q 42:716–749
Kilduff M, Brass DJ (2010) Organizational social network research: core ideas and key debates. Acad Manage Ann 4:317–357
Kilduff M, Krackhardt D (2008) Interpersonal networks in organizations. Cambridge University Press, Cambridge
Labun A (2012) Social networks and informal power in organizations. ICS Dissertation series, Groningen, p 194
Lazega E (2001) The collegial phenomenon: the social mechanisms of cooperation among peers in a corporate law partnership. Oxford University Press, Oxford
McCall MW (1979) Power, authority, and influence. In: Kerr S (ed) Organizational behavior. Grid, Columbus, pp 185–206
Mechanic D (1962) Sources of power of lower participants in complex organizations. Adm Sci Q 7: 349–364
Mehra A, Kilduff M, Brass DJ (2001) The social networks of high and low self-monitors: implications for workplace performance. Adm Sci Q 46:121–146
Pettigrew AM (1972) Informational control as a power resource. Sociology 6:187–204
Pfeffer J (1981) Power in organizations. Pitman, Marshfield
Podolny JM (2001) Networks as the pipes and prisms of the market. Am J Sociol 107:33–60
Podolny JM, Baron JN (1997) Resources and relationships: social networks and mobility in the workplace. Am Sociol Rev 62:673–693
Seibert SE, Kraimer ML, Liden RC (2001) A social capital theory of career success. Acad Manage J 44: 219–237
Wei L, Chiang FFT, Wu L (2012) Developing and utilizing network resources: roles of political skill. J Manage Stud 49:381–402
Xiao Z, Tsui AS (2007) When brokers may not work: the cultural contingency of social capital in Chinese high-tech firms. Adm Sci Q 52:1–31

Recommended Reading

Buskens V, Van de Rijt A (2008) Dynamics of networks if everyone strives for structural holes. Am J Sociol 114:371–407
Ellwardt L, Steglich CEG, Wittek R (2012) The co-evolution of friendship and gossip in workplace social networks. Soc Netw 34:623–633
Kilduff M, Crossland C, Tsai W, Krackhardt D (2008) Network perceptions versus reality: a small world after all? Organ Behav Hum Decis Process 107:15–28
Mercken L, Snijders TAB, Steglich CEG, Vartiainen E, de Vries H (2010) Dynamics of adolescent friendship networks and smoking behavior. Soc Netw 32:72–81
Snijders TAB, Van de Bunt GG, Steglich CEG (2010) Introduction to stochastic actor-based models for network dynamics. Soc Netw 32:44–60
Steglich CEG, Snijders TAB, West P (2006) Applying SIENA: an illustrative analysis of the co-evolution of adolescent's friendship networks, taste in music, and alcohol consumption. Methodology 2:48–56

Structural Measure, Link Mining Metric

▸ Role Discovery

Structural Roles

▸ Querying Volatile and Dynamic Networks

Structuralism

▸ Network Analysis in French Sociology and Anthropology

Subgraph Count

▸ Motif Analysis

Subgraph Discovery

▸ Subgraph Extraction for Trust Inference in Social Networks

Subgraph Evolution

▸ Motif Analysis

Subgraph Extraction for Trust Inference in Social Networks

Yuan Yao[1], Hanghang Tong[2], Feng Xu[1], and Jian Lu[1]
[1]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, Jiangsu, China
[2]CUNY city college, New York, NY, USA

Synonyms

Interaction network; Subgraph discovery; Trust evaluation; Trust network; Trust prediction

Glossary

Social Network A graph in which the nodes represent the participants in the network and the edges represent relationships

Trust-Based Social Network A directed weighted graph in which the nodes represent the participants in the network, the edges represent trust relationships and the weight on each edge indicates the local trust value derived from the historical interactions

Trust Inference A mechanism to build new trust relationships based on existing ones

Subgraph A subgraph of graph G is a graph whose node set is a subset of that of G, and whose edge set is a subset of that of G restricted to the node subset

Subgraph Extraction Discovery of a subgraph from a whole graph

Definition

Trust-based social networks might contain a large amount of redundant information, making existing trust inference suffer from the scalability and usability issues. Therefore, it is natural to apply subgraph extraction as an intermediate step to speed up as well as to interpret the trust inference process.

Introduction

Trust inference, which aims to infer a trustworthiness score from the trustor to the trustee in the underlying social network, is an essential task in many real-world applications including e-commerce (Xiong and Liu 2004), peer-to-peer networks (Kamvar et al. 2003), and mobile ad hoc networks (Buchegger and Le Boudec 2004).

To date, many trust inference algorithms have been proposed, which can be categorized into two main classes (see the next section for a review): (a) path-based inference (Mui et al. 2002; Wang and Singh 2006; Hang et al. 2009; Wang and Wu 2011) and (b) component-based inference (Guha et al. 2004; Massa and Avesani 2005; Ziegler and Lausen 2005; Zhou and Hwang 2007).

Despite their own success, most of the existing inference algorithms have two limitations. The first challenge lies in *scalability* – many existing algorithms become very time-consuming or even computationally infeasible for the graphs with more than thousands of nodes. Additionally, some algorithms assume the existence of a subgraph while how to construct such a subgraph remains an open issue (Wang and Wu 2011). The second challenge is the *usability* of the inference results. Most, if not all, of the existing inference algorithms output an abstract numerical trustworthiness score. This gives a quantitative measure of *to what extent* the trustor should trust the trustee but gives few cues on *how* the trustworthiness score is inferred. This usability/interpretation issue becomes more evident when the size of the

underlying graph increases, since we cannot even display the entire graph to the end users (see Fig. 9 for an example).

In this article, we propose subgraph extraction to address these challenges. The core of our subgraph extraction consists of two stages: *path selection* and *component induction*. In the first (path selection) stage, we extract a few, important paths from the trustor to the trustee. In the second (component induction) stage, we propose a novel evolutionary algorithm to generate a small subgraph based on the extracted paths. The outputs of these two stages are then used as an intermediate step to speed up the path-based inference and component-based inference algorithms, respectively. Our experimental evaluations on real graphs show that the proposed method can significantly accelerate existing trust inference algorithms (up to 2,400× speedup) while maintaining high accuracy (P-error is less than 0.04). In addition, the extracted subgraph provides an intuitive way to interpret the resulting trustworthiness score by presenting a concise summarization on the relationship from the trustor to the trustee. To the best of our knowledge, we are the first to propose subgraph extraction for trust inference. We believe that our work can improve most of the existing trust inference algorithms by (1) scaling up as well as (2) delivering more usable (i.e., interpretation-friendly) inference results to the end users.

Historical Background

We review the historical background in this section, which can be categorized into two parts: trust inference algorithms and subgraph extraction.

Trust Inference

We categorize existing trust inference algorithms into two main classes: path-based trust inference and component-based trust inference.

In the first class of path-based inference, trust is propagated along a path from the trustor to the trustee, and the propagated trust from multiple paths can be combined to form a final trustworthiness score. For example, Wang and Singh (2006, 2007) as well as Hang et al. (2009) propose operators to concatenate trust along a path and aggregate trust from multiple paths. Liu et al. (2010) argue that not only trust values but social relationships and recommendation role are important for trust inference. However, these algorithms are only suitable for small networks due to their complexity. Some other path-based trust inference algorithms, such as Mui et al. (2002) and Wang and Wu (2011), assume the existence of an extracted subgraph while how to construct such a subgraph remains an open issue (Wang and Wu 2011).

In the second class of component-based inference, EigenTrust Kamvar et al. (2003) tries to compute an objective trustworthiness score for each node in the graph. In contrast to EigenTrust, our main focus is to provide support for subjective trust metrics where different trustors can form different opinions on the same trustee. In contrast to path-based trust inference algorithms, there is no explicit concept of paths in component-based trust inference. Instead, existing subjective trust algorithms, including Guha et al. (2004), Massa and Avesani (2005), Ziegler and Lausen (2005), and Nordheimer et al. (2010), take the initial graph as input and treat trust as random walks on a Markov chain or on a graph (Richardson et al. 2003). For example, in MoleTrust (Massa and Avesani 2005) and Appleseed (Ziegler and Lausen 2005), trust propagates along the edges according to the trust values on the edges. Our subgraph extraction method not only can speed up many of these algorithms but also can provide interpretive result which is not considered by the existing algorithms.

