A

Accessibility

▸ Analysis and Planning of Urban Networks

Accuracy

▸ Quality of Social Network Data

Actionable Information in Social Networks, Diffusion of

Cindy Hui[1], William A. Wallace[2], Malik Magdon-Ismail[3], and Mark Goldberg[3]
[1]Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA
[2]Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY, USA
[3]Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, USA

Synonyms

Information propagation; Large-scale network; Social relationships

Glossary

Diffusion Spread of an idea, product, or behavior in a social system

Social Ties Connections between individuals on which information is passed

Definition

Information diffusion is the process whereby information spreads through a social system through interactions among its members. Actionable information refers to information that requires the individual members to make a decision or perform an action.

Introduction

Diffusion can be described as "the process by which an innovation is communicated through certain channels over time among the members of a social system" (Rogers 1995). The innovations can be ideas, products, or behaviors but in general can refer to anything that is perceived as new or novel by the individuals of the social system. The innovations spread from originating sources to prospective users and individuals. Diffusion theories and models are used to describe how these innovations may spread and are adopted by the individuals over time. This entry presents an

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

overview of the research on information diffusion followed by a framework that generalizes this research and can be used to position ongoing and proposed diffusion of information research.

Early models stem from communication theories that try to explain the communication patterns that are created by the flow of information among individuals. Early communication models focused mainly on mass media channels and the communications between opinion leaders and followers. One of the early mass communication models is the "hypodermic needle model" or the "magic bullet" theory. This model suggested that mass media has a direct and immediate influence on the behavior of individuals. Later on, the two-step flow model was introduced, which suggested that ideas flow from mass media to opinion leaders and from opinion leaders to the rest of the population. The two-step flow model was introduced in the 1944 study which focused on looking at the effect of mass media on voting behavior during the presidential election campaign (Katz 1957). In this study, researchers found that when they surveyed individuals about influential sources on their voting intentions, personal contacts were mentioned more frequently than mass media, such as radio or newspapers. Information from mass media reaches opinion leaders and influential individuals who pass the original content along with their interpretation of the information to other members of society. The assumption is that the people with most access to the media have a better understanding of the information and explain and diffuse the information to the other members of the population. The two-step flow model set the foundation for the multistep flow theory of communication and study of the diffusion of innovations.

Later diffusion theories also considered the importance of interpersonal channels and personal networks in the diffusion process. Rogers showed that personal networks had a stronger impact than mass media in the diffusion of innovation (Rogers 1995). Granovetter introduced the concept of strength of personal ties and its impact on the access to nonredundant resources in the context of obtaining information on jobs (Granovetter 1983). Researchers began to look at individual behaviors and individuals as decision makers and focus on the social processes that influence individual decision-making.

The diffusion of innovation theory, introduced by Everett Rogers in 1962, describes diffusion as the process where an innovation is spread through a social system by communication channels over time (Rogers 1995). The innovation is considered to be any material, i.e., product, or nonmaterial object, i.e., idea or practice, that is considered new. The theory suggests that there is a decision-making process that occurs when individuals are considering the adoption of the innovation and defines stages of product adoption process: knowledge, persuasion, decision, implementation, and confirmation. The characteristics of the individual along with the attributes of the innovation and the communication channel will influence the adoption process.

The model classifies adopters into five categories: innovators, early adopters, early majority, late majority, and laggards. The theory suggests that the adoption curve follows an S curve, in which a small proportion of individuals initially adopt the innovation, followed by relatively quick adoption by the early and late majority, and then levels off as the laggards finally adopt. The main feature of this theory is the concept that for most individuals in the social network, the decision to adopt the innovation is dependent upon the other individuals in the network. That is, early adopters have a profound effect on the adoption decisions of the later adopters. Recent research utilized the categories of adopters as introduced in Roger's theory to analyze how the adoption process affects the information flow of product recommendation (Song et al. 2006).

An individual's decision to spread an idea or adopt a product is highly affected by social influences and interactions that occur over time. In marketing, the concept of "word of mouth" is commonly used. It builds on the observation that a consumer's decision to accept a new product depends on what they hear from others (Goldenberg et al. 2001). Previous theories of innovation diffusion focused on the studying the diffusion rate and the extent of the diffusion. More recent theories try to incorporate the concept of bandwagons.

Bandwagons occur when individuals adopt an innovation not because of personal preferences, but as a result from observation or pressure from their social contacts (Bikhchandani et al. 1992). This is a form of information cascade which occurs when the individual observes the actions of other individuals and follows the behavior of others, independently of their own information and perspectives. The assumption here is that the individual acts rationally and conforms to what others are doing. The basic model of how cascades occur is described as follows. There is a set of individuals in the system. Each individual decides whether to adopt or reject an idea or behavior. The decisions are made in a sequence where each individual can observe the decisions of those ahead of them. The decision to adopt or reject the idea or behavior is probabilistic based on the decisions of previous individuals.

There are two perspectives to the bandwagon effect that occurs in information cascades (Abrahamson and Rosenkopf 1997). One perspective is rational efficiency. These theories assume that when people adopt an innovation, they provide information on the efficiency and value of the innovation. As more people adopt the innovation, they spread more knowledge about the innovation. This knowledge is made available to the rest of the people and influences potential adopters to adopt the innovation. On the other hand, fad theories suggest that individuals adopt an innovation simply because many others have adopted the innovation, regardless of how efficient or useful the innovation may be.

Granovetter's threshold model is one of the earliest models of collective behavior (Granovetter 1978). The model suggested that each individual has a different "threshold" which represents their capacity to resist social pressure. The threshold is the number of other individuals in the network that have adopted the behavior or innovation before an individual will adopt. The thresholds are distributed over the population based on a probability distribution. This theoretical model asserts that the thresholds are dependent on the context of what is being diffused. The threshold model can be formalized to model many social phenomena, such as the diffusion of innovation, rumors, diseases, strikes, and voting behavior (Granovetter 1978; Rogers 1995).

Granovetter's threshold model has been extended by many researchers, and there are many variations of threshold models that are used for studying collective behavior. Some variations of the model differentiate the social influence from the individual's personal networks and the external influence from the rest of the social system (Delre et al. 2006; Valente 1996). An individual will decide to adopt a behavior or innovation when the proportion of adopters in their personal networks surpasses their threshold. Threshold models can be deterministic or stochastic (Strang and Macy 2001). Thresholds can be assigned depending on the role or characteristic of the individuals, e.g., opinion leaders can have lower thresholds and influence individuals/followers with higher thresholds (Valente 1996). The distribution of thresholds can affect bandwagon dynamics and the effect is greatest when there is a balance of similar and dissimilar thresholds between individuals and their network neighbors (Chiang 2007).

There are two general classes of models in the research literature that are used for modeling the spread of information or influence. In both models, the social network can be represented by a directed graph where each node may be either inactive or active. The linear threshold model and the independent cascade model both serve as a basis for many diffusion models and extensions to these models, developed to study different diffusion processes (Goldenberg et al. 2001; Kempe et al. 2003; Leskovec et al. 2006).

In the linear threshold model (Granovetter 1978; Watts 2002), there is a weight $w_{u,v}$ on the edge between two nodes u and v, which defines a measure of influence. Each node u has a threshold value, which is drawn randomly from a specified probability distribution. This threshold determines how many neighboring nodes have to be activated before the node itself becomes active. If the sum of the weights of all active neighbors exceeds the threshold, then the node will become active. In the independent cascade

model (Goldenberg et al. 2001; Kempe et al. 2003), when a node u becomes active, it has a single chance to influence each of its inactive neighbors v with a given probability of success $p_{u,v}$. The probability is assumed to be independent of the history of other node influences. If the transmission is successful, the neighbor will become active at the next time step. However, regardless of the success, each node can only attempt to influence its neighbor once and cannot make another attempt at a later time step. This process continues until there are no more possible transmissions.

Research on information diffusion is an active research area with work on the diffusion of innovation and technology, viral marketing, spread of political opinions and news in political science, diffusion of information on blogs and the Internet, the spread of computer viruses, and the spread of diseases. Common research questions include studying how different types of information spread, e.g., recommendations and opinions in social networks (Goldenberg et al. 2001; Leskovec et al. 2006; Richardson and Domingos 2002), how structural properties of the network affect the diffusion process, identifying influential nodes either to promote diffusion or detect early outbreaks (Domingos 2005; Eubank et al. 2004; Kempe et al. 2003, 2005; Leskovec et al. 2007), and developing efficient strategies to select targeted nodes to exploit the characteristics of certain network structures, such as hubs in networks with scale-free properties (Duan et al. 2005).

Methodology

This section describes a general diffusion framework that can be used (1) to study how actionable information may spread and (2) to analyze algorithms for promoting or inhibiting such spreads (Hui et al. 2010a, b). Using the framework, we can investigate how structural properties of the network and attributes of the individuals in the network affect the diffusion process and examine how the diffusion process is influenced by the existence of influential nodes, highly connected information hubs, and nodes that bridge information between groups of individuals (Hui et al. 2009a, b). We assume that there is a network defined for some particular context and the network becomes dynamic over time as information flows (Hui et al. 2008). The attributes of the individuals can influence the diffusion process. When individuals interact with each other, certain properties of the individual might change depending on the information that was conveyed. These properties would affect their decision to initiate an action or not. Once the individual has decided on the action, the action can also have an effect on information flow through the network, i.e., individuals could leave the network, could become disconnected but do not leave, and may disrupt the flow of information at later times.

The model incorporates trust between individuals and trust in information sources and propagators by defining a weight on each edge in the network. The weight represents a measure of trust in the information that is passed between the two connected nodes, i.e., the likelihood that a message will be believed as it is passed from one node to another (Kelton et al. 2008). Depending on the direction of information flow between the two nodes, the trust value may be different.

The diffusion model describes (1) what happens to the message as it is propagated, (2) how the nodes process the information they receive, and (3) how the nodes update their properties based on their interactions and the information they received.

Messages are introduced into the network through external source nodes. When a message is initiated by a source, the message is composed of a source-value pair $\{S, V\}$, which stores the identification of the message's original source and the original information value of the message. When a message is passed from one node to another, the information value of the message at the receiver node is a function of the social relationship, i.e., trust, between the sender and recipient, and this value is nonincreasing. If (S, V) is a source-value pair at node a which is propagated to node b, then the source-value pair at node b is $(S, \alpha(a, b) * V)$, where

$0 \leq \alpha(a,b) < 1$ is the propagation loss from a to b. $\alpha(a,b)$ quantifies the trust relationship between nodes a and b.

Nodes merge the information values from all the messages that they receive by first combining information values from messages that originate from the same source (Part **A**) and then merging all information values into a single information fused value at the node (Part **B**).

A. If a source S_i appears in multiple incoming messages with values $V_i^1, V_i^2, \ldots$, the information from this particular source, V_i^*, is fused into the single source-value pair (S_i, V_i^*), where $\max_k V_i^k \leq V_i^* \leq \sum_k V_i^k$. The value V_i^k corresponds to the information value of source i at node k.

B. Suppose that node k has source-value pairs $(S_1^k, V_1^k), (S_2^k, V_2^k), \ldots$. The fused information value at node k is computed as follows:

$$\text{fused}_k = \lambda * \sum_i V_i^k + (1-\lambda) * \max_i V_i^k, \quad (1)$$

where $\lambda \in [0,1]$. The value fused_k is at least the $\max_i V_i^k$ and at most $\sum_i V_i^k$.

After computing the fused information value, the node will determine its state and behavior based on whether the information value exceeds certain thresholds. Each node can be in one of five possible states: uninformed, disbelieved, undecided, believed, and removed/evacuated. Initially, all the nodes are uninformed. When nodes become exposed to the information, they can enter into one of three states: disbelieved, undecided, or believed. Each node has two defined threshold levels, a lower bound, which lies between the disbelieved and uninformed states, and an upper bound, which lies between the undecided and believed states:

$$0 \leq \text{LowerBound} \leq \text{UpperBound} \leq 1 \quad (2)$$

The thresholds determine the boundaries for when the node will acknowledge the message and/or take an action. If the node's combined information value exceeds one of the thresholds, the node will enter a new state. Table 1 summarizes the possible node states along with its corresponding behaviors.

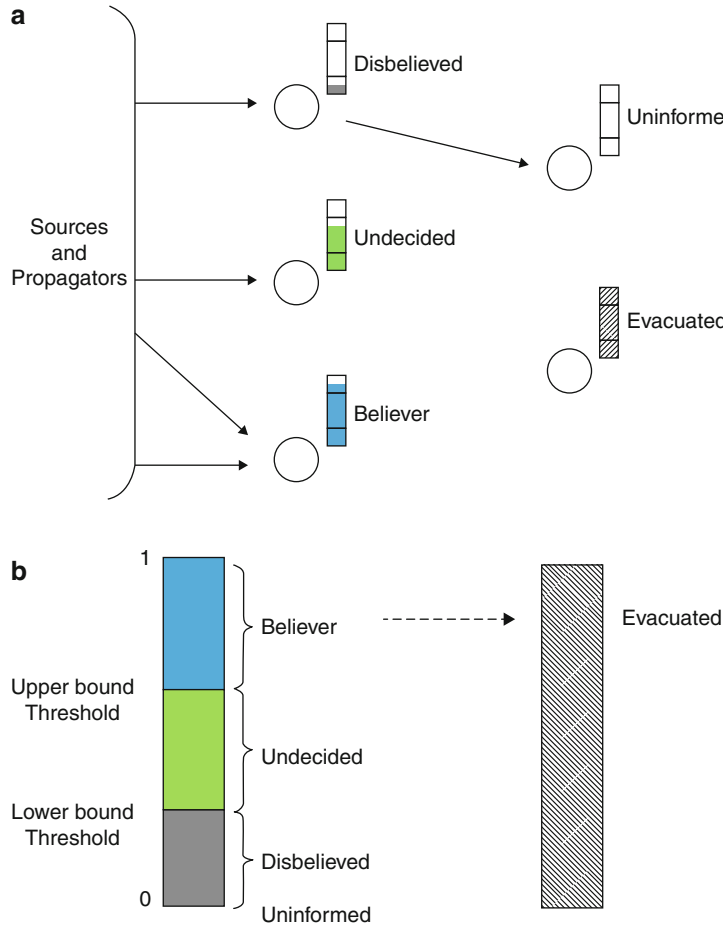
Actionable Information in Social Networks, Diffusion of, Table 1 Description of node states in model

State	Description	Behavior
Uninformed	Node has not received any messages	No action
Disbelieved	Node has received a message but does not believe the message	No action
Undecided	Node has received the message and is uncertain of what to do	Query neighbors in the network
Believed	Node has received the message and believes the value of the message	Spread the message to its neighbors and is removed from the network after x time steps
Removed	Node is no longer in the network	No action

When a nodes tries to query for information or spread a message, there is a probability p_{success} that the message will be received or shared. When an undecided node queries a neighbor who is uninformed or removed/evacuated, the node receives no additional information. When an undecided node queries a disbelieved, undecided, or believed node, the node receives information with probability $p = p_{\text{success}}$.

Key Applications

The framework presented may be used to represent various diffusion scenarios by specifying the parameters to fit the particular context. For example, we can calibrate the model parameters using data from various sources to model the spread of evacuation warnings (Hui et al. 2008, 2009b, 2010b). We can utilize demographic and event data to construct models that



Actionable Information in Social Networks, Diffusion of, Fig. 1 Depiction of message propagation and node states

can be used to investigate questions of interest regarding diffusion in large-scale networks, see Fig. 1. An example of how network properties affect the dissemination of evacuation warnings is provided by (Hui et al. 2009a, 2010a). The modeling approach can be used to study at the household level while simulating large-scale warnings diffusion.

We applied the diffusion model to the context of evacuation warnings in large-scale networks. We used demographic and event data of the San Diego firestorms that occurred in 2007 to simulate the spread of evacuation warnings on a million node social network of households. After configuring the model, we simulate various diffusion scenarios to study how social group structure, distribution of trust, and existence of weak ties affect the dissemination of evacuation warnings (Hui et al. 2009b, 2010a). The procedure is as follows:

1. Construct a social network of households.
2. Script the events of San Diego firestorms.
3. Configure model parameters.
4. Validate the model configurations by obtaining results close to the actual reported number of evacuated households.
5. Simulate the spread of evacuation warnings on the constructed networks.
6. Study how social groups and trust distribution affect the evacuation.

Future Directions

We can expand the framework to incorporate additional features and model various diffusion contexts. The current model captures information-seeking behavior, as a form of seeking confirmation, where the node tries to gather more information from their neighbors in the network. Confirmation behavior generally refers to seeking reinforcement for a decision or action. We can explore other forms of confirmation, such as visual cues, where observing other's actions may increase or decrease their likelihood to act.

The current framework models the situation where the information being diffused requires a decision or an action. The framework can be extended to incorporate the scenario where there is a need to take back the information and instead inform the individuals in the network to not perform the suggested action. The idea would be to impede the diffusion of the previously broadcasted message by sending an abort message at a later time to remove the false information from the network. Exploratory experiments looked at the performance of various strategies for selecting seed nodes for the broadcast of actionable information and abort information. Preliminary experiments demonstrate that there is a measurable trade-off between a fast effective spread of an action and the ability to effectively retract or counter the actionable information (Hui et al. 2011a, b).

Future work focuses on studying algorithms for optimally spreading or impeding spread of information under given network characteristics and developing dynamic strategies for selecting seeds to broadcast information. Currently, nodes are either selected based on the initial social or communication network or predefined at the beginning of the simulation. It would be of interest to develop mechanisms for selecting seed nodes over time that consider network dynamics and changes due to information flow. At the present we do not consider how individuals evaluate opposing information – this could lead to investigating other mechanisms for information fusion. The order in which messages arrive at the node may affect how the node processes the information and determines its state and behaviors. The time between messages may also affect the node's decision-making process. Timing issues such as these could be addressed.

Acknowledgments

This material is based upon work sponsored by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053 at Rensselaer Polytechnic Institute and by the Department of Homeland Security through the Command, Control, and Interoperability Center for Advanced Data Analysis Center of Excellence at Rutgers University. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the US Government.