Overall, our subgraph extraction is motivated to address the two common challenges (i.e., scalability and usability) shared by most of these existing trust inference algorithms.

Subgraph Extraction

Several end-to-end subgraph extraction algorithms are developed to solve different problems.

In the field of graph mining, Faloutsos et al. (2004) refer to the idea of electrical current where trust relationships are modeled as resistors and try

to find a connection subgraph that maximizes the current flowing from source to target. Later, Tong et al. (2007) generalize the connection subgraph to directed graphs and use the subgraph to compute proximities between nodes. Similar to Tong et al., Koren et al. (2006) also try to induce a subgraph for proximity computation. In addition, Koren et al. search the k-shortest paths to provide a basis for measuring the proximity.

Recently, several algorithms are proposed for reliable subgraph extraction. Among them, Monte Carlo pruning (Hintsanen and Toivonen 2008) measures the relevance of each edge by Monte Carlo simulations and tends to remove the edge of lowest relevance one by one. The most related work is perhaps the randomized Path Covering algorithm (Hintsanen et al. 2010) which also consists of two stages of path sampling and subgraph construction. However, both Monte Carlo pruning and Path Covering tend to find a subgraph with highest probability to be connected, while we aim to find a subgraph to address the scalability and usability issues in trust inference.

The Proposed Subgraph Extraction Method

In this section, we first formalize the subgraph extraction problem for trust inference in social networks and then introduce our proposed solution which consists of two stages: path selection and component induction.

Problem Definition

Following the standard notations in the existing trust inference algorithms, we model the trust relationships in social networks as a weighted directed graph (Barbian 2011; Yao et al. 2011). The nodes of the graph represent the participants in the network, and the weight on each edge indicates the local trust value derived from the historical interactions.

We then categorize the existing trust inference algorithms into two major classes: *path-based trust inference* and *component-based trust inference*.

Definition 1 Path-Based Trust Inference
Path-based trust inference includes the approaches, which are started by the trustor, to evaluating the trustworthiness of the trustee, through a set of paths from the trustor to the trustee in the network.

Definition 2 Component-Based Trust Inference
Component-based trust inference includes the approaches, which are started by the trustor, to evaluating the trustworthiness of the trustee, through a connected component from the trustor to the trustee in the network.

Both classes belong to the subjective trust metrics (Ziegler and Lausen 2005), where different trustors can form different opinions on the same trustee. Accordingly, path-based trust inference such as Mui et al. (2002), Wang and Singh (2006), Liu et al. (2010), Hang et al. (2009), and Wang and Wu (2011) and component-based inference such as (Guha et al. 2004), Massa and Avesani (2005), Ziegler and Lausen (2005), and Zhou and Hwang (2007) all belong to trust inference algorithms. Although the main focus of this article is on the subjective metrics, our proposed subgraph extraction can also be applied to the objective trust metrics.

Despite the success of most existing inference algorithms, they share the scalability and usability limitations. To address these issues, we propose subgraph extraction for trust inference. The core of our subgraph extraction consists of two stages. The first stage, which serves for path-based trust inference, selects a set of paths from the trustor to the trustee. The second stage aims to produce a connected component between the trustor and the trustee for component-based trust inference. In addition, the second stage of our subgraph extraction produces a relatively small subgraph which can be clearly displayed and help the end user better understand the inference result.

We now formally define the subgraph extraction problem for trust inference. In accordance to the corresponding two stages, the problem is divided into two subproblems: *path selection problem* and *component induction problem*.

Definition 3 Path Selection Problem

Given: a weighted directed graph $G(V, E)$; two nodes $s, t \in V$; and an integer K

Find: a set C with K paths from s to t that minimizes the error function $f(C)$

Definition 4 Component Induction Problem

Given: a set C of paths from s to t and an integer N

Find: an induced component $H(V', E')$ with at most N edges that minimizes the error function $g(H)$, where $V' \subseteq \{v|(u, v)$ or $(v, u) \in P, P \in C\}$ and $E' \subseteq \{e|e \in P, P \in C\}$

We next discuss the error function in the definitions. The error function $f(C)$ in Definition 3 indicates the goodness of the extracted paths, and $f(C)$ reaches its minimum value when C contains all the possible paths from s to t. Similarly, the error function $g(H)$ in Definition 4 reaches its minimum value if $H = G$. In this article, we use P-error, which is defined as follows, as the error function for both subproblems, i.e., $f = g =$ P-error.

Definition 5 P-error

For a given trustor-trustee pair, the error function P-error is defined as

$$P\text{-error} = |p_{\text{sub}} - p_{\text{whole}}|,$$

where p_{sub} is the trustworthiness score inferred from the subgraph and p_{whole}, which serves as a ground truth, is the trustworthiness score inferred from the whole graph.

Path Selection

In the path selection stage, we aim to extract a few paths from the trustor to the trustee as an intermediate step to speed up path-based trust inference algorithms. These extracted paths will also serve as the input for the component induction stage.

There are two preprocessing steps in our extraction method. First of all, trust is interpreted as the probability by which the trustor expects that the trustee will perform a given action. This interpretation of trust is adopted by many existing trust inference algorithms, and it allows trust to be multiplicatively propagated along a path (Liu et al. 2010). Second, we transform probability into weight by negative logarithm. Namely, the local trust value on the edge e is interpreted as probability $p(e)$, and the probability $p(e)$ is transformed to weight $w(e) = -\log(p(e))$. Based on these two steps, the weight of a path P can be presented as

$$\begin{aligned} w(P) &= \sum_{e \in P} -\log(p(e)) = -\log(\prod_{e \in P} p(e)) \\ &= -\log(Pr(P)). \end{aligned}$$

As a result, finding a path of high trustworthiness in the original network is equivalent to finding a short path in the transformed network. We will use this transformed weighted graph $G(V, E)$ as the input of our method.

Then, the path selection problem becomes to extract top-k short paths from the trustor to the trustee in the transformed graph $G(V, E)$. Many existing algorithms can be plugged into this stage, such as Yen's k-shortest loopless paths (KS) (Yen 1971), and path sampling (PS) (Hintsanen et al. 2010). In our experiments, we found that KS algorithm performs best even if the multiplicative property of the interpretation does not hold, and we therefore recommend KS in this stage. A brief skeleton of the KS algorithm is shown in Algorithm 4, and the detailed algorithms for KS and PS are presented in the appendix for completeness.

Algorithm 4: KS algorithm (see the appendix for the details)

Input: Weighted directed graph $G(V, E)$, two nodes $s, t \in V$, and a parameter K of path number
Output: Set C with K paths from s to t
1: C = k-shortest(G, s, t, K)
2: **return** C

Algorithm Analysis

The worst-case time complexity of KS is $O(K|V|(|E| + |V|\log|V|))$, which is known as the best result to ensure that k-shortest loopless paths can be found in a directed graph (Hershberger et al. 2007). However, the actual

wall-clock time of KS on many real graphs is often much better than such worst-case scenario (Martins and Pascoal 2003). In fact, based on our experiments, we find that it empirically scales near linearly wrt the graph size $|V|$ in the chosen datasets.

Component Induction

In the component induction, we take the output of path selection stage (i.e., a set of K paths) as input and output a small connected component from the trustor to the trustee. The output of the component induction stage not only acts as an intermediate step to speed up component-based trust inference algorithms but also helps to improve the usability of trust inference by interpreting the inference results for the end users. Notice that although our upcoming proposed algorithm EVO could also be applied on the whole graph, we do not recommend it in practice for the following two reasons: (1) most trustworthy paths have already been captured by the path selection stage (i.e., KS), and (2) applying EVO on the whole graph would cost more memory and time to achieve high accuracy. We will present more detailed experimental evaluations to validate this in the next section.

In general, our proposed EVO algorithm (shown in Algorithm 5) belongs to the so-called evolutionary methods (Bäck 1996). It aims to minimize P-error under the constraint of edge number. The input component $G^c(V^c, E^c)$ is directly induced from the set C of paths from s to t, where $V^c = \{v|(u,v) \in P \text{ or } (v,u) \in P, P \in C\}$ and $E^c = \{e|e \in P, P \in C\}$. There are two implicit parameters in the algorithm, i.e., the initial vector number m and iteration number iter.