References

Abrahamson E, Rosenkopf L (1997) Social network effects on the extent of innovation diffusion: a computer simulation. Organ Sci 8(3):289–309

Bikhchandani S, Hirshleifer D, Welch I (1992) A theory of fads, fashion, custom, and cultural change as informational cascades. J Pol Econ 100(5):992–1026

Chiang Y (2007) Birds of moderately different feathers: Bandwagon dynamics and the threshold heterogeneity of network neighbors. J Math Sociol 31(1):47–69

Delre S, Jager W, Janssen M (2006) Diffusion dynamics in small-world networks with heterogeneous consumers. Comput Math Organ Theory 13(2):185–202

Domingos P (2005) Mining social networks for viral marketing. IEEE Intell Syst 20(1):80–82

Duan W, Chen Z, Liu Z, Jin W (2005) Efficient target strategies for contagion in scale-free networks. Phys Rev E 72(2):026133

Eubank S, Guclu H, Kumar V, Marathe M, Srinivasan A, Toroczkai Z, Wang N (2004) Modelling disease outbreaks in realistic urban social networks. Nature 429:180–184

Goldenberg J, Libai B, Muller E (2001) Talk of the network: a complex systems look at the underlying process of word-of-mouth. Mark Lett 12(3):211–223

Granovetter M (1978) Threshold models of collective behavior. Am J Sociol 83(6):1420–1443

Granovetter M (1983) The strength of weak ties: a network theory revisited. Sociol Theory 1:201–233

Hui C, Goldberg M, Magdon-Ismail M, Wallace WA (2008) Micro-simulation of diffusion of warnings. In: Fiedrich F, de Walle BV (eds) Proceedings of the 5th international conference on information systems for crisis response and management (ISCRAM2008), Washington, DC, pp 424–430

Hui C, Goldberg M, Magdon-Ismail M, Wallace WA (2009a) On the weak ties hypothesis in the diffusion of warnings. In: 2009 North American Association for computational social and organizational science annual conference (NAACSOS 2009), Arizona State University, Tempe

Hui C, Magdon-Ismail M, Wallace WA, Goldberg M (2009b) The impact of changes in network structure on the diffusion of warnings. In: Proceedings of the workshop on analysis of dynamic networks at the SIAM international conference on data mining, Sparks

Hui C, Goldberg M, Magdon-Ismail M, Wallace WA (2010a) Agent-based simulation of the diffusion of warnings. In: Agent-directed simulation symposium (ADS'10), as part of the 2010 Spring simulation multi-conference (SpringSim'10), Orlando

Hui C, Goldberg M, Magdon-Ismail M, Wallace WA (2010b) Simulating the diffusion of information: an agent-based modeling approach. Special issue on agent-directed simulation. Int J Agent Technol Syst 2(3):31–46

Hui C, Magdon-Ismail M, Wallace WA, Goldberg M (2011a) Aborting a message flowing through social communities. In: Proceedings of the 3rd IEEE international conference on social computing (SocialCom2011), MIT, Boston, Oct 9–11

Hui C, Magdon-Ismail M, Wallace WA, Goldberg M (2011b) Effectiveness of information retraction. In: IEEE 1st international workshop on network science (NSW 2011), West Point, June 22–24

Katz E (1957) The two-step flow of communication: an up-to-date report of an hypothesis. Public Opin Quart 21(1):61–78

Kelton K, Fleischmann KR, Wallace WA (2008) Trust in digital information. J Am Soc Inform Sci Technol 59(3):363–374

Kempe D, Kleinberg J, Tardos É (2003) Maximizing the spread of influence through a social network. In: Proceedings of the 9th ACM SIGKDD International conference on knowledge discovery and data mining. ACM Press, Washington, DC, USA, pp 137–146

Kempe D, Kleinberg J, Tardos É (2005) Influential nodes in a diffusion model for social networks. In: Proceedings of the 32nd international colloquium on automata, languages and programming (ICALP), Lisboa, Portugal

Leskovec J, Adamic LA, Huberman BA (2006) The dynamics of viral marketing. In: Proceedings of the 7th ACM conference on electronic commerce (EC06). ACM Press, New York, pp 228–237

Leskovec J, Krause A, Guestrin C (2007) Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining, San Jose, pp 420–429

Richardson M, Domingos P (2002) Mining knowledge-sharing sites for viral marketing. In: Proceedings of the 8th ACM SIGKDD international conference on knowledge discovery and data mining, Edmonton, AB, Canada, pp 61–70

Rogers E (1995) Diffusion of innovations. Free Press, New York

Song X, Tseng BL, Lin CY, Sun MT (2006) Personalized recommendation driven by information flow. In: 29th annual international ACM SIGIR conference on research and development in information retrieval, Seattle, WA, pp 509–516

Strang D, Macy MW (2001) In search of excellence: fads, success stories, and adaptive emulation. Am J Sociol 107(1):147–182

Valente TW (1996) Social network thresholds in the diffusion of innovations. Soc Netw 18(1):69–89

Watts DJ (2002) A simple model of global cascades on random networks. Proc Natl Acad Sci 99(9): 5766–5771

Recommended Reading

Albert R, Jeong H, Barabasi A (2000) Error and attack tolerance of complex networks. Nature 406(6794):378–382

Bass F (2004) A new product growth for model consumer durables. Manag Sci 50(Supplement 12):1825–1832

Brown J, Reignen P (1987) Social ties and word-of-mouth referral behaviour. J Consum Res 14(3):350–362

Chen L, Carley K (2004) The impact of countermeasure propagation on the prevalence of computer viruses. IEEE Trans Syst Man Cybern B Cybern 34(2):823–833

Gruhl D, Guha R, Liben-Nowell D, Tomkins A (2004) Information diffusion through blogspace. In: Proceedings of the 13th international conference on World Wide Web. ACM Press, New York, NY, USA, pp 491–501

Hill S, Provost F, Volinsky C (2006) Network-based marketing: identifying likely adopters via consumer networks. Stat Sci 21(2):256–276

Huckfeldt R, Sprague J (1991) Discussant effects on vote choice: intimacy, structure, and interdependence. J Polit 53(1):122–158

Java A, Kolari P, Finin T, Oates T (2006) Modeling the spread of influence on the blogosphere. In: Proceedings of the 15th international conference on World Wide Web, Edinburgh

Leskovec J, Singh A, Kleinberg J (2006) Patterns of influence in a recommendation network. In: Proceedings of the Pacific-Asia conference on knowledge discovery and data mining (PAKDD), Singapore
Macy M (1991) Chains of cooperation: threshold effects in collective action. Am Sociol Rev 56(6): 730–747
Meyers LA, Newman M, Pourbohloul B (2006) Predicting epidemics on directed contact networks. J Theor Biol 240(3):400–418
Morris S (2000) Contagion. Rev Econ Stud 67(1):57–78
Valente T (1995) Network models of the diffusion of innovations. Hampton Press, Cresskill
Wan X, Yang J (2007) Learning information diffusion process on the web. In: Proceedings of the 16th international conference on World Wide Web. ACM Press, New York, pp 1173–1174
Young HP (2003) The diffusion of innovations in social networks. The Economy as an Evolving Complex System, III, Oxford University Press

Activity Modelling

▸ Extracting Individual and Group Behavior from Mobility Data

Actor-Based Models for Longitudinal Networks

Alberto Caimo[1] and Nial Friel[2]
[1]Faculty of Economics, University of Lugano, Lugano, Switzerland
[2]INSIGHT: The National Centre for Big Data Analytics, School of Mathematical Sciences, University College Dublin, Dublin, Ireland

Synonyms

Actor-oriented modelling; Agent-based models; Stochastic actor-based models

Glossary

Actors Nodes of the network graph
Behavior Changing characteristics of actors
Covariates Variables which can depend on the actors (actor covariates) or on pairs of actors (dyadic covariates). They are considered "exogenous" variables in the sense that they are not determined by the stochastic process underlying the model
Dyad Pair of actors of the network
Dyadic Indicator Binary variable indicating the presence or absence of a tie between two actors
Effects Specifications of the objective function
Longitudinal Networks Repeated measures of networks over time
Markov Chain Stochastic process where the probability of future states given the present state does not depend on past states
Method of Moments Statistical estimation method consisting of equating sample moments of a distribution with unobserved theoretic moments in order to get an approximation to the solutions of the likelihood equations
Network Graph representing a relation on the set of actors with binary dyadic indicators which can be regarded as a state changing over time
Network Dynamics Study of longitudinal networks
Objective Function Function which determines probabilistically the dyadic changes made by the actors evaluating all information included in the connectivity structure of the actors
Rate Function Expected number of opportunities for change per unit of time
Ties Relational connections between nodes of a network graph

Introduction

The study of longitudinal networks has become a major topic of interest and dynamic modelling approaches have been pursued in much of social network analysis. Important applications range from friendship networks (see, e.g., Pearson and West 2003; Burk et al. 2007) to

interorganizational networks (see, e.g., Brass et al. 2004). However, many of the classical statistical models proposed have focused mainly on single static network analysis. One of the reasons why network dynamics was not tackled until a couple of decades ago is that the complex dependence structures that characterize networks could not allow an exact inferential calculations and estimation procedures cannot be dealt without computer simulation algorithms.

These powerful tools have allowed researchers to focus to the analysis of the underlying mechanisms that induce the characteristics of network dynamics from the "micro dynamics" such as the individual actor choices to the "macro properties" such as the network connectivity structure. Key research topics concern the structural positions of the actors, their connectivity evolution, belief development, friendship formation, diffusion of innovations, the spread of a particular behavior, etc.

Modelling the dynamics of social networks is therefore of crucial importance, but it is also extremely difficult, due to the temporal dependence, but also since network data, at any given time instance, are not composed of independent observations but each tie variable between two actors is dependent on the presence or absence of ties in the other dyads. Consequently, the network dynamics is greatly affected by the global connectivity structure. For this reason, standard statistical models cannot give an adequate representation of this dependence feature. Various models have been proposed for the statistical analysis of longitudinal social network data, and some earlier reviews were given by Wasserman (1979) and Frank (1991). An actor-oriented approach to this type of modelling was pioneered by Snijders (1996, 2001, 2005) and Snijders and van Duijn (1997) under the assumption of statistical dependence between observations evolving over time according to a continuous time Markov process. These models were originally designed to model the evolution of expressive networks consisting of individuals, but they can provide a general framework for the analysis of many different kinds of relations. Some applications were presented by van de Bunt (1999), de Nooy (2002), Huisman and Steglich (2008), and van Duijn et al. (2003). The actor-based models are a family of statistical models aiming to describe network dynamics according to some typical network dependencies such as reciprocation of ties and transitivity. They represent one of the most prominent classes of models for the analysis of network dynamics as they allow a flexible analysis of the complex social network dependencies among the actors over time. In this context, the network dynamic is assumed to be driven by different effects modelled by network statistics which operate simultaneously. The stochastic process defined by these effects can provide a good representation of the changes of the network connectivity structure over time. The model parameter estimates allow one to understand the strength of the effects included in the model. The actor-based models are flexible as they allow to incorporate a wide variety of network statistics. The main objectives of this approach consist in representing a wide variety of effects or tendencies on network evolution, estimate parameters expressing such tendencies, and test corresponding hypotheses so as to understand the structuring of social networks over time. These effects are various and can be created based on the application. The parameter estimates obtained from the inferential process can be used to simulate network structures compatible with the tendencies observed in the network changes under study. This chapter does not give a review of this literature concerning modelling approaches for longitudinal network data but it focuses only on the class of stochastic actor-based models. In this chapter, we describe the basic features of the actor-based models by providing the theoretical assumptions and methodology requirements needed. We give some basic insights concerning the inferential analysis for the parameters and goodness of fit procedures. Next, we carry out an illustration of the capabilities of these models through their application to an ethnographic study of community structure in a New England monastery by Sampson (1968). A brief discussion on the future directions is given at the end of the chapter.

Stochastic Actor-Based Models for Network Dynamics

Longitudinal social networks represent useful tools for explaining the development over time of many different kinds of social relations (such as friendship, advice, communication) within a group of actors (such as people or organizations). These relations between actors or nodes are by nature subject to change at any time. The individual properties and behavior of the nodes and the similarity characteristics of pairs of nodes can generally affect the topology structure of the network and therefore its evolution. For this reason the major difficulty in the analysis of social networks is that each connected dyad depends on the subgraph connectivity structures of its nodes. Network dynamics can be driven by several different effects such as reciprocity, transitivity and homophily. These tendencies or effects are responsible for the creation and termination of the ties between the actors of the network.

The stochastic actor-based models are a family of models designed to analyze the mechanisms which determines the change and the evolution of social networks by taking into account the strength of a wide variety of these effects described by some model parameters. The availability of statistical procedures for estimating the parameters and testing their significance allows us to understand the contribution and significativity of each single effect. Social networks can be represented by graphs and the ties represent the states of the relation between the nodes. Longitudinal networks are usually represented by panel data collected at different times.

Notation

The relational structure of a social network can be represented by a graph on a set of nodes $N = \{1, \ldots, n\}$ connected by dyadic variables represented as an $n \times n$ adjacency matrix X whose element $\{X_{ij}\}$ represents the relation between node i (sender) and j (receiver). The dyad $\{X_{ij}\}$ can take value 1 or 0 indicating the presence or absence of a tie going from node i to j. Self-ties and valued ties are not considered in this context, so by convention we set $x_{ii} = 0$. Here we use capital letters to denote random variables, and lower case letters to denote the corresponding observations.

Longitudinal dyads can be denoted by $X_{ij}(t)$ so as to indicate that the states are time dependent and are collected in the random adjacency matrix $X(t)$. There can also be other variables that may influence the network and they are regarded as covariates. These can be depending on actors (e.g. the sex of actors) and are denoted by V_i or dyads (e.g. spatial distance between two actors) and are denoted by W_{ij}.

Methodology

Stochastic actor-based models are typically concerned with directed networks whose states have the tendency to endure over time in order to satisfy the requirement of *gradual change*. This property is generally observed in many kinds of social relations such as friendship, trust, and cooperation. In this context, changes in the states of the network are generally assumed to be dependent on the current network states and not on the past ones. This means that all the relevant information for modelling the future state is assumed to be provided by the current state of the network. In statistical terms this means that the social network is a stochastic process with *Markov property*. This assumption is obviously an approximation which can be considered unrealistic in some cases.

Basic Assumptions

- The parameter t is continuous. In most applications, observations of the network are made at some discrete (generally small) number of outcomes of the process $X(t)$ which evolves in continuous time t. Actor-based models interpret this discrete time series of observed networks as the cumulative result of an observed sequence of elementary changes made by the

actors between two consecutive observations. This continuous process is not observed and it is inferred from modelling. An analogous approach was proposed by Coleman (1964).

- $X(t)$ is a Markov process, that is, the conditional probability distribution of future states of the network depends on the past states only as a function of the present states. In other words, the states of the network at time t include all the knowledge for predicting future states of the network at time $t + 1$. This assumption takes into account the tendency of ties to remain in place until some special event happens.
- At any given time t, no more than one dyadic variable $X_{ij}(t)$ can change. This assumption is related to the fact that changes of ties are mutually dependent only because tie changes will depend on the current global connectivity configuration of the network. The process is assumed to be decomposable into sequence of smallest possible changes.
- The process is actor based in the sense that dyadic changes are made by the actors who create or drop a tie on the basis of their covariate attributes and their position in the network. The actor-based process is characterized by two stochastic subprocesses:
  - The *change opportunity process* models the frequency of the dyadic changes made by actors. The *rate function* is denoted by λ and represents the probability of the occurrence of a change in a given small time interval. This frequency rate can be constant or dependent on the nodal covariates or on the current network connectivity structure.
  - The *change determination process* models the choice of the *ego actor* who gets the opportunity to make a dyadic change. The ego actor may create or drop one outgoing tie or make no changes. The probability of the choice is modelled by the *objective function* $f_i(x^0, x, v, w)$ which depends on the current state of the network x^0, the potential new state of the network x and the covariate attributes v (nodal) and w (dyadic).

Model Specifications and Estimation

Specification of the Objective Function

The objective function has a crucial role in actor-based models as it determines the rules of the actor behavior in the network. When an opportunity for actor i occurs at a rate λ_i, the actor can change one of the outgoing dyadic variables $X_{ij}, \ \forall j \in N$ leading to a new state of the network x. The probability of this new state of the network is given by

$$p_i(x|x^0, v, w) = \frac{\exp\{f_i(x^0, x, v, w)\}}{\sum_{x' \in C(x^0)} \exp\{f_i(x', x, v, w)\}}, \tag{1}$$

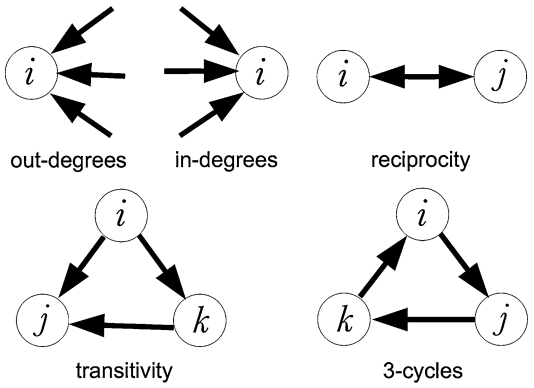
where $C(x^0)$ denotes the set of all possible networks resulting from a change in the current network x^0. The changes can be considered as determinations of new dyadic states of X_{ij} according to a multinomial logistic regression model. Similar to generalized linear models, the objective function is specified as a linear combination of a set of network statistics called *effects* $s_{ki}(x)$ (Fig. 1) which correspond to "tendencies" driving the network dynamics:

$$f_i(\beta, x) = \sum_k \beta_k s_{ki}(x), \tag{2}$$

where β_k are statistical parameters indicating the impact of the corresponding network statistic on the dynamic process: the greater the value of the parameter, the higher the probability of the corresponding network statistics to have an impact on the network dynamics and vice versa if the value of the parameter is negative. The choice of the set of effects to include in a model is generally theory guided and it often depends on the application context. In the following section, we will describe some of most used effects included in the objective function.

Basic Effects

Outdegree-Indegree These effects correspond to the number of outgoing and incoming ties for actor i, respectively. These network statistics can be defined as



Actor-Based Models for Longitudinal Networks, Fig. 1 Some of the most common network effects for stochastic actor-based models

$$s_{Oi}(x) = \sum_j x_{ij},\ s_{Ii}(x) = \sum_j x_{ji},\ i \neq j. \tag{3}$$

They give a measure of the position and popularity of actor i. If the corresponding parameters have positive values, the importance of these effects will tend to increase or stay high over time.

Reciprocity This effect corresponds to the number of reciprocated ties for actor i. The network statistic is defined as

$$s_{Ri}(x) = \sum_j x_{ij} x_{ji}, \qquad i \neq j. \tag{4}$$

It gives a measure of the tendency toward reciprocation of choices referring to mutual tied: if i connects to j, then j connects to i and vice versa. This effect shows significant positive evidence in many kinds of networks such as friendship relational data.

Other degree-based effects may be formulated by taking into account more complicated features of the actor's behavior. For example, actors may have preferences to connect to other actors based on both their own and other's degrees. In many contexts the degrees reflect status of an actor and therefore can play a crucial part in the tie formation and evolution.

Triadic Structures These family of effects incorporate information involving three actors and they can measure two important features of the network such as *transitivity* and *transitivity closure*. Several network statistics can be formulated to analyze transitivity.

The *transitive ties* effect is defined as

$$s_{TEi}(x) = \sum_h x_{ih} \max_j (x_{ij} x_{jh}),\ i \neq j \neq h, \tag{5}$$

and measures the number of configurations involving three actors in which $x_{ij} = 1$, $x_{jh} = 1$, and $x_{hi} = 1$. This effect models the tendency toward generalized exchange.

The *transitive triplets* effect is one of these and it is defined as

$$s_{TTi}(x) = \sum_j x_{ij} \sum_h x_{ih} x_{hj}, \quad i \neq j \neq h, \tag{6}$$

and measures transitivity for actor i by counting the number of actors j for which there is at least one intermediary h forming a transitive triplet.

The *three-cycles* effect is defined as

$$s_{Ci}(x) = \sum_j x_{ij} \sum_h x_{jh} x_{hi}, \qquad i \neq j \neq h, \tag{7}$$

and measures the number of configurations involving three actors in which $x_{ij} = 1$, $x_{jh} = 1$, and $x_{hi} = 1$. This effect models the tendency toward generalized exchange.

Covariate-Based Effects These are effects that take into account the information on the nodes (v_i) or ties (w_{ij}). For example, the *similarity* or *homophily* effect measures whether ties tend to occur more often between actors with similar values of V. The *ego* effect reflects the propensity of the actor to send ties, leading to a correlation between v_i and outdegrees. The *alter* effect models the tendency of the popularity of the actor for receiving ties, leading to a correlation between v_i and indegrees. It is also possible to include dyadic attributes expressing different kinds of features such as

meeting opportunities or spatial propinquity. More formulae of effects can be found in Snijders (2012).

Network and Behavior Coevolution

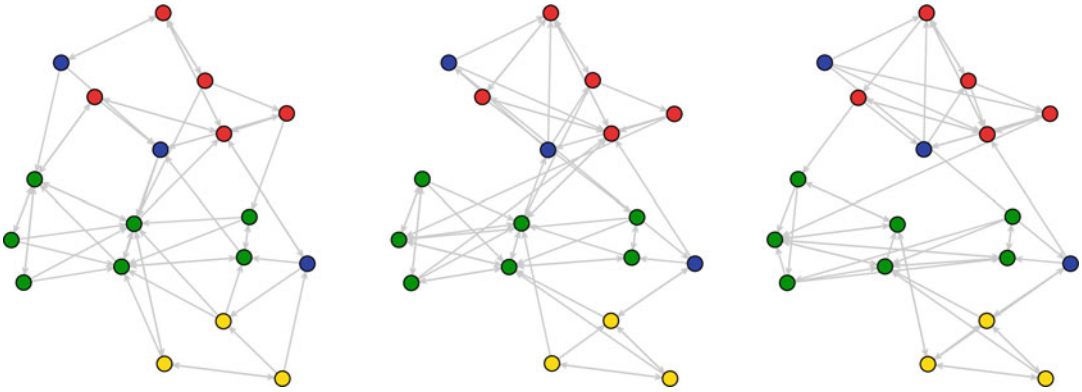
The network structure can have an impact on the behavior and performance of the actors. For example, actors can be influenced by their neighbors because of many different factors such as competition and cooperation. and, in turn, influence other actors. This type of changing attributes is referred to as *behavior*. The vector of attributes for actor i is no longer assumed to be constant over time and it is denoted by $z_i(t)$. The model for the network and behavior coevolution is defined as the stochastic process $(X(t), Z(t))$. Now the structure changes of $X(t)$ will depend on both the current network states $X(t)$ and behavior states $Z(t)$, and, similarly, the behavior changes of $Z(t)$ will depend on both the current network states $X(t)$ and behavior states $Z(t)$. The behavior process is characterized by a rate function λ^Z driving the frequency of changes and objective function f_i^Z which defines the probabilities of changes. The process $(X(t), Z(t))$ follows the assumptions of gradual change over time. At any given moment t no more than one of the set of variables $X_{ij}(t)$ and $Z_{ih}(t)$ can change. This means that there is no direct coordination between changes in ties and changes in behavior and the dependence is made by the influence they have to each other. The objective function defined in Equation 2 can be used to model the *behavior* of the focal actor i or some of his neighbors and on his network position, etc. The so-called shape, influence, and position-dependent effects allow to model the behavior change. It is generally quite challenging to estimate a model which includes both dynamic and behavior effects; for this reason, in many cases, it is advisable first to focus on fitting a good network dynamic model and subsequently include terms for the behavior evolution process.

Data Requirements and Modelling Issues

In order to use stochastic actor-based models for longitudinal networks, the assumptions described above should be plausible from a practical point of view. For example, the total number of changes between two consecutive network observations has to be "large enough" so that the changes can actually provide the necessary information for estimating the parameters. At the same time, this amount of change does not have to be too high so as not to violate the assumption of gradually changing states. One of the implicit assumptions of the model is that the actors of the network have a homogeneous behavior toward change so specific outlying behaviors are not taken into account.

Statistical Estimation

The distribution of the parameter estimates of the parameters β_k in the objective function is approximately normally distributed, and therefore the significance of the parameters can be tested using the t-ratio testing context. Estimation procedures are possible by a variety of simulation-based methods. The method of moments proposed by Snijders (2001) operates by selecting a vector of statistics, one for each parameter coordinate to be estimated, and determining the parameter estimate as the parameter value for which the expected value of this vector of statistics equals the observed value. This estimation procedure is iterative and based on variants of the Robbins-Monro algorithm (Robbins and Monro 1951). The convergence of this algorithm is generally quite good but it can be affected by the initial value. In many cases good starting points are obtained from estimates of simple models. Generally fitting complicated models may affect the convergence and the performance of the estimation algorithm and this may result in obtaining poor estimates. Moreover, network statistics can be highly correlated by definition and this implies that also the parameter estimates can be strongly correlated. For these reasons forward selection procedures are preferred to backward ones. The estimation procedure is based on the principle that the first observed network is considered a starting point of the dynamic process so it is not modelled directly.



Actor-Based Models for Longitudinal Networks, Fig. 2 Sampson monastery network graph at three time points. Colors represent the actors membership covariates: *green* actors belong to the Young Turks, *red* actors belong to the Loyal Opposition, *yellow* actors belong to the Outcasts, and *blue* actors belong to the interstitial group

The parameters β_k of the objective function can be interpreted as weights of the model effects. It is important to check if the estimated model is able to describe the overall features of the stochastic process. Goodness of fit for stochastic network models is generally based on the comparison of a set of data simulated from the estimated model with the observed data (Hunter et al. 2008). For each network statistic, the percentile at which the observed value is located in the distribution of simulated network statistics is used as a test. Obviously it is recommended to consider also effects that are not directly included in the model estimated so as to have an overall picture of the fit of model to the data.

Example: Dynamics of Social Relations in a Monastery

The Sampson monastery network dataset (Sampson 1968) consists of social relations among a set of 18 monk novitiates preparing to enter a monastery in New England. There are three separate adjacency matrices representing liking relations at three points in time (Fig. 2). Based on observations and analyses, the monks can be partitioned into four groups: Young Turks, Loyal Opposition, Outcasts, and an interstitial group. The Loyal Opposition consists of the novices who entered the monastery first. The Young Turks arrived later, in a period of change. They questioned practices in the monastery, which the members of the Loyal Opposition defended. Some novices did not take sides in this debate, so they are labelled "interstitial." The Outcasts are novices who were not accepted in the group.

In this explanatory example we propose two models. In the first model we decided to include only structural network effects: `outdegree` measuring the network graph connection density, `reciprocity` measuring the importance of mutual ties, `transitive triplets` measuring transitivity, and `3 cycles` measuring the tendency toward clustering. The second model includes the previous network effects plus two nodal covariate-based terms which take into account the group membership of the actors: covariate-related tendency for mutual ties (`group similarity x reciprocity`), defined by the number of preferences for mutual ties with actors that have similar values on a certain individual level nodal covariate, and a covariate-related identity (`same group`), defined by the number of ties of actor i to all other actors j's who have exactly the same value on the covariate, in this case, group membership. All the calculations were done using the RSiena package version 1.1–212 (Snijders 2012), and the parameter estimates and standard errors of both models are reported in Table 1.

Actor-Based Models for Longitudinal Networks, Table 1 Parameter estimates with standard errors of the actor-based models used for modelling the Sampson monastery longitudinal networks

Effect	Estimate	S.E.
Model 1		
Network rate function		
`Rate parameter` (period 1)	3.5692	0.7472
`Rate parameter` (period 2)	2.6102	0.4637
Network objective function		
`Outdegree`	–1.4080	0.2131
`Reciprocity`	1.2313	0.2880
`Transitive triplets`	0.2898	0.1350
`3 cycles`	–0.1475	0.2213
Model 2		
Network rate function		
`Rate parameter` (period 1)	3.7309	0.7143
`Rate parameter` (period 2)	2.8616	0.5490
Network objective function		
`Outdegree`	–1.5110	0.2023
`Reciprocity`	0.8109	0.3030
`Transitive triplets`	0.1486	0.1389
`3 cycles`	–0.3796	0.2294
`Group similarity x reciprocity`	0.4154	0.7772
`Same group`	1.2887	0.3253

The analysis of the first model confirms that there is an overall low level of outdegrees and a high level of reciprocity, as indicated by the two significant `outdegree` and `reciprocity` parameter estimates; there is no evidence of transitive closure as indicated by the estimates of `transitive triplets`. In the second model, the covariate-based effect concerning the `group similarity x reciprocity` is significant, meaning that reciprocity is high between actors belonging to the same group. Rate parameters indicate that the liking relations between monks have a peak in the first period between the first two observations and then decrease slightly in the second period. These differences are obviously reflecting the amounts of change observed between two consecutive network observations. It can be concluded that monks tended to like monks belonging to the same group and this tendency tends to reinforce itself over time determining the loss of across-group ties and the creation of within-group ties.

In this simple example, we have considered that dropping a tie is the opposite of creating a new one. However, in many contexts, this is not plausible. In order to take this into account, it is possible to consider another component of the objective function: the *endowment* function. This operates only for the termination of ties and not for their creation.

Tie-Based Approaches

The modelling approach presented in this chapter is actor based in the sense that the dynamic process and the statistical inference are driven by the behavior of each single actor of the network. However, a tie-based version of the longitudinal model was proposed by Snijders (2006) and corresponds to exponential random graph models (ERGMs, see Robins et al. 2007 for a review). These models are best suited for analyzing the global topological properties of static networks who are assumed to be in equilibrium. This approach makes it difficult to infer the development of longitudinal networks when the structural trend represented by tie-based network statistics is not discernible. In other words, the tie-based network statistics used in the exponential random graph modelling context are inadequate to microstructural changes which are not affecting the global topology of the graph. For this reason, the actor-based models represent a richer class of models which do not require the sequence of observed networks to be in equilibrium and model the global features of the evolving network by taking into account its micro changes over time. Moreover, the

computational cost of fitting exponential random graph models can be substantial, particularly for large networks.

Modelling Multiple Networks

The actor-oriented models can also be applied for longitudinal multiple network observations. Suppose that at each time point M different dependent relational structures are observed on the same set of N actors. An example is friendship and trust relations for the same set of individuals. The actor-based models now are defined as

$$Y(t) = (Y_1(t), \ldots, Y_M(t)). \qquad (8)$$

Rate and objective functions can be defined separately for each dependent network. Some effects expressing dependencies between multiple networks based on composition of relations are as follows: *direct dependence* measuring the propensity that actors who are connected by a certain relation tend to be connected by another relation and; *cross-network dependence* measuring the propensity that some relational ties tend to be reciprocated by ties of a different kind.

Future Directions

The area of statistical modelling of network dynamics is of growing in development. This chapter has given an introduction to the flexible family of stochastic actor-based models for analyzing longitudinal networks. These models can be used to formulate and test hypotheses concerning the evolution of networks in order to obtain a useful representation of the dynamic behavior of the network structure by measuring the strength of various effects and estimating the corresponding parameters.

The advantage of this modelling approach is that parameter estimates identify a model which provides an easy interpretation of the effects driving the dynamic change in the network structure that are clear reflections of the patterns and regularities that can be derived from the analysis of static social networks. For the purposes of statistical inference, actor-based models provide an important tool for representing the dependencies between the ties of the network and the behavior of the actors over time.

Several applications have demonstrated the usefulness of these models and have provided a better understanding of the interpretability of the results. In particular actor-based models have been proven to be useful in the context of questions about selection and influence in social relational data. Various modelling extensions have recently been proposed, for example, by Checkley and Steglich (2007) and van de Bunt and Groenewegen (2007). From an inferential point of view, other estimations procured have been recently proposed: maximum likelihood estimation by Snijders et al. (2010) and Bayesian estimation by Koskinen and Snijders (2007). The software SIENA ("simulation investigation for empirical network analysis") (Ripley and Snijders 2011), and its R version RSiena (Snijders 2012), for the statistical analysis of network data provides a very useful tool for practitioners and applied scientists.

Acknowledgments

Nial Friel's research was supported by a Science Foundation Ireland Research Frontiers Program grant, 09/RFP/MTH2199.

Cross-References

▸ Exponential Random Graph Models
▸ Modeling and Analysis of Spatiotemporal Social Networks

References

Brass D, Galaskiewicz J, Greve H, Tsai W (2004) Taking stock of networks and organizations: a multilevel perspective. Acad Manage J 47:795–817

Burk W, Steglich Cr, Snijders Tr (2007) Beyond dyadic interdependence: actor-oriented models for co-evolving social networks and individual behaviors. Int J Behav Dev 31:397–404

Checkley M, Steglich C (2007) Partners in power: job mobility and dynamic deal-making. Eur Manage Rev 4(3):161–171
Coleman J (1964) Introduction to mathematical sociology. The Free Press of Glencoe, New York
de Nooy W (2002) The dynamics of artistic prestige. Poetics 30:147–167
Frank O (1991) Statistical analysis of change in networks. Statistica Neerlandica 45:283–293
Huisman ME, Steglich CEG (2008) Treatment of non-response in longitudinal network data. Soc Netw 30:297–308
Hunter DR, Goodreau SM, Handcock MS (2008) Goodness of fit for social network models. J Am Stat Assoc 103:248–258
Koskinen JH, Snijders TA (2007) Bayesian inference for dynamic social network data. J Stat Plan Inference 137(12):3930–3938
Pearson MA, West P (2003) Drifting smoke rings: social network analysis and Markov processes in a longitudinal study of friendship groups and risk-taking. Connections 25(2):59–76
Ripley R, Snijders T (2011) Manual for SIENA version 4.0. http://www.stats.ox.ac.uk/siena/
Robbins H, Monro S (1951) A stochastic approximation method. Ann Math Stat 22(3):400–407
Robins GL, Pattison PE, Kalish Y, Lusher D (2007) An introduction to exponential random graph (p^*) models for social networks. Soc Netw 29:173–191
Sampson SF (1968) A novitiate in a period of change: an experimental and case study of social relationships. PhD thesis, Cornell University
Snijders TAB (1996) Stochastic actor-oriented dynamic network analysis. J Math Soc 21:149–172
Snijders TAB (2001) The statistical evaluation of social network dynamics. Sociol Methodol 31(1):361–395
Snijders TAB (2005) Models for longitudinal network data. In: Carrington PJ, Scott J, Wasserman S (eds) Models and methods in social network analysis. Cambridge University Press, Cambridge/New York, pp 215–247
Snijders TAB (2006) Statistical methods for network dynamics. In: Luchini SR et al (eds) Proceedings of the XLIII scientific meeting, Italian Statistical Society. CLEUP, Padova, Italy, pp 281–296
Snijders TAB (2012) Siena in R: RSiena. http://www.stats.ox.ac.uk/~snijders/siena/siena_r.htm
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, Koskinen J, Schweinberger M (2010) Maximum likelihood estimation for social network dynamics. Ann Appl Stat 4(2):567–588
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, Groenewegen P (2007) An actor-oriented dynamic network approach: the case of interorganizational network evolution. Organ Res Methods 10(3):463–482
van Duijn M, Zeggelink EPH, Huisman M, Stokman FN, Wasseur FW (2003) Evolution of sociology freshmen into a friendship network. J Math Soc 27:153–191
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

Actor–Network Theory

▸ Networks in Geography

Actor-Oriented Modelling

▸ Actor-Based Models for Longitudinal Networks

Actors, Nodes

▸ Role Discovery

Addition and Subtraction of Matrices

▸ Matrix Algebra, Basics of

Adult Entertainment Industry

▸ Pornography Online

Adversarial Knowledge

▸ Anonymization and De-anonymization of Social Network Data

Advisory Systems

▸ Recommender Systems: Models and Techniques

Agent-Based Models

▸ Actor-Based Models for Longitudinal Networks

Agent, User, Peer

▸ Computational Trust Models

Aggregation Strategies

▸ Group Representation and Profiling

Algebraic Path Problem

▸ Semirings and Matrix Analysis of Networks

Algorithm Evaluation

▸ Benchmarking for Graph Clustering and Partitioning

Alliances

▸ Inter-organizational Networks

Almost Network Data

▸ Sources of Network Data

AI Planning

▸ Web Service Composition

Analysis

▸ GUESS
▸ NodeXL: Simple Network Analysis for Social Media
▸ Social Network Analysis in a Digital Age

Analysis and Mining of Tags, (Micro)Blogs, and Virtual Communities

Lisa Kaati
Department of Information Technology, Uppsala University, Uppsala, Sweden

Synonyms

Social media analysis; Web analysis; Web mining

Glossary

Buzz Monitoring Collection and analysis of the voice of the public
Targeting Analysis of social media with the goal to detect individuals
Alias Internet alias that is used for communication on the Internet
Part of Speech Tagging Marking words in a text according to lexical categories
Syntactic Parsing The process of analyzing a string of symbols according to a grammar
Sentiment Analysis Text analysis techniques used to classify the polarity of written text

Introduction

Analysis and mining of various social media sites has become an important task for many different reasons. By using a variety of state-of-the-art techniques including web crawling, text analysis, computational linguistics, and other algorithms, online content from social media sites can be gathered and analyzed. Analysis of social media can be performed using many different approaches depending on the goal of the analysis and the specific media that is considered. The goal of the analysis can, for example, be to get information about the public opinion regarding an event, a feature, or a brand. The analysis can also be predictive, where the goal is to predict the future. In some cases the goal of the analysis is to detect possible threats towards the security of the society posed by individuals or groups. Another aspect of the analysis can be to obtain information about a social network and analyze the network in order to identify social roles and relations. This information can, for example, be used for marketing or for investigation of criminal networks.

Definition

Analysis and mining of social media refers to the process of gathering data from various social media and analyzing the gathered data. Depending on the goal of the analysis, different methods and techniques can be used. In this work we have divided the analysis into four different categories: buzz monitoring, network analysis, prediction, and targeting. Each category has a different goal for the analysis.

Key Points

In this work we focus on mining and analysis of various forms of social media. The analysis is divided into four different categories that we call buzz monitoring, network analysis, prediction, and targeting. Each category differs in the goals of the analysis.

Historical Background

We spend more and more of our lives on the Internet using various forms of social media. The development of the Web 1.0 into Web 2.0 has led to increased user-generated content, and it allows users to communicate and interact with each other using social media services. Due to the large amount of user-generated content, analyzing social media has become interesting to find out public opinions, to detect trends, and to market new products. The analysis can be done manually, but since the amount of available data is vast, computerized methods are much more efficient and feasible. To be able to computerize the analysis, the information needs to be gathered automatically and analyzed using various text mining techniques, natural language processing, machine learning, statistics, and network algorithms.

Analysis of Social Media

Social media refers to the means of interactions among people in which they create, share, and exchange information and ideas in virtual communities and networks online. There are many different kinds of social media and new forms of social media are continuously being developed and used. Internet forums (also called discussion boards, message boards, Web forums, etc.) are one way of communicating on the Internet. An Internet forum is a web application that is used to publish user-generated content under the form of a discussion. Discussions considering particular subjects are called threads or topics. Internet forums have an important social aspect. Many forums are active for a long period of time and attract a group of users that builds a community. There are a lot of different forums dedicated to very possible aspect of human activity, which allows users to find a forum that suits their interests and needs. Internet forums that are dedicated to products are a source of information about consumers' opinions about products and companies are increasingly exploiting this fact. Another form of social media are blogs (or web logs)

which are online journals consisting of different posts written by one or several authors. The majority of blogs allows readers to leave comments and this interactivity distinguishes blogs from other static websites. A blog usually consists of a set of posts displayed in reverse chronological order with the most recent post appearing first. The most common blog is the personal blog, which can be resembled with an online journal containing reflections on life. Other blogs focus on particular subjects such as politics, fashion, gardening, food, or music. Mining and analyzing blogs might reveal information about how ideas are spread, how trends are set, and public opinions. Microblogs are similar to blogs, but they only allow users to send and read text-based messages of up to 140 characters. The most famous microblogging service is Twitter. Microblogs are sources for analysis regarding prediction of trends or early warnings as well as for getting information about the public opinion. Many forms of social media allow tagging. Tags express keywords or labels, for example, when they are assigned to videos, messages, or photos. Tags can be used to search for content that have a common topic. On Twitter, people use the hashtag symbol before a relevant keyword or phrase in their messages to categorize messages and make them searchable in Twitters search engine. Hashtagged words that become popular can be seen as trends in topics, and by gathering and analyzing tags, it is possible to observe and analyze trends. Most online social media enable users to create and join groups. Users can post messages to the group and also upload shared content to the group. Some groups are moderated and a moderator controls admission and postings to the group. Other groups are unrestricted and allow any member to join and post messages or content. There are several challenges that arise when analyzing social media. One challenge is the large amount of available data that can be collected and processed. Another challenge is the fact that most of the data is user generated and contains misspellings and a specific language with emotion symbols and abbreviations. This makes the analysis process more difficult since it is harder to construct computerized tools and techniques that can analyze this kind of data. Text analysis (or text mining) is an important technique that is used when analyzing social media. Text analysis is used to derive information from unstructured text. The process includes structuring the text, parsing, and deriving patterns from the structured text. Examples include extracting entities from text, recognizing patterns, relation extraction, sentiment analysis, part of speech tagging, syntactic parsing, and other kinds of linguistic analysis. There are a number of text analysis tools on the market, both commercial and open source. In this work we divide the analysis of social media into four different categories: buzz monitoring, network analysis, prediction, and targeting. Each category has different goals with the analysis, and different methods are used to obtain the goal.