We now explain EVO in detail. The first step of EVO is to establish a one-to-one correspondence between the edges in G^c and the elements in vector B. Each element of B is a 0/1 bit where 1 indicates that the corresponding edge exists and 0 indicates otherwise. The vector has exactly $|E^c|$ bits where $|E^c|$ is the edge number of G^c. In the second step, the algorithm generates m vectors $B_1, B_2, \ldots, B_m$, and each of them has at most N 1-bits. In our implementation, we apply a constant-time search in C to find a subset of paths with minimized P-error. In the following steps, EVO adopts *mutation* on each of these vectors to separately generate m new vectors $B_{m+1}, B_{m+2}, \ldots, B_{2m}$. In the mutation from B_i to B_{i+m}, each bit of B_i is changed with probability $1/|E^c|$. If the resulting vector has more than N 1-bits, the mutation operation is redone. The error function, which is P-error in our case, is then computed on each of these $2m$ vectors, and the m vectors with smallest P-error are kept to the next iteration. For efficiency, the P-error computation on vector B herein means computing the P-error between $G^c(V^c, E^c)$ and the component corresponding to the vector B. Namely, we use the input component $G^c(V^c, E^c)$ as an approximation of the ground truth in this stage.

Algorithm 5: EVO algorithm

Input: Set C of paths from s to t and the directly induced component $G^c(V^c, E^c)$, as well as a constraint N of the edge number
Output: Induced component $H(V', E')$ with at most N edges

```
define 0/1 vector B of size |E^c| where each element in B stands for the existence of a corresponding edge in G^c
initialize m vectors S ← {B_1, B_2, ..., B_m}, with at most N 1-bits for each vector
while iter > 0 do
    for each vector B_i in S do
        repeat
            mutate B_i to B_{i+m} with mutation probability 1/|E^c|
        until the number of 1-bits in B_{i+m} ≤ N
    end for
    compute P-error results for the 2m vectors {B_1, B_2, ..., B_{2m}}
    S ← the best m vectors from the 2m ones
    iter ← iter - 1
end while
B_final ← the best vector in S
return the corresponding component H(V', E') of B_final
```

Algorithm Analysis

The time complexity of EVO is summarized in the following lemma, which basically says that the expected time complexity of EVO scales linearly wrt both initial vector number m and iteration number iter.

Subgraph Extraction for Trust Inference in Social Networks, Table 1 High-level statistics of advogato datasets

Graph	Nodes	Edges	Avg. degree	Avg. clustering	Avg. diameter	Date
Advogato-1	279	2,109	15.1	0.45	4.62	2000-02-05
Advogato-2	1,261	12,176	19.3	0.36	4.71	2000-07-18
Advogato-3	2,443	22,486	18.4	0.31	4.67	2001-03-06
Advogato-4	3,279	32,743	20.0	0.33	4.74	2002-01-14
Advogato-5	4,158	41,308	19.9	0.33	4.83	2003-03-04
Advogato-6	5,428	51,493	19.0	0.31	4.82	2011-06-23

Lemma 1 *The average-case time complexity of EVO is* $O(\text{iter} \cdot m(|E^c|/N + \theta))$*, where* θ *is the time complexity of the error function computation.*

Proof In the mutation step of EVO, with mutation probability $1/|E^c|$, the expected number of bit changes is 1. This step is expected to be redone only when the number of 1-bits is N and the bit change is from 0 to 1. Under this condition, the probability of bit change from 0 to 1 is $(|E^c| - N)/|E^c|$. Therefore, the expected iteration number of the mutation step is $|E^c|/N$. Therefore, the whole expected time complexity of EVO is $O(\text{iter}(m \cdot |E^c|/N + m\theta)) = O(\text{iter} \cdot m(|E^c|/N + \theta))$, which completes the proof. □

Experimental Evaluation

In this section, we first describe the experimental setup and then present the results.

Experimental Setup

We first describe the datasets and the representatives of path-based and component-based trust inference algorithms. All algorithms are implemented in Java and have been run on a T400 ThinkPad with 1,280 m jvm heap space. Few other activities are done during the experiments.

Datasets Description

We use the advogato (http://www.trustlet.org/wiki/Advogato_dataset) datasets in our experiments, because advogato is a trust-based social network and it contains multilevel trust assertions. There are four levels of trust assertions in the network, i.e., "Observer," "Apprentice," "Journeyer," and "Master." These assertions can be mapped into real numbers in [0,1]. In our experiments, we map "Observer," "Apprentice," "Journeyer," and "Master" to 0.1, 0.4, 0.7, and 0.9, respectively. The statistics of the datasets is shown in Table 1.

Trust Inference Representatives

To evaluate our subgraph extraction method, we need to apply trust inference algorithms on the whole graph and on our extracted subgraph to compare their effectiveness and efficiency. We chose *CertProp* (Hang et al. 2009) as the representative of path-based inference algorithms, and *Appleseed* (Ziegler and Lausen 2005) as the representative of component-based inference algorithms.

P-error computation in CertProp needs to first compute the ground truth p_{whole} by finding all paths from the trustor to the trustee in the whole graph. This computation, however, easily causes the overflow of the jvm heap space even on the advogato-1 graph. Following the suggestions in the original CertProp (Hang et al. 2009), we apply the fixed search strategy and search all paths whose length is not longer than seven as an approximation of the ground truth. For CertProp, we define *collapsed samples* as the trustor-trustee pairs of which the P-error computation either exceeds the range of Java.lang.Double or runs out of the jvm heap space. We randomly select 100 node pairs out of 122 samples, where the rest 22 of them are collapsed samples. Our experimental results are all based on the average of these 100 samples. Notice that, as discussed in the path selection section, the multiplicative property of the probability interpretation does not hold in CertProp. As to Appleseed, we apply linear normalization on the outputs, since the algorithm can produce arbitrary trustworthiness scores.

Experimental Results

We now present the experimental results of our subgraph extraction method. In our experiments, the effectiveness, efficiency comparisons, and interpretation results are all based on the advogato-1 graph, as we found CertProp on the whole graph becomes computationally infeasible on all the other larger datasets. We evaluate the scalability of our method using all the datasets (i.e., advogato-1 to advogato-6). As for EVO, we set $m = 5$ and iter $= 10$ unless otherwise specified. The edge constraint N is set as $K/2$.

Effectiveness

For effectiveness, we first study how CertProp and Appleseed perform on the KS subgraph (the output of path selection stage) and EVO subgraph (the output of component induction stage), respectively. The results are shown in Fig. 1. We can observe that all the P-error values of CertProp and Appleseed are less than 0.04, indicating that our extracted subgraphs, which are based on a small set of carefully selected paths and an evolutionary strategy, provide high accuracy for the trust inference algorithms.

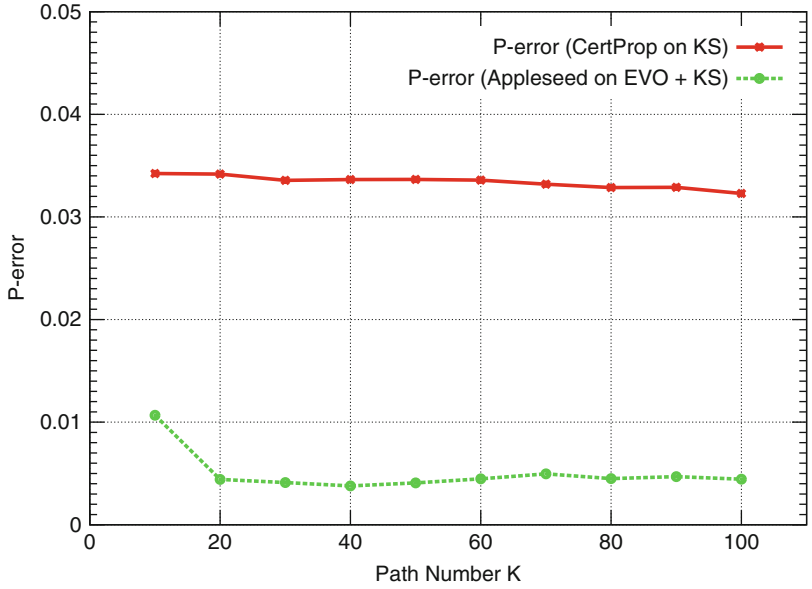
Remember that the proposed EVO is always applied on the output of the path selection stage (referred to as "EVO + KS"). Here, for comparison purpose, we also apply EVO on the entire graph (referred to as "EVO + whole graph"). With the same parameter setting, the results are shown in Fig. 2. It can be seen that EVO on KS outperforms EVO on the whole graph. The reason is as follows. As an evolutionary algorithm, EVO (either on KS or on the entire graph) finds a local minima. By restricting the search space to those highly trustworthy paths (i.e., the output of KS), it converges to a better local minima in terms of P-error.