Buzz Monitoring

Buzz monitoring is when the voice of the public is collected and analyzed. One example when this is used is when companies analyze social media to gain knowledge about the public opinion regarding their products. The fact that data can be analyzed in real time means that new trends and opinions can be spotted fast. However, monitoring social media in real time is technically challenging. Opinions of individuals and groups are typically expressed as informal communication and hidden in a large amount of irrelevant information arising from all forms of social media. To determine users' opinions regarding a certain topic or products, a technique called sentiment analysis can be used. The basic idea is to decide whether a written text expresses negative or positive opinions regarding a particular topic. One way to do this is to classify the polarity of the written text as positive, negative, or neutral. This has been done on, for example, movie reviews, hotel recommendations, and restaurant reviews. Sentiment analysis can be done automatically using machine learning, statistics, and natural language processing. Using more advanced methods it is also possible to recognize moods and feelings in written text. Examples of feelings that can be of interest are angry, sad, scared, and happy. Social media is being increasingly used during crises.

During ongoing crises large amounts of user-generated content, including tweets, blog posts, and forum messages, are created. This information can be used for detecting that an emergency event has taken place, to understand the scope of a crisis, or to find out details about a crisis. By collecting and processing relevant information from social media, crisis responders have the ability to get more information about the situation and help troubled people. Tagging communication with appropriate hashtags allows users to find relevant information and also to make sure that their own messages can be found. As mentioned earlier buzz monitoring can be used when a company wants to find out consumers' opinion regarding a new product. Instead of doing a traditional market research investigation, discussion boards where products are reviewed and discussed or microblogs such as Twitter can be analyzed. Using sentiment analysis it is possible to find out whether consumers have negative or positive feelings regarding the product. The analysis can be done in real time and is usually more cost efficient than doing a market research investigation.

Network Analysis

Since almost all online social media allow networking, one interesting and challenging problem is analyzing the networks. Social networks are a theoretical model that traditionally is used to study relations between individuals, groups, and organizations (Wasswerman and Faust 1994). Social networks can be used to visualize actors and relations but it can also be of interest to analyze the networks. The actors in a network are usually represented by nodes, while the relation between actors is represented as edges between the nodes. Social networks may be very complex, both due to the fact that they may be large and also due to the fact that there might be different kinds of actors and relations in a network. Networks can be constructed using different kinds of relations. In most online social networks, the most common relation is the friend relation, that is, two users are friends. This kind of relation is a mutual relation. In Twitter the relations are somewhat different since you can "follow" another user. Following is not mutual, you are free to follow any user and they do not have to approve or follow back. The following relation is not similar to a friend relation since it does not reflect off-line relationship. To represent a network with relations that are not mutual requires edges to have a direction. When constructing networks from Internet forums, information such as the fact that two users write in the same thread or about the same subject can be used as relations. It is also possible to use relations that describe if one user has cited other users. In the blogosphere many virtual communities have emerged. Networks describing these communities can be created using membership or subscription linkages as relations. There are mainly two categories for analysis of social networks. The first category computes a numerical value for each actor in the network; the other category uses clustering methods to group actors that are similar according to some definition. Centrality is one of the most used measurements in social network analysis. Centrality is a measure of how central a person is in a network. The most commonly used centrality measures are degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality. Degree centrality is a measurement on the number of direct links to other actors in the network. Actors with a high degree centrality are assumed to be important for the network since they are independent of other actors that reach great parts of the network. Closeness centrality can be described as a measure on how close the other actors are. An actor may be important if it is relatively close to the rest of the actors in the network. Betweenness centrality is a measurement that describes how important a person is as a link between different networks. Eigenvector centrality is a measurement on how central an actor is in a network with regard to the global structure of the network. The different centrality measures are commonly used to identify different kinds of key actors in a network. Centrality measurements are usually not enough to analyze a social network. To find the most influential persons in a network, the structure of the network needs to be analyzed. Identifying influential persons is usually done by analyzing information about their position or role in a group

as well as their relations to other groups. Clustering algorithms are used to create groups (or clusters) in a network where each cluster consists of the actors that are related to each other in some sense. There are several different clustering algorithms that divide a network into clusters by maximizing the number of edges within a cluster. By constructing groups in a network, it is possible to target actors that are likely to be influential persons in the social network. Companies that are interested in marketing a new product are often interested in finding the most suitable individuals that can be targeted with the right information and promotions in a cost-effective way.

Prediction

Social media can be used to predict trends or detect early warnings for certain events. Typically, emerging events, breaking news, and general topics that attract the attention of the public drive trends. Predictions are usually done by monitoring social media and detect small shifts in opinions or emerging topics. By analyzing these shifts carefully, trends can be predicted in real time. One of the major problems with detecting trends is that it is difficult to separate what is actually going to be a trend and what is just an anomaly. The difficulty lies in the prediction since it is hard to foresee if something is going to be a trend or if it will disappear as quickly as it was detected. Mining and analyzing tags is one way to get information about emerging trends in topics. Knowledge of emerging trends is particularly important to individuals and companies who are charged with monitoring a particular field or business. Information on how trends spread in a network as well as information about trends and shifts in public opinion can be used for marketing and business purposes. Analysis of messages from Twitter has been used to forecast box-office revenues for movies (Asur and Huberman 2010), and prediction of spikes in book sales by analyzing blogs is done in Gruhl et al. (2005).

Targeting

Analysis of social media can also be used for targeting of individuals. The basic idea with targeting is to automatically identify messages that reveal threats, certain knowledge, plans, or possibly even psychological behavior markers. These messages can be identified and may be subject for more detailed investigations. One problem that arises when analyzing online social media with the aim of detecting individuals or groups is the problem that users may communicate using several different aliases. The use of several aliases becomes problematic when analyzing messages since it is the combined information reveled in all messages written by one user that is interesting. The problem of users that have multiple aliases is addressed in Novak et al. (2004), Narayanan et al. (2012), Chen et al. (2004), and Dahlin et al. (2012). One way to overcome this problem is to analyze the writing style of different users and try to find similarities that indicate that the same Internet user can actually write messages written by different aliases. This kind of analysis is called stylometric analysis (Zheng et al. 2006). Stylometric analysis uses various features from texts, such as the frequency of specific words, lexical features, parts-of-speech tags, and syntactic features, to identify the author of a text. Several algorithms and features for stylometry-based author identification have been proposed throughout the literature (Stamatatos 2009; Abbasi and Chen 2008; Juola 2006). Even if it was possible to detect Internet users that communicate using different aliases, it is still (in general) impossible to find the physical person behind an alias. With existing technology it is easy for a user to disguise him or her using various forms of anonymization techniques. An example when targeting can be used is when detecting individuals that might pose a threat to society and requires further investigation. It could, for example, be school shooters (Veijalainen et al. 2010) who reveal their plans before they take action in a discussion board or terrorists who leak information about a planned terror attack (Brynielsson et al. 2012). Targeting individuals using techniques such as author identification poses a severe threat towards privacy and personal integrity. Before using these kinds of methods, these issues should be considered further.

Key Applications

The ideas presented here may be used to obtain information about activities of individual users, analyze communities, detect new trends and shifts in the public opinion, or predict the future of certain events. Gathering information on how people communicate regarding particular products can be helpful when designing marketing and advertising campaigns. Analysis of social media is cost effective compared to market research investigations and it can be done in real time.

Future Directions

New forms of online social media are constantly being developed, and more and more people use social media in their daily life to communicate. Mining and analysis of various forms of social media has become an important task, and resent research shows many possibilities where mining and analysis of social media can be useful. One problem that arises when analyzing social media is ethical and integrity-related issues. Future research in this area will not only focus on developing new methods and techniques for analyzing social media, it also needs to consider ethical and integrity issues.

Acknowledgments

This research was financially supported by Vinnova through the Vinnmer programme.

Cross-References

- ▸ Classical Algorithms for Social Network Analysis: Future and Current Trends
- ▸ Mapping Online Social Media Networks
- ▸ Network Data Collected via the Web
- ▸ Sentiment Analysis in Social Media
- ▸ Twitter Microblog Sentiment Analysis
- ▸ User Sentiment and Opinion Analysis

References

Abbasi A, Chen H (2008) Writeprints: a stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Trans Inf Syst 26(2):7:1–7:29

Asur S, Huberman BA (2010) Predicting the future with social media. In: Proceedings of the 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology – volume 01, WI-IAT'10, Toronto. IEEE Computer Society, Washington, DC, pp 492–499

Brynielsson J, Horndahl A, Johansson F, Kaati L, Mårtenson C, Svenson P (2012) Analysis of weak signals for detecting lone wolf terrorists. In: Proceedings of the European intelligence and security informatics conference 2012, Odense, pp 197–204

Chen HC, Goldberg MK, Magdon-Ismail M (2004) Identifying multi-ID users in open forums. In: Intelligence and security informatics. Springer, Berlin/New York, pp 176–186

Dahlin J, Johansson F, Kaati L, Mårtenson C, Svenson P (2012) Combining entity matching techniques for detecting extremist behavior on discussion boards. In: ASONAM, Istanbul, pp 850–857

Gruhl D, Guha R, Kumar R, Novak J, Tomkins A (2005) The predictive power of online chatter. In: Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining, KDD'05, Chicago. ACM, New York, pp 78–87

Juola P (2006) Authorship attribution. Found Trends Inf Retr 1(3):233–334

Narayanan A, Paskov H, Gong N, Bethencourt J, Stefanov E, Shin E, Song D (2012) On the feasibility of internet-scale author identification. In: 2012 IEEE symposium on security and privacy (SP), San Francisco, pp 300–314

Novak J, Raghavan P, Tomkins A (2004) Anti-aliasing on the web. In: Proceedings of the 13th international conference on world wide web, WWW'04, New York. ACM, New York, pp 30–39

Stamatatos E (2009) A survey of modern authorship attribution methods. J Am Soc Inf Sci Technol 60(3): 538–556

Veijalainen J, Semenov A, Kyppö J (2010) Tracing potential school shooters in the digital sphere. In: ISA, Miyazaki, pp 163–178

Wasserman S, Faust K (1994) Social network analysis: methods and applications. Cambridge University Press, Cambridge/New York

Zheng R, Li J, Chen H, Huang Z (2006) A framework for authorship identification of online messages: writing-style features and classification techniques. J Am Soc Inf Sci Technol 57(3):378–393

Analysis and Planning of Urban Networks

Andres Sevtsuk
City Form Lab, Singapore University of Technology and Design, Singapore

Synonyms

Accessibility; Centrality urban design; City planning; GIS; Spatial networks; Urban form

Glossary

Urban Form The physical pattern of urban infrastructure and buildings

Land Use Pattern The spatial distribution of human activities accommodated within urban form

Built Environment A combination of urban form and land-use mix of an area

Accessibility Property of a location that describes the ease with which the location can be accessed from surrounding urban form and land-use attractions

Centrality Refers to metrics that describe how centrally an event is located in a spatial network (see ▶ Centrality Measures)

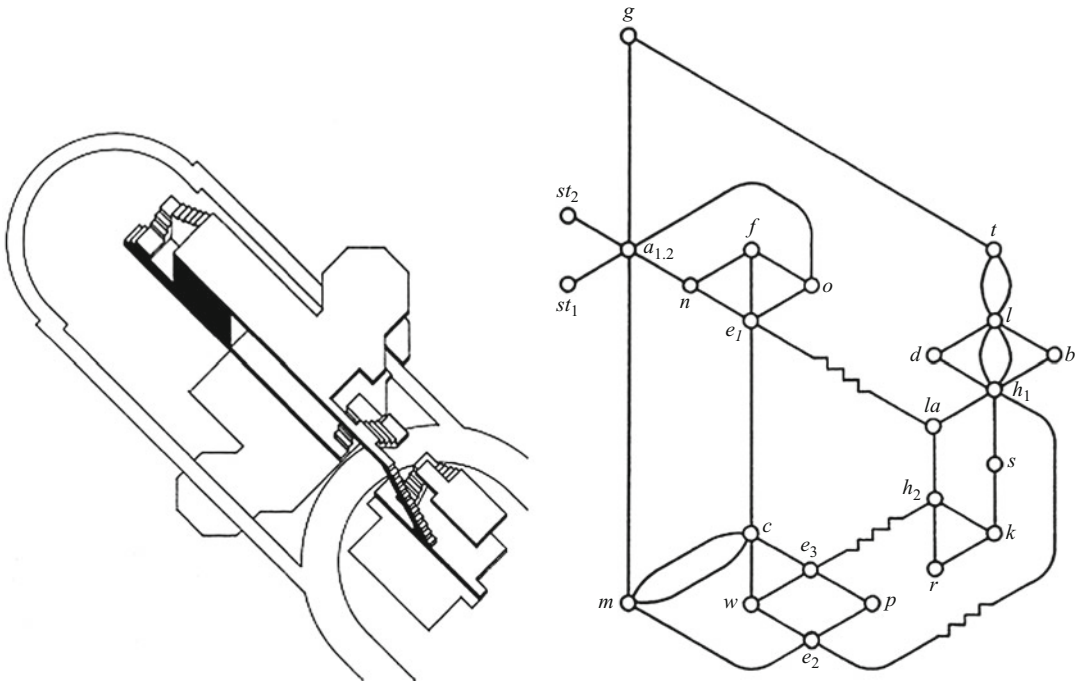
Introduction

Network analysis concepts have been used in the design and planning of cities for several decades. Until recently, however, they were common in only highly specialized applications – disaster planning problems, critical facilities' location problems, and costly utility and transportation infrastructure design problems. Efforts to apply network analyses to the design of ordinary buildings, public spaces, and urban districts go back to the 1960s, but only in the recent decade have the necessary tools and data for their widespread use become available to architects and planners.

Most work in urban network analysis has relied on methods that were originally developed for social networks. There are, however, important adjustments in applying social network analysis methods on urban space. In social networks, for instance, connections between network elements are generally described topologically – as degrees of separation between people in a network, for example – where geography and geometry of relationships have little importance. Scholars of the built environment, on the other hand, are more often interested in precise geographical relationships of a spatial network, where distances, angles, and travel times are critical to describing adjacencies and proximities between places. Second, whereas in social networks the weighting of network elements – people – according to personal characteristics has until recently been rare, weighting is often critical in spatial network studies. A tree-lined street with small single-family homes has a different effect on a neighborhood than a street lined with high-rise office buildings. These particularities have led researchers to customize both the representation of urban networks and the metrics applied thereon.

Historical Background

The spread of graph theory among planners in the latter part of the twentieth century was catalyzed by the appearance of numerous applied graph theory publications after the Second World War (Berge 1962; Harary 1969). Spatial applications of graphs were quickly adopted in transportation research, where the precedent for applying graph measures to large-scale road

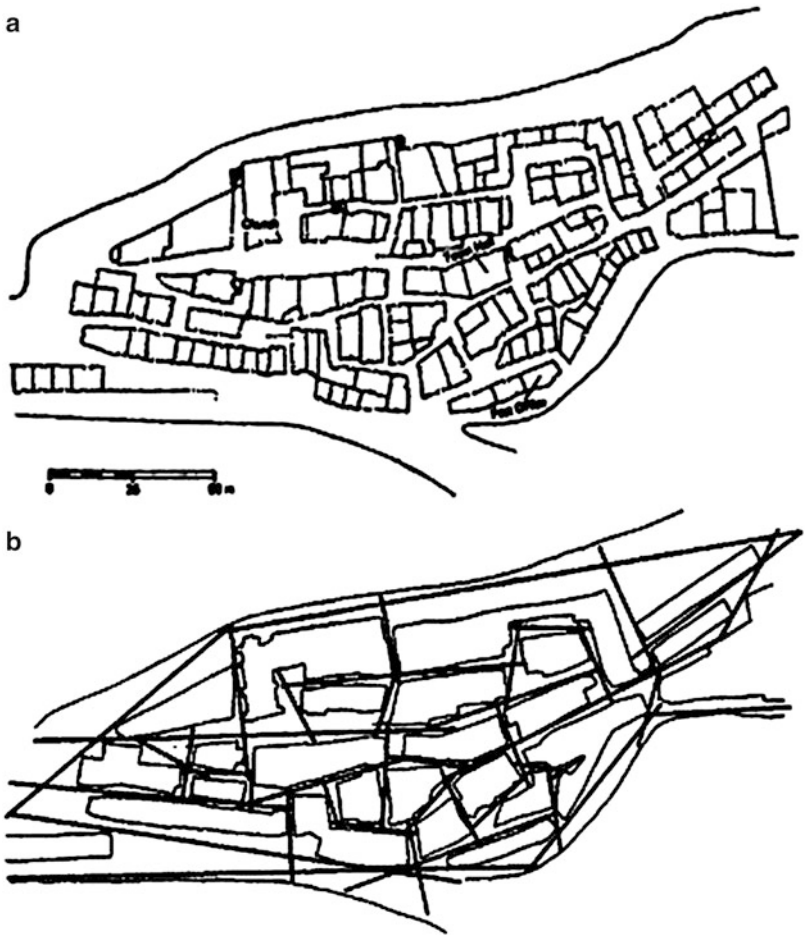


Analysis and Planning of Urban Networks, Fig. 1 Adjacency graph for Frank Lloyd Wright's Aline Devin House (Source: March and Steadman 1971: 259–261)

and rail networks was first established (Garrison 1960; Garrison and Marble 1962; Kansky 1963; Haggett and Chorley 1969). Architects soon also adopted graph representation for the study of building plans (Levin 1964; Casalania and Rittel 1967; Rittel 1970; March and Steadman 1971), typically representing each room by a node and the availability of a direct circulation connection between two rooms by a link (Fig. 1). Representing buildings with graphs opened up new opportunities for distinguishing common layouts of architectural plans using graph indices, some of which are discussed below (Tabor 1976).

Hillier and Hanson applied the representation and analytic tools of graphs to street networks establishing the now well-known Space Syntax methodology (Hillier and Hanson 1984; Hillier 1996). Over the last three decades, their work has argued that the spatial configuration of street networks is related to diverse social phenomena including the rates of pedestrian flow, the geography of crime, and the distribution of business establishments.

Hillier and Hanson have chosen to represent streets not with centerlines, as in most transportation studies, but rather with *axial lines*. Axial lines are defined as the fewest and longest lines of sight that can be drawn through the open street spaces of a study area (Hillier and Hanson 1984). This approach has led to some criticism, since the specification of axial lines is subjective (there is more than one solution) and poorly applicable to sparsely built-up streets (Ratti 2004). Unlike typical transportation applications of graph theory, Space Syntax researchers have also adopted a so-called *dual* graph representation, where streets are represented as nodes and intersections as edges. Since most graph theory indices have been designed to focus on the properties of nodes (e.g., in social networks, nodes can represent people), this inverted form of graph representation allows the Space Syntax analysis to focus on streets (axial lines). Whereas "degree centrality" in social networks indicates how many direct links (e.g., kinship ties or acquaintances) connect to a node of interest (e.g., a person), an analogous



Analysis and Planning of Urban Networks, Fig. 2 (**a**) Plan drawing of Gassin, a hill town in Southern France. (**b**) Axial lines overlaid on its street network (Source: Hillier and Hanson 1984)

measure in Space Syntax describes the number of neighboring axial lines that intersect with a particular axial line of interest.