Finally, to compare EVO with existing component induction algorithms, we implement the *Monte Carlo pruning (MC)* method (Hintsanen and Toivonen 2008) and the *proximity extraction (PE)* method (Koren et al. 2006). As mentioned in the historical background, MC is proposed for the reliable subgraph extraction problem. The key idea of MC is to measure a relevance score for each edge by Monte Carlo simulations and then remove the edges of lowest relevance scores. On the other hand, PE is proposed for the proximity computation problem where a small set of paths are selected to maximize the proposed proximity objective function. We plot the comparison results in Fig. 3. Again, we can see that EVO outperforms both MC and PE wrt P-error. In fact, MC induces a component by successively deleting edges (edge-level component induction), while PE only selects a smaller set of paths (path-level component induction). Our EVO algorithm outperforms MC and PE because EVO combines these two levels of component induction by searching a smaller set of paths in the initial step and then evolving the resulting component on the edge level.

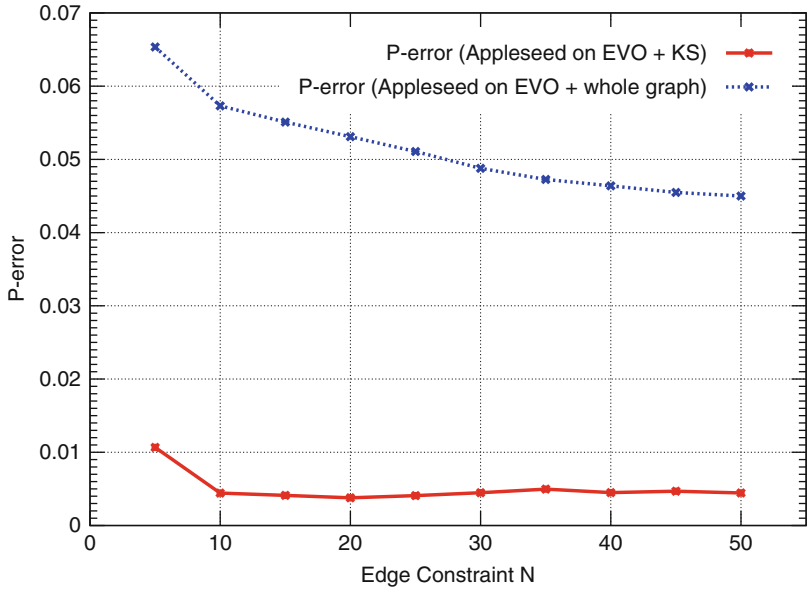
Efficiency

First, we compare the different algorithmic choices in the path selection stage. To this end, we compare the wall-clock time of KS with an alternative path selection algorithm *path sampling (PS)* (Hintsanen et al. 2010). The results are shown in Fig. 4. Note that the y-axis is of log scale. As we can see from the figure, although PS is slightly faster than KS when $K = 5$, the wall-clock time of PS is much longer than that of KS when K is greater than 30. For example, the wall-clock time of PS is more than 170× longer than that of KS when $K = 100$. Therefore, we recommend using KS for path selection.

Next, we study the computational savings by applying the proposed subgraph extraction as the intermediate steps for the existing trust inference algorithms. To this end, we report the wall-clock time of CertProp on the output of the path selection stage and Appleseed on the output of the component induction stage, respectively. The results are shown in Fig. 5 where the y-axis is of log scale. Notice that the reported time includes the wall-clock time of both subgraph extraction and trust inference. In the figure, we also plot the wall-clock time of CertProp and Appleseed on the entire graph for comparison. We can see that our subgraph extraction method saves the wall-clock time for both path-based trust inference and component-based trust inference, especially for the former one. For example,



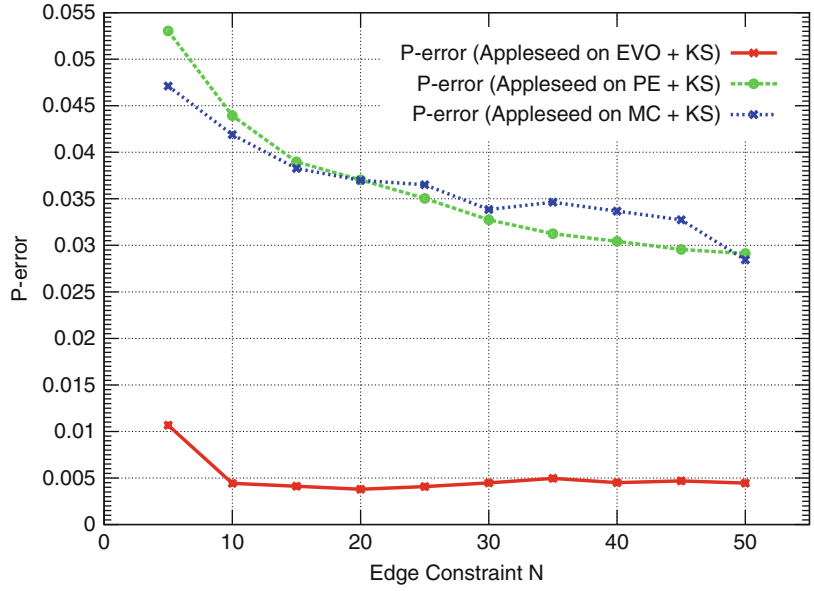
Subgraph Extraction for Trust Inference in Social Networks, Fig. 1 Effectiveness of our subgraph extraction method with edge number constraint $N = K/2$. In all cases, the P-error is less than 0.04



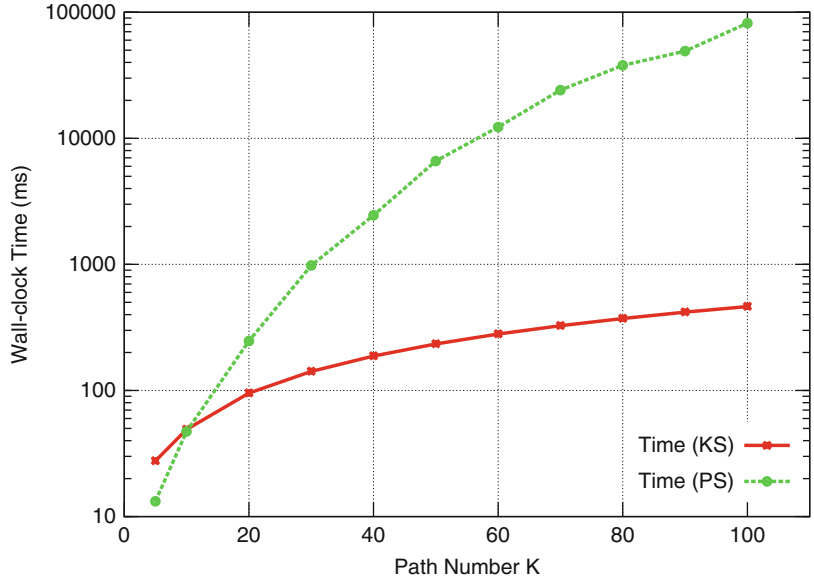
Subgraph Extraction for Trust Inference in Social Networks, Fig. 2 Comparison of EVO on KS vs. EVO on the whole graph with edge number constraint $N = K/2$. EVO on KS outperforms EVO on the whole graph

when $K = 10$, our subgraph extraction method achieves up to 2,400× and 5.4× speedup for CertProp and Appleseed, respectively. Even when K grows to more than 60, our method can still achieve 200 – 400× speedup for CertProp.