Though useful for centering the analysis on streets, the dual representation also introduces a well-known problem to the Space Syntax methodology. If streets are represented as nodes, then both long and short streets alike reduce to dimensionless points, thus effectively eliminating metric distance from the analysis. Space Syntax researchers address this problem by measuring travel from one line to another across the graph in topological terms, using the count of lines traversed (i.e., degrees of separation) as a metric of proximity. This metric, commonly referred to as *depth*, is central to most Space Syntax analysis. It is used as a kind of distance measure, which represents the minimum number of axial lines needed to go from an origin to any other axial line in the network. The depth measure leads to another central metric in Space Syntax literature: *integration* (Hillier 1996). The integration measure is simply a relative description of each axial line's depth with respect to all other axial lines in the graph (Fig. 2). It is obtained by repeating the depth measure from each line to all other lines in the system and normalizing the obtained sums for each line by the total number of lines in the graph. In mainstream network analysis terms, the *integration* analysis is analogous to the *Closeness* metric, with the difference that distance is being calculated on the basis of topological turns instead of metric units. If *integration* is computed with a radius of only one turn (also referred to as one *step* in Space Syntax literature), then the result simply

shows how many axial lines intersect with a given line of interest, analogous to the familiar *degree* centrality of nodes in graph theory.

Several other approaches to graph analysis of street networks have appeared in the recent years. Among those, Porta and Xie have implemented a number of spatial graph indices that rely on *primal* representation of spatial networks (Porta et al. 2005; Xie and Levinson 2007). A number of freely accessible software tools have been developed to operationalize spatial network analyses including the Axwoman toolbox (Jiang et al. 1999), the SANET toolbox (Okabe and Shiode 2001; Okabe and Sugihara 2012), the Urban Network Analysis Toolbox (Sevtsuk and Mekonnen 2012), and other custom-built applications for GIS (Miller and Wu 2000; Peponis et al. 2008). In the following we will primarily rely on the approach introduced by Sevtsuk and Mekonnen (2012), which allows us to describe urban spaces of multiple morphologies and scales using a representational framework that is common in transportation studies.

Representational Framework

Most spatial network studies to date have represented networks using two types of network elements – nodes and edges. In the case of urban street networks, edges typically represent street segments, and nodes the junctions where two or more edges intersect (Porta et al. 2005). As already discussed, the Space Syntax approach inverts these elements. The Urban Network Analysis (UNA) framework (Sevtsuk and Mekonnen 2012; Sevtsuk 2010) has recently introduced two important modifications to this framework. First, it adds buildings (or other location instances, such as land parcels, transit stations) to the representation, adopting a tripartite system that consists of three basic elements: edges, representing paths along which travellers can navigate; nodes, representing the intersections where two or more edges intersect; and buildings, representing the locations where traffic from streets enters into indoor environments or vice versa. The unit of analysis thus becomes a building, enabling the different graph indices to be computed separately for each building. Should the analyst wish to compute the graph centrality measures for nodes of the network instead of buildings, then the nodes themselves can be used as inputs instead of buildings. This allows a user to account for both uneven building densities and land use patterns throughout the network, neither of which are addressed in most previous urban network analysis methods. The UNA representation assumes that each building connects to a street segment (edge) that lies closest to it along the shortest perpendicular connection. This network representation framework is illustrated in Fig. 3. The left side of the figure presents a fragment of Harvard Square in Cambridge MA, in plan drawing. The same plan drawing is shown in graph form on the right.

If the spatial configuration of the environment under study cannot be represented in a two-dimensional graph – as may be the case if the network contains underpasses, overpasses, or three-dimensional circulation routes inside buildings – then a similar graph can also be represented three dimensionally, using vertical z-axis values on each of the network elements.

The three-element representation of spatial networks is well suited for mapping urban and regional networks of various typologies and scales. At the finest architectural scale, the third network element can be used to represent individual establishments or rooms within buildings (Fig. 4). Each establishment, room, or floor of a building can be described with an appropriate weight that captures the quantity or quality of activities it houses. At a larger scale, units of analysis can instead represent whole buildings, allowing the user to weigh the analysis by substantive attributes of each building (Fig. 3). At an even larger scale, the units of analysis can become whole city blocks that contain multiple buildings or zip codes that include multiple blocks. The third network element thus offers a flexible container for analyzing the urban built environment at various scales with consistent methods.

It is important to aggregate the units of analysis at substantively justified levels. Numerous studies have shown that the choice of aggregation can itself affect analysis results. This issue, which has become known as the Modifiable Areal Unit Problem or MAUP in literature

Analysis and Planning of Urban Networks, Fig. 3 *Top*: plan drawing of Harvard Square in Cambridge, MA. *Bottom*: graph representation of the same plan (Source: author)

Analysis and Planning of Urban Networks, Fig. 4 The network of urban spaces aggregated to individual business establishments around the Bugis Junction in Singapore. Each establishment is marked by a point at its entrance. Drawing by City Form Lab (Onur Ekmekci and Farre Nixon)

(Openshaw 1984), is defined as "a *problem arising from the imposition of artificial units of spatial reporting on continuous geographical phenomenon resulting in the generation of artificial spatial patterns"* (Heywood 1998). In order to avoid MAUP issues, it is important to favor behaviorally justified choices of aggregation to ad hoc choices of aggregation. One sensible way of aggregating the units of analysis in urban networks is to choose aggregation levels according to *spatial control* boundaries (Habraken and Teicher 1998). Individual rooms in a building, for instance, have a separate control structure from the building as a whole – changing access to the room does not affect access to the entire building. But changing access to the building does affect access to all rooms within the building. It is often convenient to graphically mark the third network elements at the entrances of spatial control boundaries (e.g., doors) of the units of analysis (e.g., business establishments), as shown in Fig. 3. Just like rooms in a building, individual business establishments can have autonomous spatial control boundaries, as can floors in a building, buildings in a block, and blocks in a district. It can be practical to keep track of different aggregation levels by adding an extra digit to the location ID at each successive aggregation level (e.g., building ID = 26; floor ID = 26-3; room ID = 26-3-1). Aggregation allows each of these elements to be both a parent and a child to other elements. Using spatial control boundaries as a basis for aggregation avoids the hazard of biasing

analysis results to arbitrary data groupings and makes the collected data useful for cross-scalar analysis and the results easy to explain to professionals of different disciplines.

Metrics

We can broadly distinguish two types of network analysis indices on urban networks. The first type – inter-network indices – captures the properties of a spatial graph or subgraph as a whole (March 1976; Rodrigue et al. 2006). Their results become meaningful if compared to other networks. The second type – intra-network indices – characterizes the relative relationship of each network element (e.g., node or building) to other surrounding elements in that network.

Inter-network Indices

Inter-network indices can be used to analyze the overall properties of an area's spatial network. These indices are most commonly applied to traditional two-element networks consisting of nodes and edges, where nodes can represent street intersections and edges street segments (Tabor 1976). But they can be equally applied to the circulation layouts of buildings or to three-element networks that contain other network instances. In this case, each building or other network instance can be represented as a node that is connected to the street network via a link to the nearest street segment.

The *Gamma Index* illustrates the extent to which a spatial network resembles a fully connected *diamond graph*, where each node is directly connected to every other node in the graph (Fig. 5). It is calculated as follows:

$$Gamma\ Index = e/[(v^2 - v)/2]$$

where e is the number of edges and v is the number of vertices in the graph. The higher the index, the higher the internal connectivity of the observed street network and the more directly a traveller can commute between intersections in the street network.

The *Cyclomatic Number* forms another index, which shows the availability of alternative, rather than unique routes between nodes in the network. A *cycle* is analogous to a *hole* in a fishnet, and the index is defined as follows:

$$Cyclomatic\ Number = e - v + g$$

where g the number of connected components in the network (Fig. 5).

In urban street layouts, a grid of four blocks surrounded by streets on all sides produces four cycles (Fig. 6). A *cul-de-sac* network, on the other hand, has no cycles at all and is referred to as a *tree* graph. Tree graphs always have $v - 1$ edges, and therefore only a single shortest path is available between any pair of nodes in a tree (Fig. 7). From an urban design perspective, this means that tree networks, which are commonly seen in suburban settings, require the least asphalt to connect a given set of locations. But they also favor hierarchical patterns of organization and limit choice in travel paths.

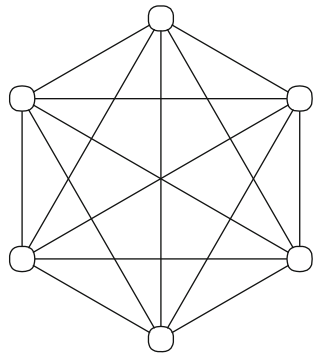
The maximum number of cycles for a given number of vertices is calculated as maximum possible edges in a graph minus edges in a tree graph with the same number of vertices:

$$Max.\ Cycles = [(v^2 - v)/2] - (v - 1)$$

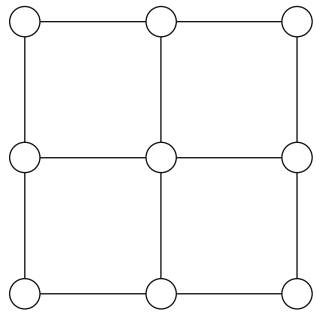
The *Redundancy Index* shows the vulnerability of the network to divisions. The index is defined as a ratio between the number of observed cycles and the number of maximally possible cycles:

$$Redundancy\ Index = (e - v + g)/[[(v^2 - v)/2] - v + 1]$$

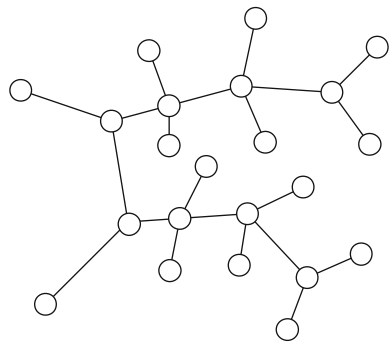
When the *Redundancy Index* is zero, then the network is one of several trees; when the *Redundancy Index* is one, then the network is totally connected (e.g., diamond). The index can be used to study how volatile a street network is to ruptures that may be caused by natural disasters, such as floods and mudslides. Different indices describe different properties of a network, and a careful, hypothesis-driven choice of indices can



Analysis and Planning of Urban Networks, Fig. 5 Diamond graph with 6 nodes and 15 edges



Analysis and Planning of Urban Networks, Fig. 6 An urban grid of four city blocks, where e = 12; v = 9. *Gamma Index* = 1/3; *Cyclomatic Number* = 4; *Maximum Cycles* = 28; *Redundancy Index* = 1/7



Analysis and Planning of Urban Networks, Fig. 7 A "tree" network, where e = 21, v = 22. *Gamma Index* ≈ 0.09; *Cyclomatic Number* = 0; *Maximum Cycles* = 210; *Redundancy Index* = 0

lead to a complementary set of metrics that provide a holistic description of the urban area under study (see Network Representations of Complex Data 00012).

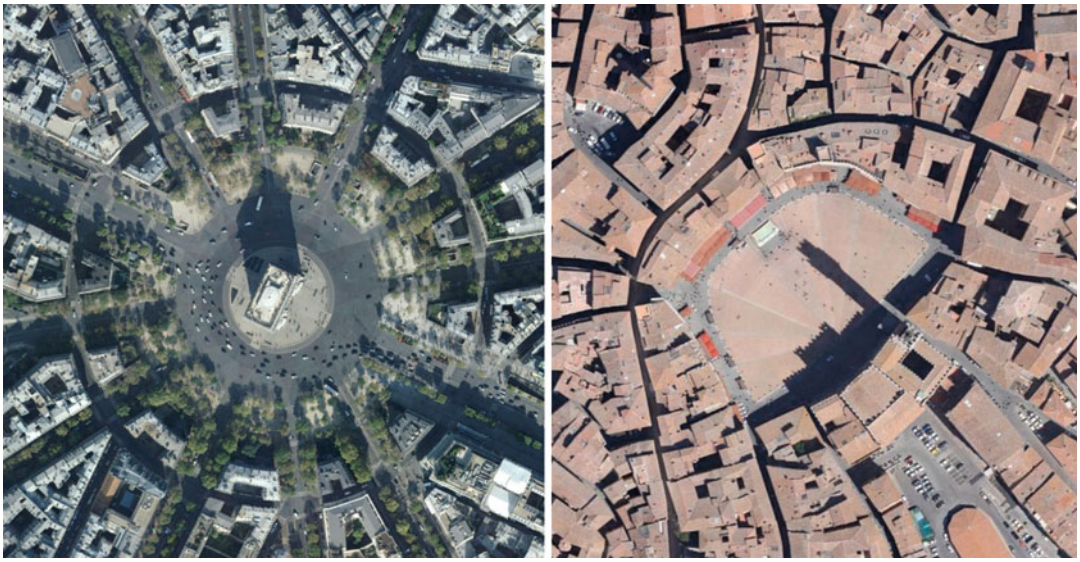
Intra-network Indices

Intra-network indices describe the relative importance of individual nodes or buildings in a network. These indices allow us to compare how well different elements of the same network are connected to the rest of the network. In the following we briefly outline six indices that are readily applicable to the analysis of urban infrastructure networks. The open-source Urban Network Analysis Toolbox allows an interested user to implement most of these indices in ArcGIS (Sevtsuk and Mekonnen 2012).

The *degree* centrality of a street intersection indicates the number of street segments that intersect at the given node. The vast majority of intersections in cities have three or four intersecting street segments. A typical intersection at the corner of an avenue and a street in the Manhattan grid, for instance, has four intersecting edges. But the maximum number of intersecting streets can be much higher. The Charles de Gaulle Etoile (a.k.a. Arc de Triomphe) intersection in Paris forms an atypically connected node, where 12 radial streets collide. One of the most admired public spaces of Europe – the Campo di Siena in Italy (Fig. 8) – also draws foot traffic from 12 paths that enter the Campo. But even nodes with five or six streets tend to stand out as exceptional places in urban circulation networks. Such atypical intersections have been found to be memorable in people's mental maps of a city (Lynch 1960; Mohsenin 2011).

While the degree centrality captures the connectivity of a node to its immediately adjacent street segments, a number of centrality indices have also been designed to describe a location's connectivity to a set of destinations within a larger access radius.

The *Reach* index describes the total number of destinations that can be reached within a given network radius from any street intersection or building. The reach centrality $Reach^r[i]$ of a node i in a graph G at a search radius r describes the number of other nodes in G that are reachable from i at a shortest path distance of at most r. It is defined as follows:



Analysis and Planning of Urban Networks, Fig. 8 *Left*: plan of Charles de Gaulle Etoile traffic square in Paris. *Right*: plan of the pedestrian Campo di Siena in Italy. The degree centrality of both squares is 12 (Source: Google Maps)

$$Reach^r[i] = \sum_{j \in V(G)-\{i\}; d[i,j] \le r} W[j]$$

where $[i, j]$ is the shortest path distance between nodes i and j in G, and $W[j]$ is the weight of a destination node j. The weights can represent any numeric attribute of the destination buildings – their size, the number of employees they contain, the number of residents they accommodate, etc. Using weights allows the analyst to compute how many of such attributes (e.g., residents, jobs) can be reached from each building within a given network radius. The choice of input network and search radius r allow the user to model the index from the perspective of different transportation modes (e.g., walking, biking, driving). The measure may be interpreted as an alternative to areal density measures (e.g., households per acre or jobs per square kilometer). It accounts for opportunities that are reachable along the actual street network as perceived by a pedestrian, bike, or vehicle, producing a unique result for each origin location.

Whereas the *Reach* measure simply counts the number of destinations around each building within a given search radius (optionally weighted by building attributes), the *Gravity* measure additionally factors in the spatial impedance required to reach each of the destinations (Hansen 1959). The *Gravity* index, $Gravity^r[i]$ of a node i in graph G at a radius r, postulates that centrality is inversely proportional to the shortest path distance between i and each of the other nodes in G that are reachable from i within a geodesic distance r. It is defined as follows:

$$Gravity^r[i] = \sum_{j \in V(G)-\{i\}; d[i,j] \le r} \frac{W[j]}{e^{\beta \cdot d[i,j]}}$$

where β is the exponent that controls the effect of distance decay on each shortest path between i and j and $W[j]$ is the weight of a particular destination j that is reachable from i within the radius threshold r. If the buildings in G are weighted, then the *Gravity* measure is proportional to the weight of each of the other buildings that can be reached within the given search radius.

The exponent β in the *Gravity* index controls the shape of the distance decay function, that is, how strongly the distance between i and its neighboring destinations j affects the result. The specification of β should thus be set according to

the mode of travel assumed in the analysis (e.g., walking, cycling, driving), as well as the units of distance measurement. An empirical study of pedestrian trips to convenience stores in Oakland, CA, by Handy and Niemeier (1997) has suggested that for walking distances, measured in minutes, β is approximately 0.1813. The *Gravity* index offers a powerful measure that combines the number of destinations, the attractiveness of the destinations, and the travel costs of reaching them into a single value.

The Reach and *Gravity* indices describe how conveniently each location can be accessed from a set of surrounding locations. For some purposes, however, it may be more important to estimate the ease with which a location can be accessed *en route* while travelling between other locations. Newspaper kiosks, for instance, might find it less desirable to locate at places that are closest to people's homes or jobs and more desirable at places where people tend to pass by while travelling between other destinations. The potential of passersby at different locations of a spatial network can be estimated using a *Betweenness* measure (Freeman 1977).

The *Betweenness* centrality, $Betweenness^r[i]$, of a building i in graph G counts the number of times i lies on shortest paths between pairs of other reachable buildings in G that lie within the network radius r. If more than one shortest path is found between two buildings, as is frequently the case in a rectangular grid of streets, then each of the equidistant paths is given equal weight such that the weights sum to unity. *Betweenness*, in the context of spatial networks, is thus defined as follows:

$$Betweenness^r[i] = \sum_{j,k \in V(G)-\{i\}; d[j,k] \le r} \frac{n_{jk}[i]}{n_{jk}} \cdot W[j]$$

where n_{jk} is the number of shortest paths from building j to building k in G and $n_{jk}[i]$ is the number of these paths that pass through i, with j and k lying within the network radius r from i, and $W[j]$ is the weight of a particular destination j. If the analysis is weighted by demographics of a certain type in the surrounding buildings for instance, the *Betweenness* centrality can capture the potential number of passersby for that particular demographic at building i.

Figure 9 illustrates the *Betweenness* measure applied on two common types of urban layouts: the grid and the cul-de-sac subdivision. For illustration purposes, both layouts have the same number of buildings. The analysis shows that the peak *Betweenness* values are twice as high in the cul-de-sac plan than in the grid, with the range of values also much wider in the former. Since the grid offers multiple routes between any pair of locations, not all paths need to pass a particular link or building. The *Betweenness* values are more equal and distributed in the grid, producing a lesser spatial hierarchy between different locations.

The *Closeness* centrality of a building i is defined as the inverse of the total distance required to reach from i to all surrounding destinations j within the given access radius r (Sabidussi 1966):

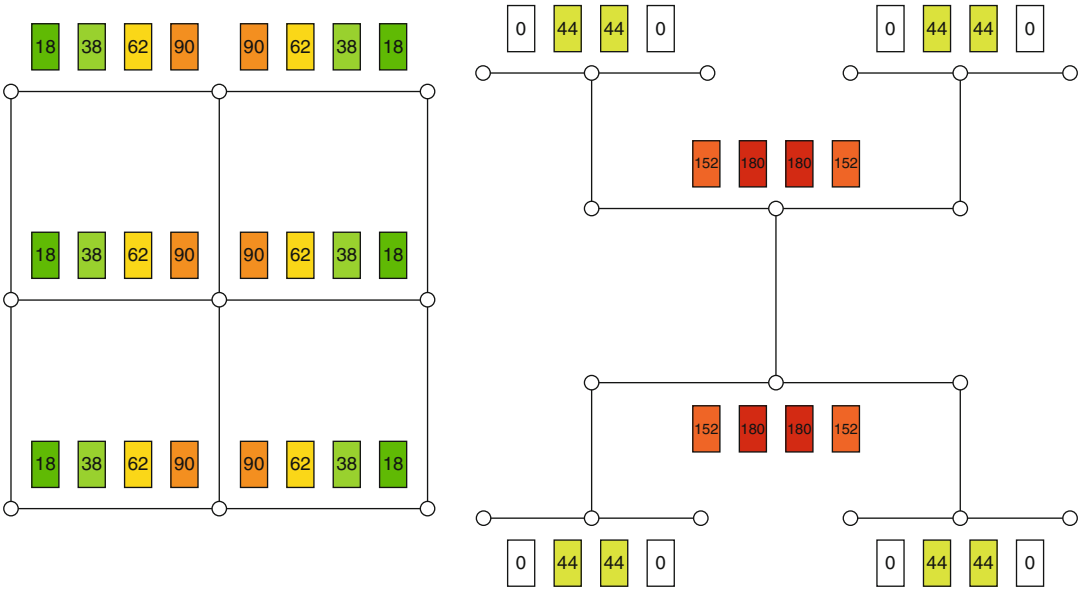
$$Closeness^r[i] = \frac{1}{\sum_{j \in V(G)-\{i\}, d[i,j] \le r} (d[i,j] \cdot W[j])}$$

The *Closeness* measure illustrates how close each of these locations is to all other surrounding locations within a given network access radius. The index is therefore best suited for analyzing the relative proximity of a set of locations to surrounding resources in a city.