Next, we compare the efficiency between applying EVO on KS and applying EVO on the whole graph. With $N = K/2$, the results are shown in Fig. 6. As we can see, the wall-clock time of EVO on KS (which includes the



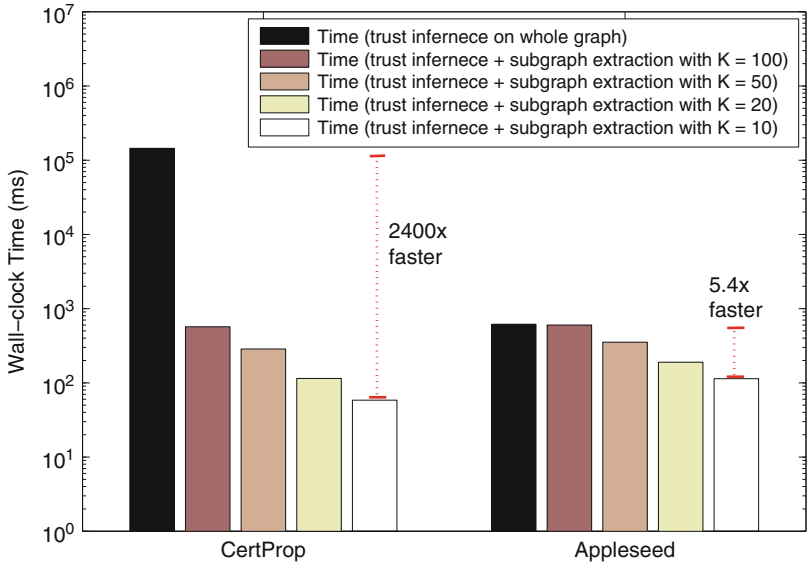
Subgraph Extraction for Trust Inference in Social Networks, Fig. 3 Comparison of different component induction algorithms with edge number constraint $N = K/2$. EVO outperforms the existing component induction algorithms



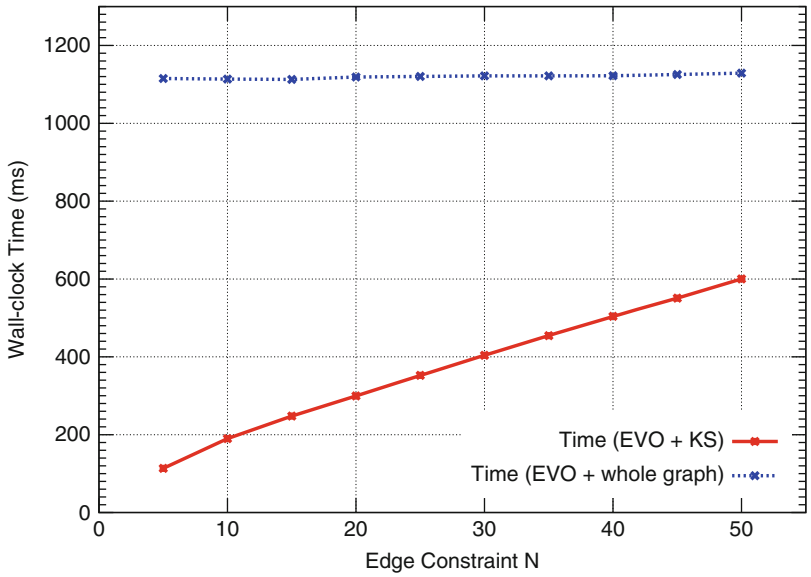
Subgraph Extraction for Trust Inference in Social Networks, Fig. 4 The average wall-clock time of KS and PS. The average wall-clock time of KS is much faster than that of PS when K is greater than 30

wall-clock time of both EVO and KS) is much faster than EVO on the whole graph. Together with the effectiveness results (Fig. 2), we recommend running EVO on the KS subgraph in practice.

Finally, we evaluate how the parameters m and iter in EVO affect the wall-clock time. In this experiment, we fix $K = 20$ and $N = 10$, and the results are shown in Fig. 7. We can observe that the wall-clock time of EVO scales linearly



Subgraph Extraction for Trust Inference in Social Networks, Fig. 5 The average wall-clock time of CertProp on KS and Appleseed on KS + EVO. We achieve up to 2,400× speedup



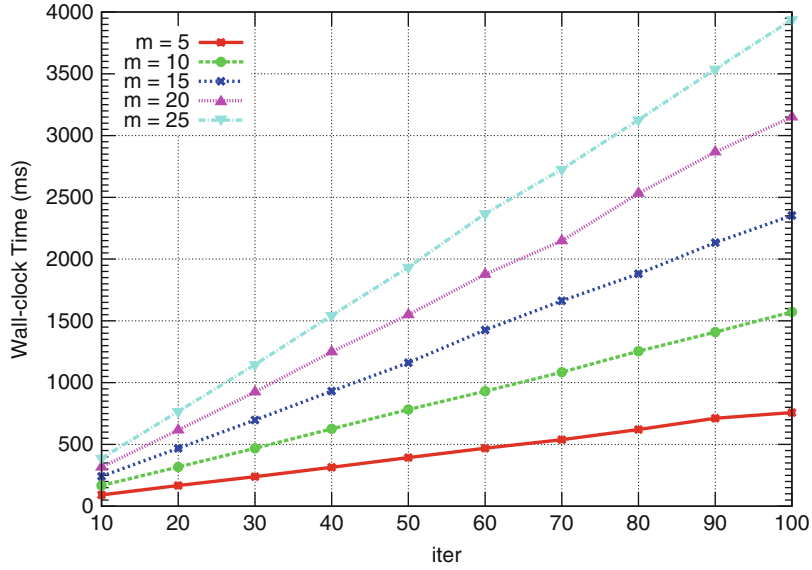
Subgraph Extraction for Trust Inference in Social Networks, Fig. 6 The average wall-clock time of EVO on KS and EVO on the whole graph with edge number constraint $N = K/2$. EVO on KS is much faster

wrt iter for any fixed m, which is consistent with the time complexity analysis shown before.

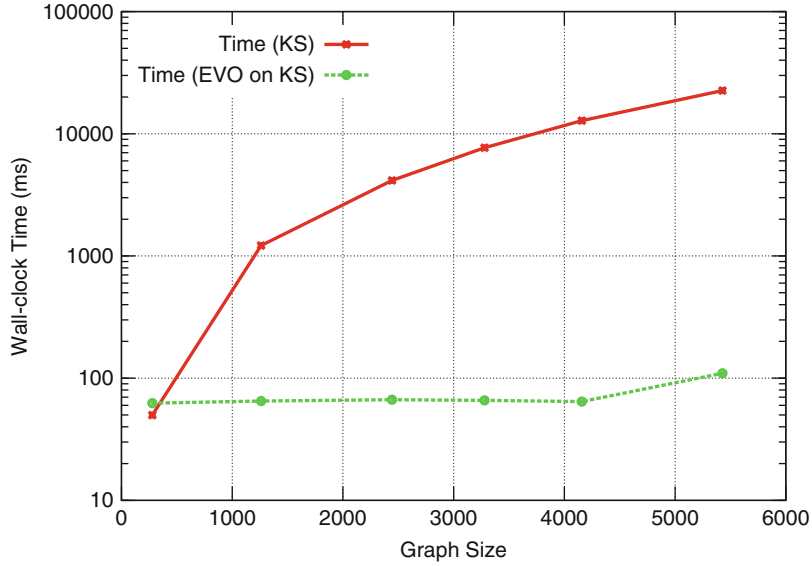
Scalability

We now evaluate the scalability of our method on datasets advogato-1 to advogato-6. Figure 8 shows the results, where the y-axis is of log scale. In this experiment, we fix $K = 10$ and $N = 5$.