Finally, the *Straightness* centrality $Straightness^r[i]$ of a building i estimates how closely the shortest network distances between i and its surrounding buildings j that are reachable within radius r resemble straight Euclidean distances (Vragovic et al. 2005; Porta et al. 2005):

$$Straightness[i]^r = \sum_{j \in V(G)-\{i\}; d[i,j] \le r} \frac{\delta[i,j]}{d[i,j]} \cdot W[j]$$

where $\delta[i,j]$ is the as-a-crow-flies distance between buildings i and j, $d[i,j]$ the shortest



Analysis and Planning of Urban Networks, Fig. 9 Comparison of the *Betweenness* measure on two layouts – the grid and the tree – with the same number of buildings (24) and the same linear length of streets. Values on buildings indicate how many times a building is passed (Source: author)

network distance between the same buildings, and $W[j]$ the weight of destination j. As a ratio between the Euclidian distance and the geodesic distance, *Straightness* can only be estimated if the units of impedance are in linear distance (e.g., miles), not time (e.g., minutes). The index can be used to estimate how directly a set of residential apartment buildings connect to their nearest bus stops, for instance.

Areas of Application

As urban networks typically represent the built environment in a way that ties activities and spaces together with a connecting web of travel paths, urban network analysis is well suited to social, economic, and transportation questions that involve accessibility and the movement of people, goods, and information in a city. Unlike idealized Euclidian distances, proximity measurements on networks model the environment with the constraints imposed by the geometry of streets and other man-made or natural boundaries that closely approximate the human experience of navigating a city. When analyzing how close or accessible households are to public transit stations, for instance, conventional Euclidian buffer distances risk connecting places that are in reality separated by highways, water bodies, or fences, resulting in overestimations of accessibility. An accurate network representation of the same area can integrate these constraints and provide a more reliable estimate.

Research has shown that urban network analysis measures can be useful predictors for a number of interesting phenomena. Porta and Sevtsuk have used network analysis to study the location patterns of retail commerce on urban street networks and found that retailers typically locate in places that are more "between" surrounding destinations (Porta et al. 2009), close to jobs, transit stations, and built density (Sevtsuk 2010). This research has developed a new direction for urban economics, moving from economic geography to economic geometry. A similar approach can be applied to studying the distribution of other land uses, land values, rents, commercial revenue, and other spatial economic indicators.

A number of researchers have also studied how the geometry and topology of urban street networks affect pedestrian and motorized traffic patterns (Hillier et al. 1987; Hillier 1996). Using a third network element to describe buildings or establishments, shown above, allows an analyst to investigate how an addition of a new building or business affects the accessibility of existing buildings or businesses. It can be used to show, for instance, how much additional foot traffic a newly proposed development could add to individual street segments or where opportunities might be created for new businesses.

Future Trends

From the point of view of planners, a central shortcoming of current network analysis methods is their inability to generate alternative geometric configurations. Most spatial network analysis approaches are good at quantifying existing geometric networks, but poor at suggesting alternative solutions that could improve an existing network with respect to given constraints (König et al. 2012; Raford 2010). Thus an analyst will typically test before and after scenarios of a proposed urban intervention and use network analysis to illustrate the improvements achieved by the proposed changes. It is less clear how the results could be used to improve the design. This is not a shortcoming of network analysis per se, but of all spatial analysis methods in general. Spatial analysis tools do not design, they analyze existing designs.

A potentially promising development on this issue has recently appeared in procedural urban models (Parish and Muller 2001; Vanegas et al. 2009). These models generate geometric configurations of urban form on a fly, based on a set of input parameters and are capable of illustrating the geometric results so as to achieve a more desirable combination of the input parameters. If one of the parameters of a residential development model is to achieve particular levels of network accessibility to surrounding jobs, for instance, then the model could iterate through a large number of possible geometric configurations and search for solutions that yield the highest home-to-jobs accessibility. This looks like a promising research direction in the coming years. But since all good urban design involves hundreds, if not thousands of variables that a designer intuitively balances, it is unlikely that any truly sophisticated generative models will appear in the near future.

Cross-References

▸ Centrality Measures
▸ Network Representations of Complex Data

References

Berge C (1962) The theory of graphs and its applications. Methuen, London

Casalania V, Rittel H (1967) Generating floor plans from adjacency matrices. MIT, Cambridge

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

Garrison WL (1960) Connectivity of the interstate highway system, papers and proceedings, vol 6. Regional Science Associations, Philadelphia, pp 121–137

Garrison WL, Marble DF (1962) The structure of transportation networks, no. 62-II. U.S. Army transportation command technical report, pp 73–78

Habraken NJ, Teicher J (1998) The structure of the ordinary: form and control in the built environment. MIT, Cambridge

Handy S, Niemeier AD (1997) Measuring accessibility: an exploration of issues and alternatives. Environ Plan A 29:1175–1194

Hansen WG (1959) How accessibility shapes land use. J Am Plan Assoc 25(2):73–76

Haggett P, Chorley JC (1969) Network analysis in geography. Butler & Tanner Ltd, London

Harary F (1969) Graph theory. Addison-Wesley, Reading

Heywood (1998) Introduction to geographic analysis. Addison Wesley Longman, New York

Hillier B (1996) Space is the machine: a configurational theory of architecture. Cambridge University Press, Cambridge/New York

Hillier B, Hanson J (1984) The social logic of space. Cambridge University Press, Cambridge

Hillier B, Burdett R, Peponis J, Penn A (1987) Creating life: or, does architecture determine anything? Archit Comport Archit Behav 3(3):233–250

Jiang B, Claramunt C, Batty M (1999) Geometric accessibility and geographic information: extending desktop GIS to Space Syntax. Comput Environ Urban Syst 23(2):127–146

Kansky KJ (1963) Structure of transportation networks: relationships between network geometry and regional characteristics. University of Chicago, Chicago
König R, Schneider S, Bielik M (2012) The parametric exploration of spatial properties – coupling parametric geometry modeling and the graph-based spatial analysis of urban street networks. In: Proceedings of the symposium on simulation for architecture and urban design, Orlando, pp 123–129
Levin PH (1964) The use of graphs to decide the optimum layout of buildings. Archit J 7:809–815
Lynch K (1960) The image of the city. MIT, Cambridge, p 194
March L (1976) The architecture of form. Cambridge University Press, Cambridge
March L, Steadman P (1971) The geometry of environment: an introduction to spatial organization in design. RIBA Publications, London, p 360
Miller HJ, Wu Y-H (2000) GIS software for measuring space-time accessibility in transportation planning and analysis. Geoinformatica 4(2):141–159
Mohsenin M (2011) The impact of urban geometry on cognitive maps. Massachusetts Institute of Technology, SMArchS thesis
Okabe A, Shiode S (2001) SANET: a toolbox for spatial analysis on a network. J Geogr Anal 38(1):57–66
Okabe A, Sugihara K (2012) Spatial analysis along networks: statistical and computational methods. Statistics in practice. Wiley, Hoboken, p 296
Openshaw S (1984) The modifiable areal unit problem. Geo Books, Norwick
Parish Y, Muller P (2001) Procedural modeling of cities. In: ACM SIGGRAPH, Los Angeles
Peponis J, Bafna S, Zhang Z (2008) Connectivity of streets: reach and directional distance. Environ Plan B Plan Des 35:881–901
Porta S, Crucitti P, Latora V (2005) The network analysis of urban streets: a primal approach. Environ Plan B 35(5):705–725
Porta S, Strano E, Iacoviello V, Messora R, Latora V, Cardillo A, Wang F, Scellato S (2009) Street centrality and densities of retail and services in Bologna, Italy. Environ Plan B Plan Des 36:450–465
Raford N (2010) Social and technical challenges to the adoption of Space Syntax methodologies as a planning support system (PSS) in American urban design. In: Proceedings of the 7th international Space Syntax symposium, Stockholm, Sweden, pp 090:1–090–12
Ratti C (2004) Space Syntax: some inconsistencies. Environ Plan B- Plan Des 31:487–499
Rittel H (1970) Theories of cell configuration: Emerging methods in environmental design and planning. In: Moore GT (ed) MIT, Cambridge
Rodrigue J-P, Comtois C, Slack B (2006) The geography of transport systems. Routledge, Abingdon/New York
Sabidussi G (1966) The centrality index of a graph. Psychmetrika 31:581–603
Sevtsuk A (2010) Path and place: a study of urban geometry and retail activity in Cambridge and Somerville, MA. MIT, Cambridge
Sevtsuk A, Mekonnen M (2012) Urban network analysis toolbox. Int J Geomat Spat Anal 22(2):287–305
Tabor P (1976) Networks distances and routes. The architecture of form. In: March L (ed) MIT, Cambridge, pp 366–367
Vanegas C, Garcia-Dorado I, Aliaga DG, Benes B, Waddell P (2012) Inverse Design of Urban Procedural Models, unpublished
Vragovic I, Louis E, Diaz-Guilera A (2005) Efficiency of information transfer in regular and complex networks. Phys Rev E 71:026122
Xie F, Levinson D (2007) Measuring the structure of road networks. Geogr Anal 39(3):336–356

Analysis and Visualization of Dynamic Networks

Faraz Zaidi[1], Chris Muelder[2], and Arnaud Sallaberry[3]
[1]College of Computing and Information Sciences, Karachi Institute of Economics and Technology (KIET), Karachi, Pakistan
[2]Computer Science Department, University of California at Davis, Davis, CA, USA
[3]Computer Science Department, LIRMM – Université Montpellier 3, Montpellier, France

Synonyms

Evolving networks or graphs; Graph mining; Information visualization; Longitudinal network analysis; Network or graph visualization; Temporal networks or graphs; Time-stamped graphs; Time-varying graphs; Visual analytics; Visual data mining

Glossary

Network or a Graph A mathematical structure to represent objects and their interactions. Objects are represented by nodes or vertices (often denoted by a set V), and interactions are represented by links or edges (often denoted by a set E). Mathematically, a graph G is defined as a tuple $G(V, E)$. Mathematicians use the term graph, whereas scientists from

other disciplines usually use the term network to refer to the same concept. Throughout this text, we use these terms interchangeably

Social Network A network where objects represent people and their interactions represent some sort of relationship among people. For example, two individuals may be connected to each other if they have studied at the same school or play for the same football team

Clusters A group of nodes (representing objects) that are densely connected to each other and sparsely connected to other nodes in the network. Formally, a clustering of a static graph $G = (V, E)$ is defined by a set C of subsets of V: $C = \{c_1, c_2, \ldots, c_l\}$ such that $V = c_1 \cup c_2 \cup \ldots \cup c_l$

Small-World Network A graph with two characteristic properties. The average path length, i.e., the number of nodes needed to traverse from one node to another on average, is low, as compared to an equivalent size random graph. The second characteristic is the high transitivity among nodes, i.e., many sets of three nodes are connected to each other with three vertices

Scale-Free Network A graph whose degree distribution follows a power law where the power law coefficient is usually between [2, 3]. In other words, this means that most nodes have only a few connections (low degree) and few nodes have many connections (high degree) in the network

Definition

Network science has emerged as an interdisciplinary field of study to model many physical and real-world systems. A network, although consists of only a set of nodes and edges, is a very powerful structure to represent a wide variety of systems such as people related through social relations, airports related through flights, and computers connected through the Internet. The world we live in increasingly becomes a collection of such networks, and scientists from various disciplines are combining efforts to develop sound theories and fundamental concepts governing this new and exciting field of study.

A subpart of network science which has attracted a lot of attention, practical applications, and research interest is social network analysis (SNA). SNA focuses on the relationships and the interconnected behavior of different entities such as objects, people, and organizations. A more specialized area associated with SNA is the study of dynamic behavior of networks formally known as dynamic network analysis (DNA). More often than not, social networks (and even networks in general) exhibit structural changes over time, i.e., addition or deletion of nodes or edges. For example, social relations are a function of time as they appear and disappear with respect to social events taking place in the society. Similarly studying air transportation networks or data packet traffic over the Internet, all are examples of dynamic networks where time plays an important role in the evolution, analysis, and understanding of the entire network.

Research on dynamic networks ranges from analytical and algorithmic models studying their evolution processes to the study of specialized topics such as the role of individuals in epidemics of infectious diseases. The study of these networks helps predict the behavior of many natural and real-world systems, how individuals or groups of individuals influence, control, and mold the shape of the entire network.

Along with methods to analyze networks, visualization has become an integral component of this research area (Freeman 2004). The visual representation of these networks exploits the human cognitive and perceptual capabilities to build hypothesis, validate theories, and create a better understanding of networks around us.

Formally, we can define a dynamic network as a network which undergoes structural changes over time. The analysis and visualization of these networks is the study of algorithms, methods, tools, and techniques which helps us understand these networks and extract applicable knowledge from them. The study of dynamic networks forms a new and cross-disciplinary area of study with research opportunities and applications in many diverse fields.

Introduction

Networks are all around us. Wherever we look, we can find an interconnected web of objects participating in an interdependent system. Traditional science studies a system of similar objects by modeling an individual and then generalizing its behavior for the entire system. Network science is different, as its bases are in the holistic nature of systems that cannot be modeled as individuals but only through the study and analysis of interconnectedness. This holistic approach helps in understanding the complex behavior of real-world systems and allows us to predict and forecast how these networks evolve over time. The dynamic behavior of many of these networks has made this field both interesting and challenging for researchers. Subsequently this focus area of networks has been dubbed dynamic network analysis (DNA).

Dynamic network analysis is the study of change occurring in networks with the passage of time (Moody et al. 2005). These dynamics occur due to processes either inherent to the system or through some external change processes forced into the system. An example of an inherent process for a social network would be aging which results in death and thus change in social structure, simply due to the passage of time. In contrast, an external change process would be the development of a new airport in a city or the introduction of a technologically advanced hub in a communication network which induces structural change in these types of networks. The granularity of time with which change occurs depends on the system being modeled and the context in which it is studied. For example, the study of air traffic networks is meaningful in minutes and hours, whereas the data packet traffic in communication networks is usually studied in micro- and nanoseconds.

People have only recently started to realize the potential of dynamic analysis and look closely at how it can be useful. Often, the temporal dimension has been completely ignored in earlier studies. The reasons for this fairly late thrust in this area (as compared to early work such as Sampson's monastery study in 1969) are the immaturity of the field itself and the unavailability of reliable and substantial time-stamped datasets. Along with other datasets, the exponential growth of online social networks has boosted the availability of large, accurate, and complete network datasets that include a temporal dimension and have facilitated research in this area. This availability of dynamic network datasets has helped scientists move from theoretical framework to a more applied approach where they can actually formulate problems and test and validate their solutions pertaining to dynamic network analysis.

DNA is applied in scenarios where temporal ordering of events, temporal duration of processes, or the rate of change in interactions taking place in a network is important. For example, researchers have studied the network of romantic relationships among school students to understand and model how sexually transmitted diseases spread in a network. A static view of the entire network fails to exhibit relationship changes over a time period. A partial view of the network of any specific day consists of mostly disconnected pairs of nodes as only a negligible number of students are involved in multiple relationships at the same time (Moody et al. 2005). For these kinds of networks, questions such as when and in what order did the relationships change and how frequently an individual changes a partner all become significant, thus enforcing the necessity of methods for the analysis of dynamic networks. Other examples of networks with dynamic behavior commonly found include data communication networks, social communication networks, transportation networks, ecological networks, and biological networks.

DNA differs from classical SNA as it studies the change phenomena rather than group-level coherence, attribute-level features, and optimization issues. Questions related to the efficiency and stability of the network become more pertinent than the structural properties and organization of the network. As suggested by M. Trier, dynamic network analysis is "more concerned with the activity of actors and their relationships, as compared to static analysis which concentrates on

structural issues of a network" (Trier 2008). An important point to note is that dynamic network analysis does not replace static network analysis, but it addresses a different set of issues altogether and has been found very useful in many application areas.

Visualization of networks is an integral component of the field of SNA (Freeman 2004) and especially DNA as it helps to visually comprehend the dynamics taking place in a network. Real-world systems can exhibit temporal dynamics in a number of ways. Visualizing these dynamics is an active area of research, and several different methods have been proposed to solve a variety of real-world problems. Consequently, methods that go beyond traditional SNA are needed to cater the growing requirements of DNA. Network analysis and visualization tools such as Pajek, UCINET, Gephi, Tulip, and SoNIA (Social Network Image Animator) are all contributing towards the development of methods to support dynamic network analysis and visualization along with traditional SNA.

Historical Background

The history of static SNA dates back to 1933 from the field of sociology. Moreno (1934), a psychologist, used a sociogram (which is now known as network) to represent what the interpersonal structure of a group of people looks like. He studied an epidemic of runaways at a school where he concluded that it was the structural positioning of students in the social network that caused students to runaway. Early work in network analysis with temporal data available is by Newcomb (1961) who for a period of 16 weeks studied the acquaintance process and possible developing friendships in a group of male students who shared the same house. The first attempt to use visual analysis of dynamic networks was by Samuel F. Sampson as part of his Ph.D. thesis in 1968 (Sampson 1968). Famously known as Sampson's monastery study, Sampson studied the evolution process of community structures in a New England monastery by taking several snapshots of the same network at different time intervals for visual analysis.

Recent interest in the field of SNA and subsequently DNA was sparked by the milestone papers of Watts and Strogatz (1998) and Barabási and Albert (1999) where they studied the properties of small-world and scale-free networks. These discoveries were made in very diverse fields such as collaboration network of film actors and neural network of the worm Caenorhabditis elegans, attracting researchers from other disciplines and triggering new research horizons across many disciplines.

The study of dynamic networks presents a new and challenging area of research with its own set of problems. Domain experts relish methods that can help them visualize changes in a network resulting in better understanding and new discoveries from their network data. Although the field is in its infancy, it promises a lot for the future as more researchers focus on dynamic network analysis and visualization. Specialized journals (IEEE Network, IEEE Transactions on Visualization and Computer Graphics, Computer Graphics Forum, Information Visualization, Social Network Analysis and Mining, Journal of Social Structure) and conferences (such as IEEE VAST, IEEE InfoVis, EuroVis, IEEE/ACM ASONAM, IEEE SocialCom2011, IEEE PacificVis and Graph Drawing) are fast becoming standard platforms to share knowledge and encourage collaborative research in this area.

Dynamic Networks: Analysis and Visualization

Research in the area of dynamic networks can be broadly categorized into two partially overlapping categories: the "analytics" part and the "visualization" part. These two categories are overlapping because many analytical methods are used in conjunction with visualization techniques to facilitate the overall process of interactive extraction of knowledge. Similarly many visualization techniques are used as a preprocessing step in the analysis phase to complement analytical methods. Before we discuss these two

categories, we will give a mathematical definition of dynamic graphs followed by their general classification methods.