We can observe from the figure that even on the largest graph of 5,428 nodes and 51,293 edges, KS can help to infer the trustworthiness score within 25 s. In addition, KS scales near



Subgraph Extraction for Trust Inference in Social Networks, Fig. 7 The average wall-clock time of EVO with $K = 20$ and $N = 10$. EVO scales linearly wrt iter for the fixed m

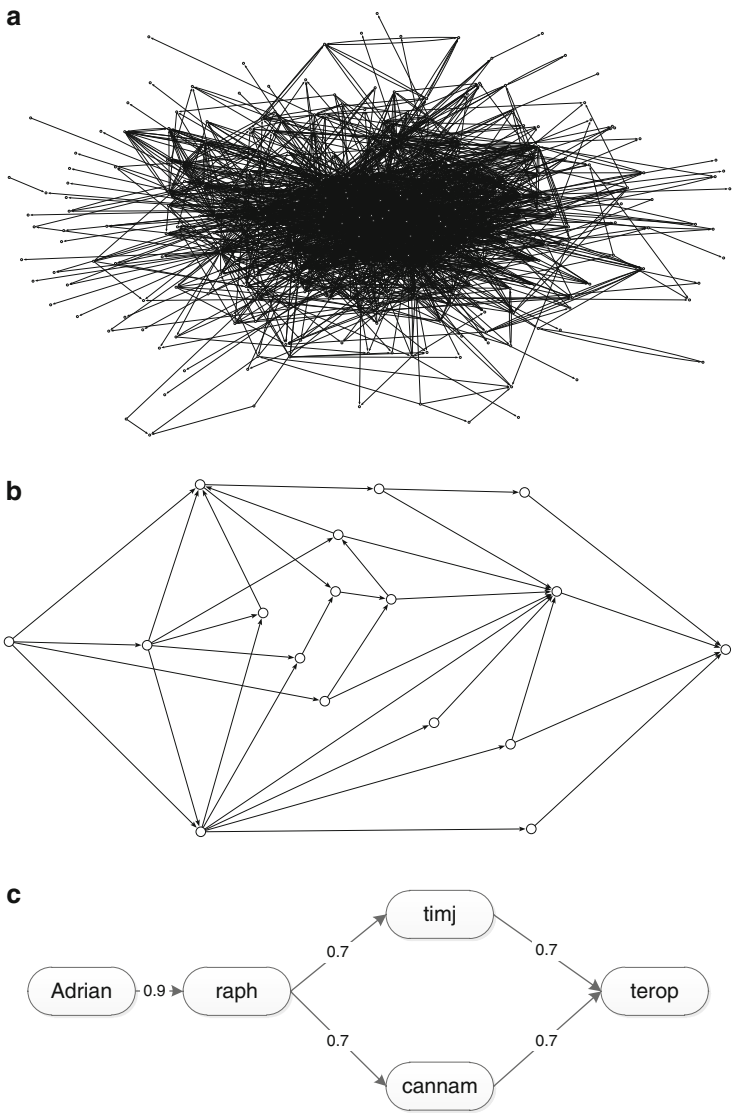


Subgraph Extraction for Trust Inference in Social Networks, Fig. 8 The scalability of our subgraph extraction method. KS scales near linearly wrt the graph size, while the wall-clock time of EVO stays almost constant

linearly wrt the underlying graph size. As to EVO, the wall-clock time stays stable in spite of the growth of the graph size. The reason is that $|E^c|$ scales near linearly to K due to many overlapping edges and N is set to $K/2$. Consequently, $|E^c|/N$ is close to a constant, and the time complexity of EVO can be approximated to $O(\text{iter} \cdot m \cdot \theta)$.

Usability/Interpretation

Another important goal of the proposed EVO is to improve the usability in trust inference by



Subgraph Extraction for Trust Inference in Social Networks, Fig. 9 The interpretation example of the whole graph, KS-20, and EVO-10 on KS-20. (**a**) The original whole graph. (**b**) KS subgraph with $K = 20$. The paths are from "Adrian" (the leftmost node) to "terop" (the rightmost node). (**c**) EVO subgraph with $N = 10$ on KS-20. The component is from "Adrian" to "terop"

interpreting the inferred trustworthiness score for end users. An illustrative example is shown in Fig. 9. The whole graph and the induced KS subgraph by the path selection stage are also plotted for comparison.

From the figures, we can see that the whole graph is hard for interpretation. As to the KS subgraph, although the number of edges has significantly decreased compared with the original whole graph, there are still some redundant edges which might diverge end users' attention. On the other hand, the EVO subgraph only presents the most important participants and their trust opinions, providing a much clearer explanation on how the trustworthiness score is inferred.

Future Directions

On one hand, much of the research in trust inference focuses on the inference accuracy, while inference efficiency is also important in real-world trust inference applications, especially in those online applications. Future work should be able to find the best trade-offs between effectiveness and efficiency according to the specific applications.

Algorithm 6: Detailed KS algorithm

Input: Weighted directed graph $G(V, E)$, two nodes $s, t \in V$, and a parameter K of path number
Output: Set C with K paths from s to t

$X \leftarrow$ shortest path from s to t
$C \leftarrow$ shortest path from s to t
while $|C| < K$ and $X \neq \emptyset$ **do**
  $P \leftarrow$ remove the shortest path in X
  $d \leftarrow$ the *deviation node* of P
  for each node v between d (inclusive) and trustee t (exclusive) in P **do**
    $pre \leftarrow$ subpath from trustor s to v in P
    $post \leftarrow$ the *deviated shortest path* from v to t
    combine pre and $post$, and add it to X
  end for
  $C \leftarrow C +$ the shortest path in X
end while
return C

Algorithm 7: PS algorithm

Input: Weighted directed graph $G(V, E)$, two nodes $s, t \in V$, and a parameter K of path number
Output: Set C with K paths from s to t

$C \leftarrow$ shortest path from s to t
while $|C| < K$ **do**
  re-decide all the edges in E
  for each path P in C **do**
    if P is decided as *true* **then**
      $F \leftarrow F + P$
    end if
  end for
  while $F \neq \emptyset$ **do**
    re-decide the most overlapped edge in F as failed
    remove failed paths from F, if there are any
  end while
  $P \leftarrow$ the shortest path among the non-failed edges from s to t
  if $P \neq \emptyset$ **then**
    $C \leftarrow C + P$
  end if
end while
return C

On the other hand, we believe that usability is becoming a new requirement for trust inference. Users start to care about not only who they should trust but also why they should trust. It is also interesting to incorporate distrust in the subgraph extraction as users may also concern about why they should not trust someone.

Acknowledgment

This work is supported by the National 863 Program of China (No. 2012AA011205), and the National Natural Science Foundation of China (No. 91318301, 61021062, 61073030). The second author was partly sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053.

Appendix

To find K short paths from graph $G(V, E)$ in the path selection stage, many existing algorithms can be used. We consider two representative algorithms from the literature. Here, we present the detailed algorithm description for completeness.

The first algorithm is *Yen's k-shortest loopless paths (KS)* algorithm (Yen 1971), which is shown in Algorithm 6.

In the algorithm, we use Dijkstra's algorithm for finding a shortest path. All the computed paths are loopless by temporarily removing visited nodes. The key idea of the KS algorithm is *deviation*. The *deviation node* d of path P is the node that makes P deviate from existing paths in the candidate set C. For each node v between d (inclusive) and trustee t (exclusive) in P, the *deviated shortest path* from node v to t is computed by temporarily removing the edge starting at v in P. The computed deviated shortest path *post* and the subpath *pre* (the path from s to v in P) are combined to form a possible path candidate. For the nodes before d, possible shortest paths are already computed and included in X. Based on deviation, KS finds the K-shortest paths from trustor s to trustee t one by one. Following Martins and Pascoal's implementation (Martins and Pascoal 2003), we compute the deviated shortest path from deviation node d to the trustee in a reverse order.

The other algorithm is the randomized algorithm *path sampling (PS)* (Hintsanen et al. 2010), which is proposed for the *most reliable subgraph problem* (Hintsanen and Toivonen 2008). While PS is proposed for undirected graphs, trust

relationships in social networks should be directed as trust is asymmetric in nature (Golbeck and Hendler 2006). Therefore, we adapt PS (as shown in Algorithm 7) for a directed graph.

PS considers the input graph as a Bernoulli random graph (Robins et al. 2007), and the algorithm is based on the *edge decision* of this random graph. An edge is randomly decided as true with probability $p(e)$, and a path is decided as true if all the edges on the path are decided as true. At the beginning of each iteration, all the edges of the graph are re-decided, and these *graph decisions* provide opportunities for distrust information to be contained. Like KS, PS first adds a shortest path into candidate set C. PS then tries to find a graph decision based on which none of the paths in C are true. To avoid the situation when this graph decision is hardly found, PS stores the true paths in C to a temporary set F and deliberately fails the most overlapping edges in F until none of the paths in F are true. Finally, based on the results of graph decision and edge failing, PS finds the shortest path P among the non-failed edges from trustor s to trustee t and adds it to C. The algorithm ends until K paths are found.