Mathematical Formulation for Dynamic Networks

A dynamic graph can be defined formally as an agglomerated graph $G = (V, E)$ and an ordered sequence of subgraphs $S = \{G_1 = (V_1, E_1), G_2 = (V_2, E_2), \ldots, G_k = (V_k, E_k)\}$ where each G_t is the subgraph of G at time t where t can be a specific time or it can be a time period. $V, V_1, V_2, \ldots, V_k$ are finite and non-disjoint sets of nodes and $E, E_1, E_2, \ldots, E_k$ are finite and non-disjoint sets of edges such that $V = V_1 \cup V_2 \cup \ldots \cup V_k$ and $E = E_1 \cup E_2 \cup \ldots \cup E_k$.

Other representations in literature only associate temporal dimension with edges and not with nodes (see, e.g., Casteigts et al. 2011). With the described representation, it is possible to represent when an object joins a system and when it leaves it. This information might be very useful in some scenarios.

Classifications of Dynamic Networks

Dynamic networks can be classified in a number of ways. We do not try to provide an exhaustive list of classification methods, but merely describe some of the most commonly used discriminators from the literature. These classification methods are briefly described below.

Online Versus Offline Dynamic Networks

Online dynamic networks are the networks where we have streaming data and we have to analyze and visualize these networks with the arrival of data streams such as climate changes. Offline networks are the networks where the complete time-stamped data is available for analysis and visualization, for example, financial transactions of a banking system or an ATM.

Continuous Versus Discrete Dynamic Networks

Discrete dynamic networks are networks where we have discrete time windows within which nodes appear and/or interactions take place, for instance, the network of players in a football game. Continuous dynamic networks do not have any such discretization, and continuous time scales are used to represent nodes and edges, for example, data packets moving over a local area computer network or the Internet.

Time-Unweighted Versus Time-Weighted Dynamic Networks

Time-unweighted dynamic networks, also known as contact sequence as defined by Holme and Saramäki (2012), are networks where interaction time among nodes is negligible, and we only consider that an interaction took place such as an email. On the other hand, time-weighted dynamic networks are those networks where interaction time is important. It represents some sort of weight that plays a vital role in the analysis and visualization of these networks, for example, the duration of a phone call or the time it took to transfer data from one location to the other.

Dynamic Graph Analytics

Analysis of dynamic graphs requires methods and metrics that can cater to the temporal dimension of these networks. The most common method to deal with the temporal aspect is to simply agglomerate the entire network over time into a single static network, but this loses almost all temporal nuances. Another common method used as preprocessing is the discretization of dynamic graphs: given a temporal graph G in a time period $t_{1,2}$ to $t_{n-1,n}$, discretization breaks G into n subgraphs $G_1, G_2, \cdots, G_n$ such that G_1 represents the state of the graph in time period $t_{1,2}$, G_2 represents the structure in time period $t_{2,3}$, and so on. We use the terms "time step" to refer to individual subgraphs according to time periods sampled from G and "time window" to refer to the time period for each subgraph. Such discretization of the network into a series of static networks can enable the extension of many static graph analysis techniques to be applicable to dynamic networks.

There are many structural analysis where the time ordering is an added constraint, for example, transitivity, i.e., the property where three

or more nodes are all connected to each other. This property may not exist in a network if the edges connecting the nodes do not appear in the same time window. However, this nuance would be lost in an agglomerate approach.

It is also possible to extend the time-step definition to a sliding time-window approach, where the discretization process for each G_a with time period $t_{p,q}$ and G_{a+1} with time period $t_{r,s}$ has some overlapping structure where the condition $r < q$ holds. This can serve to smooth out sharp changes in the network and help in the isolation of outlying abnormalities.

Dynamic Graph Models

Some graph-generating models have been proposed in the literature to mimic the dynamic processes taking place in real-world networks. Notable contributions include the work by Robins et al. (2007) and a more recent model by Kolar et al. (2010). Often topological characteristics of real-world networks are extracted and used as parameters for network models. The study of these models helps us understand and predict how networks develop and change undergoing dynamic processes. Moreover benchmark datasets can be generated for testing algorithms and data analysis techniques for extensive evaluation.

Dynamic Network Metrics

Typically SNA metrics can be divided into element level (node or edge), group-level (group of nodes) and network-level metrics (Brandes and Erlebach 2005). Traditional SNA metrics usually cannot be mapped directly to dynamic networks due to the addition of the temporal dimension. Different researchers have proposed modifications to existing static network metrics or proposed new metrics altogether to analyze dynamic networks. Even fundamental network metrics such as "degree" need modification, as the number of connections for a node across different time periods may vary substantially. Thus, new semantics must be associated with existing network metrics.

One option that is often applied to create a static agglomerate network out of the entire dynamic network weighted according to node/edge occurrence, which enables the applicability of SNA metrics. However, this ignores the dynamic nature of the network. Alternate weightings can be applied, such as weighting according to how "stable" each node/link is (i.e., how often they appear/disappear, not just how often they are there), but the result is still inherently static.

Another simple option is to simply recalculate metrics on static snapshots of the network (i.e., each time step). This can then be analyzed using time-series analysis techniques. However, it can still be naïve to analyze complex temporal patterns.

One interesting set of metrics for dynamic networks is concerned with paths and interconnectedness over time. Paths for static networks are sequences of nodes and edges such that the nodes are adjacent. In temporal graphs, instead of simple paths, the concept of time-preserving paths is introduced such that a path can only exist from one node to the other only and only if the edges appear in nondecreasing order. Mathematically, for a path $E = \{e_1, e_2, \cdots, e_n\}$ the condition holds $(e_{t_1} \leq e_{t_2}, \cdots, e_{t_n})$ where e_{t_i} denotes the time period associated with edge e_i. A simple implication of time-preserving paths is that paths are no longer symmetric as opposed to static networks, i.e., if a path from node p to q exists, a path from q to p might not necessarily exist. More details about temporal metrics can be found in Holme and Saramäki (2012).

Dynamic Community Detection

Static graph clustering has been an active area of research with well-established methods and algorithms. Schaeffer (2007) provides a good overview of some of the graph clustering methods. The purpose of such methods is to discover groups of nodes based on some similarity, either structural or based on some metric. Another successful set of methods helps to discover clusters of densely connected communities. They are generally based on an algorithm that optimizes a function such as the so-called modularity function from Newman and Girvan (2004) which represents the sum of the number of edges linking nodes of the same clusters minus the expected such sum if edges were distributed at random.

With the emergence of dynamic networks, clustering methods for static graphs fail to satisfy the new and challenging issues revolving around dynamic networks. Thus, a new area of research focuses on evolving communities within dynamic graphs: the underlying idea is to extract time-varying clusters (i.e., clusters of densely connected communities that evolve over time), by extending or adapting the algorithms developed for static graphs.

Mathematically, we define dynamic graph clustering as follows: let $G = (V, E)$ be a dynamic graph and $S = \{G_1 = (V_1, E_1), G_2 = (V_2, E_2), \ldots, G_k = (V_k, E_k)\}$ the corresponding ordered sequence of subgraphs. Following (Sallaberry et al. 2013), a time-varying clustering of a dynamic graph G is defined as a set of time-varying clusters $VC = \{VC_1, VC_2, \ldots, VC_l\}$. Each of these time-varying clusters is an ordered sequence $VC_i = \{vc_i^1, vc_i^2, \ldots, vc_i^k\}$ where k is the number of time steps and each vc_i^t is a subset of the vertices V_t at time t. That is, each time-varying cluster VC_i is a cluster whose membership can evolve over time, where vc_i^t represents the set of nodes in the cluster i at time t.

A first approach for creating a time-varying clustering is to apply directly a static graph clustering algorithm to an agglomerate of a dynamic graph $G = (V, E)$ (Hu et al. 2012). It gives a partition $C = \{c_1, c_2, \ldots, c_l\}$ of the nodes of V, which is used to create the dynamic clustering with $VC_i \leftarrow c_i$ for each time step and each $vc_i^t \leftarrow c_i \cap V_t$. As densely connected communities are extracted on the union of the time steps, it does not guarantee that each cluster of each time step is densely connected.

An alternate approach to overcome this issue is to use a static graph clustering algorithm for each time-step subgraph G_i and then associate the clusters across time to derive time-varying clusters (Sallaberry et al. 2013). In this approach, clusters are iteratively computed by comparing each time-step cluster in the current time-step pairwise with the time-step clusters of the previous time step according to a similarity index and then greedily associating the time-step clusters that most closely match into the same time-varying cluster. Once all clusters are assigned or the similarity falls below a user-defined threshold, this process terminates and moves to the next time step. Any remaining new clusters that do not have a good enough match are considered new clusters, so they start new time-varying clusters. And any remaining time-varying clusters present in the previous time step that were not assigned a cluster in the current time step were discarded, as there was no match.

Cazabet et al. (2010) also propose an interesting approach to detect dynamic communities. The algorithm is designed to detect strongly overlapping communities. The authors introduce two notions of intrinsic communities and longitudinal detection which drives the iLCD (intrinsic Longitudinal Community Detection) algorithm.

Dynamic Graph Visualizations

Visualization methods for dynamic graphs present a challenging task for researchers. Traditional methods for the visualization of static graphs cannot be applied to dynamic networks. The problem is fundamental to the static representations used as they require at least two dimensions to represent proximity; as a result, time cannot be represented on a two-dimensional plan (Moody et al. 2005). A simple approach to visualize a dynamic graph would have been as a static image, but images fail to represent change occurring in networks over time (Moody et al. 2005).

Despite the success of visualization in static networks (Freeman 2000), most of the approaches used in dynamic network analysis avoid using visualizations (Trier 2008). These approaches mostly rely on measures and metrics to establish hypothesis and base their analysis on these statistics. A major reason for this limited use of visualizations is to ensure the stability of the layouts used. A stable layout helps preserve the user's mental map as there is less movement between time steps, but sacrifices quality in terms of readability for later time steps as their layout depends on previous time steps. Many experiments have been proposed to examine the effect of preserving the mental

map in dynamic graphs visualization (Purchase and Samra 2008). The results of Purchase and Samra (2008) were quite surprising because the most effective visualizations were the extreme ones, i.e., the ones with very low or high mental map preservation: visualizations with medium preservation performed less well.

Dynamic network visualization helps to visualize dynamic network processes which in turn helps to build models that can predict future evolutionary processes. Static representation fails to accurately build a dynamic model because using graphs, we develop a static model of a phenomenon which is inherently dynamic and then we try to make inferences and predictions of how it will evolve over time.

Visualization methods for dynamic graphs can be categorized into animated visualizations, static visualizations, and hybrid visualizations which are described below.

Animated Visualizations

A common method for visualizing dynamic graphs is to animate the transitions between time steps (Frishman and Tal 2008). For example, the approaches proposed by Moody et al. (2005) and Frishman and Tal (2008) render individual network graphs segmented from entire dynamic graph. These individual graphs are then visualized as animated sequence with nodes appearing, disappearing, and moving to produce readable layout for each time step. This animation facilitates the analysis and understanding of dynamic processes taking place in a network and thus has gained a lot of popularity. One drawback of this approach is its scalability as the number of nodes and processes increases, it becomes difficult to visually comprehend the changes taking place in the network.

Approaches to visualize dynamic clustered graphs also exist in the literature such as Hu et al. (2012) who proposed a method based on a geographical metaphor to visualize clustered dynamic graphs.

All the aesthetic requirements for an animated visualization remain the same as that of static visualization such as minimizing edge crossings and node overlap. One aspect which needs to be handled for animations is the movement of nodes between time steps which needs to be meaningful and coherent with the network processes taking place.

Static Visualizations

Static visualizations have also been successfully used to visualize dynamic networks. Different techniques have been proposed in the literature such as "small multiples" (Tufte 1990) where snapshots of different time steps are placed next to each other. This technique eases the comparison of distant time steps, but the area devoted for each time step is small, and this reduces the readability of each subgraph. Another technique named "flipbook" (Moody et al. 2005) only displays edges within a time window whereas nodes maintain a fixed position. As the time window is moved, edges appear and disappear, showing interactions taking place at different time intervals.

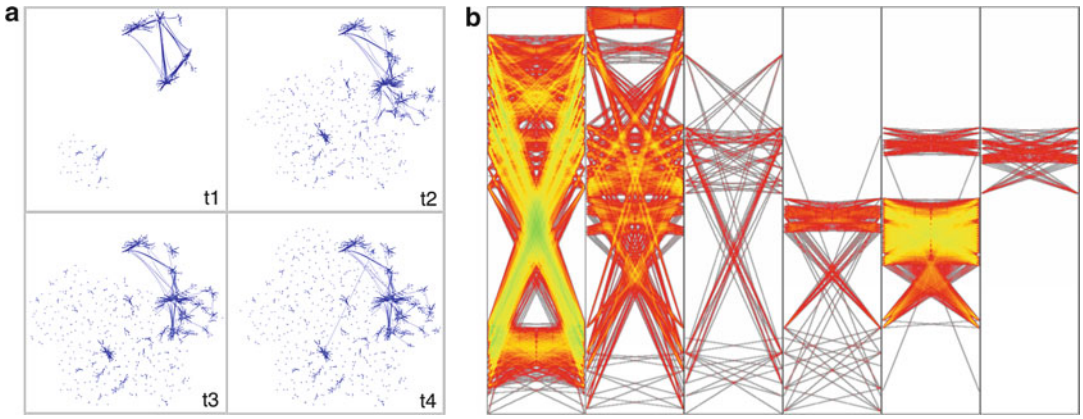
Another interesting visualization technique dealing with dynamic large directed graphs has been proposed by Burch et al. (2011) where vertices are ordered and positioned on several vertical parallel lines, and directed edges connect these vertices from left to right. Each time step's graph is thus displayed between two consecutive vertical axes (see Fig. 1b).

Hybrid Visualizations

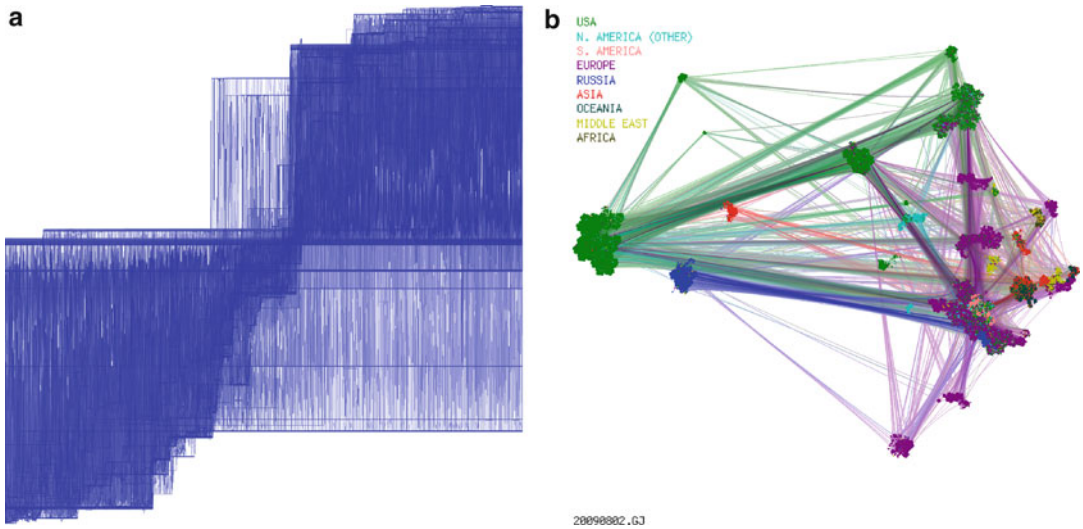
Some approaches combine both static representations with dynamic layouts to provide both a summary overview and detailed views. One such method (Sallaberry et al. 2013) computes time-varying clusters and orders both the clusters and individual nodes in 1 dimension. This is used to both define a 2-dimensional overview of the network over time and defining temporally stable layouts for any given time (see Fig. 2).

Key Applications

Biological Networks One of the most successful applications of DNA is in the area of biological networks. There are a number of visual analysis tools such as Cytoscape, VisANT, Pathway Studio, Patika, and ProVis-Tulip



Analysis and Visualization of Dynamic Networks, Fig. 1 (**a**) Small multiples: snapshots of different time steps are placed next to each other. (**b**) Parallel edge splatting for scalable dynamic graph visualization: vertices are ordered and positioned on several vertical parallel lines, and directed edges connect these vertices from *left* to *right* (Image from Burch et al. 2011 and reproduced with permission)



Analysis and Visualization of Dynamic Networks, Fig. 2 (**a**) Overview: each node is represented as a line where the x position is time and the y positions correspond the cluster the nodes belong to at each given time and its position within the cluster. (**b**) Time-step view: node-link diagram that shows the graph at any selected time step

(Suderman and Hallett 2007; Pavlopoulos et al. 2008). These tools currently focus on the dynamic behavior of biological systems like modeling cellular processes, subcellular localization information and time-dependent behavior, and interaction of protein-protein/protein-nucleotide networks (Suderman and Hallett 2007; Akhmanova and Steinmetz 2008). This recent interest is due to the huge amount of dynamic data available where visualization tools have limited capabilities to present readable images from these networks. Thus, the dynamism of these networks allows domain experts to focus on small time periods with minimal activity, making it easier for analysis and drawing conclusions.

For example, Akhmanova and Steinmetz (2008) study microtubule plus-end-tracking proteins which form dynamic networks through the interaction of a limited set of proteins modules. Taylor et al. (2009) study the dynamic

structure of human protein interaction network to predict the presence of breast cancer using structural changes.

Terrorist Networks Another highly sensitive area where DNA is widely utilized is the area of security and terrorist networks. Criminals and terrorists work as a cohesive group to achieve desired outcomes, and no single person is responsible to do it all. The roles and duties of these individuals keep changing, as well as new individuals are introduced in the network regularly. Furthermore a lot of information are missing or inaccurate, making it a challenging domain for social network researchers. A famous success story of the application of this approach is the capture of the former president of Iraq, Saddam Hussein (Borgatti et al. 2009).

One notable work is the software system named DynNetSim by Adler (2007) to model and analyze dynamic networks. The software provides a holistic view of networks responding to influences of environmental forces and disruptive events. DynNetSim also helps to study impacts of strategies to change networks and enhance different counter-terrorism activities. Gilbert et al. (2011) studied the dynamic behavior of terrorist groups given the data of cellular phones. They provide a complete framework starting from time-stamped network data to presenting visual drawings for analysis and knowledge extraction.

Computer Networks Computer security often requires monitoring of complex and highly dynamic networks, often in real time. There are many layers to computer networks, including physical connectivity, routing tables, network flows/traces, or even application level semantics such as online social networks or hyperlinks on the World Wide Web. There are numerous approaches that have been applied to these networks. Due to the highly dynamic nature of some of these networks (e.g., network flows), many approaches ignore the topology. However, some levels of this network are temporally stable enough to apply DNA (e.g., routing tables or hyperlinks). One example particularly worth noting here is Sallaberry et al. (2013), which investigates routing tables at the scale of the entire Internet.

Conclusion and Future Directions

A major reason for the popularity of the field of dynamic networks is its applicability in a number of diverse fields. The field of dynamic networks is in its infancy, and there are so many avenues that need to be explored. From developing network generation models to developing temporal metrics and measures, from structural analysis to visual analysis, there is room for further exploration in almost every dimension where dynamic networks are studied. Recently, with the availability of dynamic data from various fields, the empirical study and experimentation with real data-sets has also helped maturate the field. Furthermore, researchers have started to develop foundations and theories based on these datasets which in turn has resulted lots of activity among research communities.

While there is a growing corpus of works on dynamic graph analysis, there are still many directions for further investigation. Dynamic graph metrics are largely unexplored, as most existing dynamic graph metrics are merely applications of static graph metrics. Likewise, dynamic graph clustering is still in its infancy, as most effective dynamic graph clusterings are currently based on direct extensions to static graph clusterings. While dynamic graph visualization methods have been establishing a growing foothold, there is still much room for novel approaches. In particular, scalability will only become a more important issue, as there are very few works that can handle dynamic networks at the scale of many of today's large real-world networks. Addressing topics such as these will serve to greatly further both the field of network analysis and any other of the vastly numerous field where such dynamic networks occur.