PS allows some distrust information to be incorporated into the extracted subgraph, which could in turn lower the P-error based on our experiments. However, the time complexity of PS is difficult to estimate, since the wall-clock time depends on the graph density. In addition, as shown in our experiments, the wall-clock time of PS is especially long when K becomes sufficiently large. We conjecture that PS can be used in dense graphs where numerous paths exist between node pairs.

Cross-References

- ▸ Computational Trust Models
- ▸ Trust in Social Networks

References

Bäck T (1996) Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms. Oxford University Press, New York

Barbian G (2011) Assessing trust by disclosure in online social networks. In: Proceedings of the international conference on advances in social networks analysis and mining, ASONAM '11, Kaohsiung, pp 163–170

Buchegger S, Le Boudec JY (2004) A robust reputation system for mobile ad-hoc networks. Technical report, KTH Royal Institute of Technology, Theoretical Computer Science Group

Faloutsos C, McCurley KS, Tomkins A (2004) Fast discovery of connection subgraphs. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, ACM, KDD '04, Seattle, pp 118–127

Golbeck J, Hendler J (2006) Inferring binary trust relationships in web-based social networks. ACM Trans Internet Technol 6:497–529

Guha R, Kumar R, Raghavan P, Tomkins A (2004) Propagation of trust and distrust. In: Proceedings of the 13th international conference on world wide web, WWW '04, New York. ACM, pp 403–412

Hang CW, Wang Y, Singh MP (2009) Operators for propagating trust and their evaluation in social networks. In: Proceedings of the 8th international conference on autonomous agents and multiagent systems, AAMAS '09, Budapest, vol 2. International Foundation for Autonomous Agents and Multiagent Systems, pp 1025–1032

Hershberger J, Maxel M, Suri S (2007) Finding the k shortest simple paths: a new algorithm and its implementation. ACM Trans Algorithms 3(4):45

Hintsanen P, Toivonen H (2008) Finding reliable subgraphs from large probabilistic graphs. Data Mini Knowl Discov 17(1):3–23

Hintsanen P, Toivonen H, Sevon P (2010) Fast discovery of reliable subnetworks. In: Proceedings of the international conference on advances in social networks analysis and mining, ASONAM '10, Odense, pp 104–111

Kamvar SD, Schlosser MT, Garcia-Molina H (2003) The eigentrust algorithm for reputation management in p2p networks. In: Proceedings of the 12th international conference on world wide web, WWW '03, Budapest. ACM, pp 640–651

Koren Y, North S, Volinsky C (2006) Measuring and extracting proximity in networks. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '06, Philadelphia. ACM, pp 245–255

Liu G, Wang Y, Orgun M (2010) Optimal social trust path selection in complex social networks. In: Proceedings of the twenty-fourth AAAI conference on artificial intelligence, AAAI '10, Atlanta, pp 1391–1398

Martins E, Pascoal M (2003) A new implementation of Yen's ranking loopless paths algorithm. 4OR Q J Oper Res 1(2):121–133

Massa P, Avesani P (2005) Controversial users demand local trust metrics: an experimental study on epinions.com community. In: Proceedings of the AAAI conference on artificial intelligence, AAAI '05, Pittsburgh, pp 121–126

Mui L, Mohtashemi M, Halberstadt A (2002) A computational model of trust and reputation. In: Proceedings of the 35th annual Hawaii international conference on system sciences, HICSS '02, Big Island. IEEE, pp 2431–2439
Nordheimer K, Schulze T, Veit D (2010) Trustworthiness in networks: a simulation approach for approximating local trust and distrust values. In: Trust management IV, Morioka. IFIP advances in information and communication technology, vol 321. Springer, Boston, pp 157–171
Richardson M, Agrawal R, Domingos P (2003) Trust management for the semantic web. In: The semantic web, Sanibel Island. Lecture notes in computer science, vol 2870. Springer, Berlin/Heidelberg, pp 351–368
Robins G, Pattison P, Kalish Y, Lusher D (2007) An introduction to exponential random graph (p*) models for social networks. Soc Netw 29(2):173–191
Tong H, Faloutsos C, Koren Y (2007) Fast direction-aware proximity for graph mining. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining, KDD '07, San Jose. ACM, pp 747–756
Wang Y, Singh MP (2006) Trust representation and aggregation in a distributed agent system. In: Proceedings of the AAAI conference on artificial intelligence, AAAI '06, Boston, pp 1425–1430
Wang Y, Singh MP (2007) Formal trust model for multiagent systems. In: Proceedings of the 20th international joint conference on artificial intelligence, IJCAI '07, Hyderabad, pp 1551–1556
Wang G, Wu J (2011) Multi-dimensional evidence-based trust management with multi-trusted paths. Future Gener Comput Syst 27(5):529–538
Xiong L, Liu L (2004) Peertrust: supporting reputation-based trust for peer-to-peer electronic communities. IEEE Trans Knowl Data Eng 16(7):843–857
Yao Y, Zhou J, Han L, Xu F, Lü J (2011) Comparing linkage graph and activity graph of online social networks. In: Social informatics, Singapore. Lecture notes in computer science, vol 6984. Springer, Berlin/Heidelberg, pp 84–97
Yen J (1971) Finding the k shortest loopless paths in a network. Manag Sci 17(11):712–716
Zhou R, Hwang K (2007) Powertrust: a robust and scalable reputation system for trusted peer-to-peer computing. IEEE Trans Parallel Distrib Syst 18(4):460–473
Ziegler C, Lausen G (2005) Propagation models for trust and distrust in social networks. Inf Syst Front 7(4):337–358

Subgraph Identification

▸ Scaling Subgraph Matching Queries in Huge Networks

Subgraph Isomorphic Queries

▸ Scaling Subgraph Matching Queries in Huge Networks

Subgraph Matching

▸ Fraud Detection Using Social Network Analysis, a Case Study

Subgraph Mining

▸ Motif Analysis

Supplier Networks

▸ Inter-organizational Networks

Supply Chain Networks

Dirk Pieter van Donk
Faculty of Economics and Business, Department Operations, University of Groningen, Groningen, The Netherlands

Glossary

Supply Chain Network
Supply Chain Management

Definition

A supply chain network can be defined as a set of interconnected organizations whose different processes and activities produce value (following Slack and Lewis 2011, p. 144), which

is closely related to the definition of supply chain (management) by Christopher (1998) defined as the management of "a network of connected and interdependent organisations mutually and co-operatively working together to control, manage and improve the flow of goods and materials and information from suppliers to end users" (p. 19). In other words, it is the network of organizations that are involved, through upstream and downstream linkages, in the different processes and activities that produce value in the form of products and services in the hands of the ultimate consumer. Thus, for example, a shirt manufacturer is a part of a supply chain that extends upstream through the weavers of fabrics to the manufacturers of fibers, and downstream through distributors and retailers to the final consumer.

Supply Chain Networks

Introduction

Over the last two decades, supply chain management has become one of the major fields of attention both in organizational practice and in academia. This is reflected in a shift from concentrating on internal flows and internal processes to managing buyer-supplier relationships and even more managing relationships across the whole chain and across the network that supplies to and buys goods and services from an organization. Some of the underlying reasons are that organizations compete globally, aim at being good at one specific task, and outsource all remaining activities. Additionally, consumers demand increasingly customized products at the same prices as normal products, while requirements and customer wishes change frequently. In order to be responsive and at the same time cost effective, increased focus on the management of the supply chain or network is needed, often enabled by the use of novel ICT developments that link organizations to their suppliers and buyers. In this perspective, it is stated that "Competing is between supply chains instead of firms." Below we will sketch what supply chain networks are and how to map and understand them. The concluding remarks will relate to some empirical findings and future research directions.

Supply Chain Networks

Supply chain networks cover in principle all organizations that together produce services or products starting from basic raw materials until the final point of consumption. Such an end-to-end, integrated point of view is also reflected in statements like "from paddock to plate," "from mine to motorcar," or "from field to flower" that companies use to reflect their concern for the whole process. In order to better understand supply chain networks and to be able to map and manage them, Lambert and Cooper (2000) propose three key issues or questions:

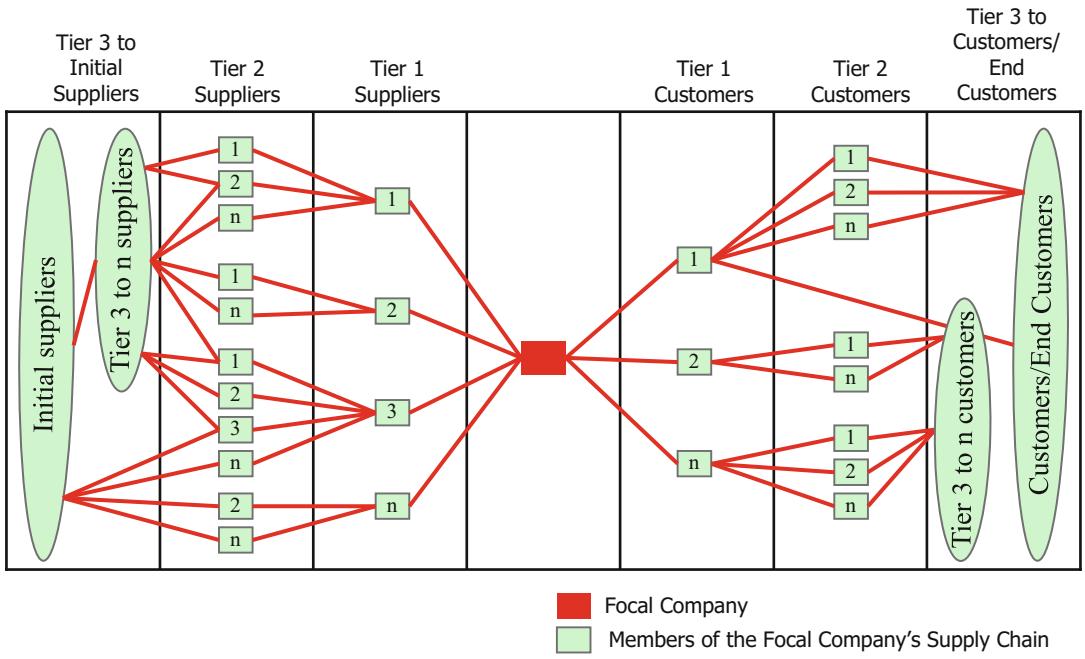
1. Who are the key supply chain members with whom to link processes?
2. What processes should be linked with each of these key supply chain members?
3. What level of integration and management should be applied for each process link?

Below we will shortly address each of these.

Structure

Companies are engaged in multiple networks such as the Chamber of Commerce, and local industrial networks. From a supply chain network perspective, we are mainly interested in those companies that directly are involved in the value-adding activities for particular customers. Other organizations are addressed as supportive, such as banks and local authorities.

Networks can be distinguished along the number of partners in the chain (the number of tiers or horizontal structure) and the number of organizations (competitors, suppliers, or customers) at each level or tier in the network (vertical structure). Probably different from social networks, supply chain networks are always considered from a specific point of view, taking the perspective of one single company that is labeled as the focal company. Therefore, in drawing a network, there is always one central point and it might be evident that the position of the focal firm in the network (being more located to the source of raw materials or more towards the final consumer), together with the vertical and horizontal position,



Supply Chain Networks, Fig. 1 Supply chain network structure (Source: Lambert and Cooper 2000)

is an important factor for the possibilities, but also the wish, to manage the – entire – network. Examples and studies often focus on large and influential companies and their networks, e.g., Wal-mart, Ford, or Toyota, which are all powerful organizations that are able to direct and manage actively their networks. This might distort a general applicable approach. It might be evident that all organizations have their own network, albeit with rather different characteristics: vertical, and horizontal structure and position in the network together indicate the complexity of the network. Figure 1 sketches a network structure with multiple suppliers/customers, but also suppliers' suppliers and customers' customers, etc. These are respectively indicated as first tier, second tier, etc.

Processes

The key process in any supply chain network is the provision of products or services to final consumers. As such the product flow is the point of departure of any network. The physical product flow involves processes such as transportation, warehousing, manufacturing, and distribution of finished goods. Mostly, depending on the nature of product and the network, each of these physical processes will be executed several times when products go from one stage to another stage in the network. In order to be able to execute those processes adequately, information processes have to be well organized as well, often enabled by ICT. Such processes relate to the physical flow directly such as ordering and purchasing processes, while others are more supportive and indirectly such as customer service management and customer relationship management. The central tenet of supply chain management is that all such processes need to be well aligned or integrated in order to be successful both within organizations (removing functional barriers) as well as along the whole network between organizations. While often formal alignment is stressed, based on formal ICT systems, there is evidence that personnel contacts between employees of different organizations in a network are important for proper functioning, as well.

Integration

A supply chain network might consist of numerous links and for each of the links different

processes have to be dealt with. In order to be able to manage the network, three types of links can be distinguished: managed links, monitored links, and non-managed links. Managed links will most likely be the links with the main first tier suppliers and first tier customers, but might also be the links with second tier suppliers or customers. For example, it is quite common that car manufacturers manage the relationship between a part manufacturer and a module manufacturer. Monitored links are links that are not actively managed but some control is needed to be sure that such links are properly managed without directly interfering with the day-to-day management of the link. Finally, all other links are non-managed as they are of less interest for the focal company. Even in managed links, it seems likely that not all processes need to be firmly tuned, as will be shortly discussed below.

Reflection and Critical Issues

While theory explains and shows the benefits of supply chain management, there is little empirical evidence of management of whole supply chain networks. Part of that stems from the complexity explained above, and partly it might stem from the difference between the rhetoric and the practice of supply chain management. As Storey et al. (2006) explain, literature (as probably in more management areas) is not always clear in distinguishing between description and prescription. Part of the reality of supply chain management is that only a limited part of the chain is managed, and mostly at the buyer-supplier relationship level and not along the whole chain or network. There is sufficient empirical evidence that shows the benefits of even such seemingly limited types of supply chain integration and management. Apart from the limited scope, there are serious barriers to supply chain integration such as misalignment of organizational and interorganizational performance measurement systems, along with misaligned – interorganizational – information systems and limited information transparency (e.g., Storey et al. 2006). In addition, different contexts (e.g., depending on product and/or market characteristics) might need different approaches, as Fisher (1997) argued. Also here, there is growing empirical evidence that shows the influence of such contextual factors (e.g., Van der Vaart and van Donk 2006; Giménez et al. 2012). Extending the benefits of buyer-supplier relationship management to the network level, while taking into account possible limitations and removing barriers is one of the challenging issues in supply chain networks and their management. Social network theory is certainly one of the theoretical stances that can help to pave the way for further exploration, and understanding of supply chain networks and their behavior.

Cross-References

- ▸ Business-to-Business Marketing
- ▸ Entrepreneurial Networks
- ▸ Inter-organizational Networks

References

Christopher M (1998) Logistics and supply chain management, 2nd edn. Pearson Education, Harlow

Fisher ML (1997) What is the right supply chain for your product? Harv Bus Rev 75(2):105–116

Giménez C, Van der Vaart T, van Donk DP (2012) Supply chain integration and performance: the moderating effect of supply complexity. Int J Oper Prod Manag 32(5):583–610

Lambert DM, Cooper MC (2000) Issues in supply chain management. Ind Mark Manag 29:65–83

Slack N, Lewis M (2011) Operations strategy, 3rd edn. Pearson Education, Harlow

Storey J, Emberson C, Godsell J, Harrison A (2006) Supply chain management: theory, practice and future challenges. Int J Oper Prod Manag 26(7):754–774

Van der Vaart T, van Donk DP (2006) Buyer-focused operations as a supply chain strategy: identifying the influence of business characteristics. Int J Oper Prod Manag 26(1):8–23

Recommended Reading

Handfield RB, Nichols EL (1999) Introduction to supply chain management. Prentice Hall, Upper Saddle River (basic text)

Skjott-Larsen T, Schary PB, Mikkola JH, Kotzab H (2007) Managing the global supply chain, 3rd edn. Copenhagen Business School Press, Copenhagen (advanced text)

Surveillance

▸ Privacy, Dataveillance, and Crime Prevention

Surveys

▸ Questionnaires for Measuring Social Network Contacts
▸ Sources of Network Data

Suspicious

▸ Social Engineering/Phishing

Symmetric and Skew-Symmetric Matrices

▸ Matrix Algebra, Basics of