Cross-References

▸ Community Evolution
▸ Community Identification in Dynamic and Complex Networks

▸ Gephi
▸ ORA
▸ Pajek
▸ Tulip III
▸ UCINET
▸ Visual Methods and Tools for Social Network Analysis
▸ Visualization of Large Networks

References

Adler RM (2007) A dynamic social network software platform for counter-terrorism decision support. In: ISI, New Brunswick. IEEE, pp 47–54

Akhmanova A, Steinmetz MO (2008) Tracking the ends: a dynamic protein network controls the fate of microtubule tips. Nat Rev Mol Cell Biol 9(4): 309–322

Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512

Bender-deMoll S, McFarland DA (2006) The art and science of dynamic network visualization. J Soc Struct 7(2):1–38

Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. Science 323(5916):892–895

Brandes U, Erlebach T (eds) (2005) Network analysis: methodological foundations. Lecture notes in computer science, vol 3418. Springer, New York

Burch M, Vehlow C, Beck F, Diehl S, Weiskopf D (2011) Parallel edge splatting for scalable dynamic graph visualization. IEEE Trans Vis Comput Graph 17(12):2344–2353

Casteigts A, Flocchini P, Quattrociocchi W, Santoro N (2011) Time-varying graphs and dynamic networks. In: Proceedings of the 10th international conference on Ad-Hoc, mobile, and wireless networks, ADHOC-NOW'11, Paderborn. Springer, pp 346–359

Cazabet R, Amblard F, Hanachi C (2010) Detection of overlapping communities in dynamical social networks. In: IEEE second international conference on social computing (SocialCom), Minneapolis. IEEE, pp 309–314

Freeman LC (2000) Visualizing social networks. J Soc Struct 1(1):1–15

Freeman LC (2004) The development of social network analysis: a study in the sociology of science. Empirical/BookSurge, Vancouver

Frishman Y, Tal A (2008) Online dynamic graph drawing. IEEE Trans Vis Comput Graph 14(4): 727–740

Gilbert F, Simonetto P, Zaidi F, Jourdan F, Bourqui R (2011) Communities and hierarchical structures in dynamic social networks: analysis and visualization. Soc Netw Anal Min 1:83–95

Holme P, Saramäki J (2012) Temporal networks. Phys Rep 519(3):97–125

Hu Y, Kobourov SG, Veeramoni S (2012) Embedding, clustering and coloring for dynamic maps. In: Proceedings of the 5th IEEE Pacific visualization symposium (PacificVis 2012), Songdo, pp 33–40

Kolar M, Song L, Ahmed A, Xing EP (2010) Estimating time-varying networks. Ann Appl Stat 4: 94–123

Moody J, Mcfarland D, Bender-demoll S (2005) Dynamic network visualization. Am J Sociol 110(4):1206–1241

Moreno J (1934) Who shall survive? Nervous and Mental Disease Publishing Company, Washington

Newcomb TM (1961) The acquaintance process. Holt, Rinehart and Winston, New York

Newman MEJ, Girvan M (2004) Graph clustering. Phys Rev E 69:026113

Pavlopoulos G, Wegener AL, Schneider R (2008) A survey of visualization tools for biological network analysis. BioData Min 1(1):12

Purchase H, Samra A (2008) Extremes are better: Investigating mental map preservation in dynamic graphs. In: Proceedings of the 5th international conference on diagrammatic representation and inference (Diagrams 2008), Herrsching. Lecture notes in computer science, vol 5223. Springer, pp 60–73

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

Sallaberry A, Muelder C, Ma KL (2013) Clustering, visualizing, and navigating for large dynamic graphs. In: Proceedings of the 20th international symposium on graph drawing (GD 2012), Redmond. LNCS 7704, Springer, Berlin/Heidelberg, pp 487–498

Sampson SF (1968) A novitiate in a period of change: an experimental and case study of social relationships. PhD thesis, Cornell University

Schaeffer SE (2007) Graph clustering. Comput Sci Rev 1(1):27–64

Suderman M, Hallett M (2007) Tools for visually exploring biological networks. Bioinformatics 23(20): 2651–2659

Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q, Wrana JL (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol 27(2):199–204

Trier M (2008) Towards dynamic visualization for understanding evolution of digital communication networks. Inf Syst Res 19(3):335–350

Tufte ER (1990) Envisioning Information. Graphics Press, Cheshire

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393:440–442

Recommended Reading

The current literature lacks a comprehensive text covering all aspects of dynamic networks. A highly informative and recent work titled "Temporal Networks" by Holme and Saramäki (2012) reviews most of the analytical part related to dynamic networks from the current literature. The authors primarily focus on dynamic network metrics, methods of representing dynamic data as static networks, models for generating temporal networks, and the spreading dynamics in these networks.

One of the landmark papers for dynamic network visualization is titled "Dynamic Network Visualization" by Moody et al. (2005) where they introduce two important concepts of network visualization, network movies, and flipbook. Bender-deMoll and McFarland's (Bender-deMoll and McFarland 2006) article "The Art and Science of Dynamic Network Visualization" is also very interesting to read as it reviews existing layout algorithms for static networks and how they can be used for visualization of dynamic networks. Trier (2008) also studies the problem of dynamic network visualization to analyze online social communities. The author uses animated graphs and measures changes to describe cluster formation processes, relate node-level analysis to network-level analysis, and measure how external events change network structures.

Analysis of Social Relations

▸ Networks at Harvard University Sociology

Analytical Models for Social Preferences

▸ Modeling Social Preferences Based on Social Interactions

Annotations

▸ Social Bookmarking

Anonymity

▸ Anonymization and De-anonymization of Social Network Data

Anonymization and De-anonymization of Social Network Data

Sean Chester, Bruce M. Kapron, Gautam Srivastava, Venkatesh Srinivasan, and Alex Thomo
Department of Computer Science, University of Victoria, Victoria, BC, Canada

Synonyms

Adversarial knowledge; Anonymity; Complexity; Graph algorithms; Privacy breach; Social network privacy

Glossary

Adversary Somebody who attempts to reveal sensitive, private information

Adversarial Model Formal description of the unique characteristics of a particular adversary

Attribute Disclosure A privacy breach wherein some descriptive attribute of somebody is revealed

Identity Disclosure A privacy breach in which a presumably anonymous person is in fact identifiable

***k-P*-Anonymity** A condition under which any instance of P appears at least k times

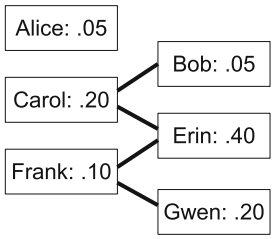
Target The particular social network member against whom an adversary is trying to breach privacy

Definition

As social networks grow and become increasingly pervasive, so too do the opportunities to analyze the data that arises from them. Social network data can be released for public research that can lead to breakthroughs in fields as diverse as marketing and health care. But with the release of data come questions of privacy. *Is there any*

information that members of the social network would not want revealed publicly? If it is released, can somebody (an adversary) attribute that information to them?

Anonymization is the modification of data so that sensitive information remains private. De-anonymization is the converse: reidentifying somebody in an anonymized network – or even simply learning something about them that was meant not to be attributable to them.



Anonymization and De-anonymization of Social Network Data, Fig. 1 Small example supervisor network. An edge (*a*, *b*) exists if a supervises b or vise versa. Vertices are also annotated with contribution percentage

Introduction

Say we the authors wanted to stimulate research on supervisory patterns among coauthors by releasing the small social network depicted in Fig. 1. The network contains an edge between two coauthors if one supervises the other, and each vertex is labelled with the percentage that the author contributed to the research. Certainly, the labels are quite sensitive, and Alice, for example, may not want it publicly known that her contribution level was lower. To protect the privacy of the coauthors, then, the social network must first be *anonymized*. In some cases, that might be a simple enough task: just remove the names and replace them with random integers (so-called *naive identity anonymization*).

Releasing the data makes it available for myriad analyses that the coauthors had not even anticipated. It also makes it available to Dean, an adversary who wishes to *de-anonymize* the data to uncover the sensitive information. In particular, he may want to reveal Alice's contribution level and may know that Alice, from another affiliation, has no supervisory relationship with any other author. Even after names have been stripped from the network, Dean can still exploit this background information about the *structure* of the social network graph to reidentify her and conclude her label (Backstrom et al. 2007); viz., she is the only isolated vertex.

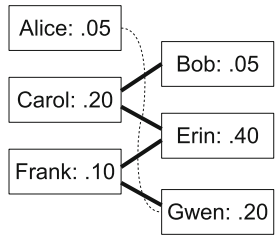
Similarly, Dean may instead know that Erin co-supervises two of the coauthors. This is sufficient structural information to reidentify Erin, because she is the only vertex in the network connected to two other vertices who also have degree two. Something more must be done to protect the identities of Alice and Erin, and that is going to have to involve distorting the network somehow, because it is the structure of the network that reveals their identities. This is social network anonymization: distorting a social network to the point that some assumed knowledge of an adversary is rendered uninformative.

Now consider Fig. 2 in which the fictitious relationship between Alice and Gwen has been added. Now, Alice and Bob have the same degree; even if Dean knows the degree to be one, he cannot distinguish between them. Erin, too, is similarly unidentifiable, because now Frank is connected to two vertices of degree two. In fact, no matter who Dean targets, even with the knowledge of the target neighbor's degrees, he is left with a $1/2$ chance of guessing who is his target once the fictitious relationship is added.

Key Points

Throughout these two examples of Alice and Erin, we knew what knowledge Dean possessed. As we continue in this chapter, we assume different levels of knowledge for Dean, and with each we ask:

- *Can* we protect ourselves from Dean's knowledge while still releasing the data?
- *When* can we do this?
- If we cannot do this, *why* not?



Anonymization and De-anonymization of Social Network Data, Fig. 2 A 2-anonymization of Fig. 1 by adding a fictitious edge among Alice and Gwen. Notice now that no vertex degree is unique

Historical Background

The need for privacy in publicly released data is not new. Relational (i.e., not social network) data has been shared for decades. Many of the ideas for social network anonymization stem from what has been researched and learned about anonymizing table data. The pivotal idea of k-anonymization that we introduce shortly originated with publishing relational data (Sweeney 2002). The privacy of individual table records can be well preserved if, under projection on *quasi-identifying* attributes (e.g., zip code, birthdate), the record is made identical to at least $k-1$ others by a series of data suppressions.

With the onset of pervasive social networking in recent years, there has been a rush to adapt some of these ideas for social network (i.e., graph) data. The task is challenging, however, because graph structure was shown by Backstrom et al. (2007) to *quasi-identify* people itself, before even considering the labels with which social networks are annotated. Since then, research has focused on what can, indeed, conceal one's identity (i.e., prevent *identity disclosure*) in a social network and what can conceal the attributes that describe you (i.e., prevent *attribute disclosure*).

Tools and Techniques for (De-)Anonymization

What It Means to Be Identical: k-Anonymization Formalized

In section "Introduction" we presented two examples of knowledge that Dean may have. We may try to protect against the sort of knowledge Dean had of Alice, viz., her degree, or that he had of Erin, viz., her 1-neighborhood. In general, we assume Dean knows some local structural property P of his target. By adding one edge, Alice and Erin were both protected because they became structurally identical to other vertices. That is to say, the graph became *k-P-anonymous*, wherein every vertex is identical to at least $k-1$ other vertices with respect to P, considering both cases where P is degree and where P is 1-neighborhood. No matter who Dean targets with his knowledge of P in a *k-P*-anonymous graph, he is left with at best a $1/k$ chance of guessing his target's identity correctly.

Definition 1 *k-P-anonymous graph.* A graph $G = (V, E)$ is *k-P-anonymous* iff the vertices can be completely partitioned into disjoint subsets such that each subset has size at least k and, within every subset, every vertex is identical with respect to P.

As a concrete example, P might be the *degree* of a vertex. The graph in Fig. 2 is 2-degree-anonymous. If an edge is added between Alice and Bob, the graph will become k-degree-anonymous for all $k \leq |V|$, since every vertex will have degree two. For a graph that is not *k-P*-anonymous, the task prior to release is to minimally distort it until it becomes *k-P*-anonymous.

Problem 1 *k-P-anonymization.* Given an input graph $G = (V, E)$, a structural property P, and a privacy threshold k, construct a graph $G' = (V, E')$ such that G' is *k-P*-anonymous, $E \subseteq E'$, and $|E'|$ is minimized.

Much of the research in literature focuses on defining appropriate properties to study or on algorithms to achieve *k-P*-anonymity, such as the two we present next.

Anonymity with Random Perturbation

A first anonymization algorithm for a graph G is to initially add m randomly chosen edges to produce an intermediate graph G_{int}, and then remove m randomly chosen edges from G_{int} to produce an anonymized graph G' (Hay et al. 2007).

The choice of m is a balance between minimizing distortion of the graph and ensuring that $\geq k$ vertices in G′ could have plausibly originated as Dean's target. By introducing randomness, Dean is forced to reason within possible world semantics and is confronted with at least k likely candidates as his target. So, although the resultant graph is not necessarily k-P-anonymous, it does leave Dean with a $1/k$ chance at guessing his target correctly.

k-Degree-Anonymization with Dynamic Programming

A second algorithm, greedy and specifically for degree-based attacks, is based on the *degree sequence* of G (Lui and Terzi 2008):

Definition 2 *Degree sequence*. Given a graph $G = (V, E)$, where the degree of a vertex v_i in V is denoted d_i, the degree sequence S_G of G is a sorted sequence of integers of length $|V|$ wherein the frequency of any integer i is exactly $|\{v_j \text{ in } V : d_j = i\}|$. If the frequency of every integer is either zero or $\geq k$, the degree sequence is k-anonymous.

A k-degree-anonymous graph G will have a k-anonymous degree sequence. The algorithm uses dynamic programming to produce a k-anonymous integer sequence nearest to the degree sequence of G and, then tries to produce a graph with a degree sequence matching that integer sequence. A graph can be produced from a sequence iff it meets the condition of the Erdos-Gallai Theorem for degree sequence realizability (Erdos and Gallai 1960); but for the anonymity problem, that graph must contain every edge in the original graph. If the sequence does not meet the Erdos-Gallai condition or if the original graph cannot be augmented to match the target degree sequence, then, repeatedly until success, some random noise is added to the degree sequence of G, a new sequence is constructed, and the conditions are rechecked.

From the work of Lui and Terzi (2008), the dynamic programming proceeds as follows. First, let C([1,d]) be the cost of anonymizing the first d integers in the sequence, and let $S([a, b]) = \Sigma_{a \leq i \leq b}(d_b - d_i)$. Then:

$$\text{For } i \leq 2k : C([1, i]) = S([1, i]).$$

$$\text{For } i > 2k : C([1, i]) = \min\{\min_{k \leq t \leq i-k} \{C([1, t]) + S([t + 1, i])\}, S([1, i])\}.$$

If δ_i is the difference between the i'th integer and the largest within the same partition, then an optimal degree sequence partitioning is one which minimizes $\Sigma\delta_i$. Minimizing $C([1, |V|])$ with this dynamic programming produces an optimal partitioning. The Lui and Terzi (2008) algorithm then checks the Erdos-Gallai condition for the new sequence constructed by increasing each i'th integer by δ_i and, when successful, adding δ_i edges to the i'th vertex.

While this algorithm has no performance guarantees, experimental comparisons (Casas-Roma et al. 2012; Ying et al. 2009) show that it typically reaches a k-degree-anonymous solution with less distortion than the random perturbation techniques. On the other hand, it is slower to reach a solution.

Broader Local Knowledge

The algorithms in section "What It Means to Be Identical: k-Anonymization Formalized" can protect a social network against an adversary Dean when Dean's knowledge is limited to the degree of his target, as he knows about Alice. But what if Dean is more powerful, as in the example of Erin? Several formalizations exist of a more powerful Dean, one who knows a more identifying property P. Correspondingly, stronger notions of k-P-anonymity are required.

Stronger Adversarial Models

To keep the examples easier to understand, we have used degree and neighborhood as the structural knowledge P possessed by Dean. The former leads to k-degree-anonymity (Lui and Terzi 2008). When Dean knows the entire neighborhood of his target (every neighbor *and* how they are connected) (Zhou and Pei 2008), as he does with Erin, privacy requires k-neighborhood-anonymity, in which the way neighbors are connected for every vertex must

be identical to at least $k-1$ other vertices. Many other models have been proposed. For example, Dean may know the *i-hop* neighborhood of his target: all the neighbors within a path of length i (Thompson and Yao 2009). Yet stronger models have been proposed, too, based on isomorphisms (Cheng et al. 2010) and symmetry (Wu et al. 2010). While achieving each progressively stronger anonymity requirement offers greater privacy protection (presumably at the cost of graph utility), one must be careful of expecting too much, because, as shown in the next section, even reasonably modest adversarial models lead to NP-hard problems.

Complexity of k-P-Anonymity

Interesting algorithms have been designed for many forms of anonymization or relation tables and social network graphs. These have been shown to perform quite well on real-world datasets but do not have any theoretical performance guarantees. That is, there is no guarantee that these algorithms distort the input optimally in order to obtain the anonymized output. So, researchers investigated if there is an efficient algorithm, running in polynomial time, that can anonymize a given table or a graph using the minimum amount of modification required.

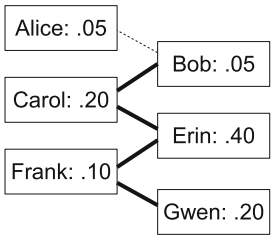
For table anonymization, a sequence of results showed that it is NP-hard to anonymize a table using the minimum number of suppressions required. These results were shown using reductions from known NP-hard graph optimization problems. Using a reduction from *hypergraph matching*, Meyerson and Williams (2004) showed that k-anonymization of tables is NP-hard provided that the number of values an attribute can assume (*alphabet size*) is larger than the number of rows in the table. This result was improved by Aggarwal et al. (2005) who showed a hardness result for a ternary alphabet using a reduction from *Partition into Triangles*. Finally, Bonizzoni et al. (2009) obtained a hardness result for binary tables via a reduction from *minimum vertex cover*.

The hardness results for anonymizing tables were, in turn, used to show NP-hardness results for graph anonymization such as 1-neighborhood anonymization (Zhou and Pei 2008) in vertex-labeled graphs and label-sequence anonymization in edge-labeled graphs (Chester et al. 2012c). We now illustrate the main idea behind the reductions in these papers using the reduction of Zhou and Pei (2008). Given a binary table T with n rows, l columns, and anonymity parameter k, build a bipartite graph $G_T = (U, V, E)$. U is a set of n vertices labeled $\{r_1, r_2, \ldots, r_n\}$ corresponding to the rows of the table. V is a set of k copies of $2l$ vertices labeled $\{c_{10}, c_{11}, c_{20}, c_{21}, \ldots, c_{l0}, c_{l1}\}$ corresponding to the columns of the table. If the (i, j)-entry of T is 0, draw k edges from vertex r_i to the k vertices labeled c_{j0}. If the (i, j)-entry of T is 1, draw k edges from vertex r_i to the k vertices labeled c_{j1}. It can be shown that T can be k-anonymized using at most s entry suppressions if and only if the graph G_T can be 1-neighborhood-anonymized using at most ks edge additions.

So, one can construct schemes to k-P-anonymize a graph, and those schemes can work well in practice and preserve the utility of the graph reasonably well. But if the objective is to construct an *optimal* anonymization – or even just one with a *fixed* level of distortion – the problem is NP-hard.

Alternative Formulations of k-Anonymity

Although most research models the problem of k-P-anonymity as in Problem 1, a few other approaches have been suggested as well. For example, one can try to achieve k-P-anonymity by adding vertices as well as edges to the input graph (Chester et al. 2012b). Also, many social networks contain vertices that do not necessarily need to be anonymous because they do not represent typical users. Consider Twitter accounts for major sports teams and celebrities, for example. In such instances, one can potentially achieve k-anonymity with very minimal distortion by aiming only for *subset anonymity* (Chester et al. 2012a). A particularly recent suggestion is to output a probabilistic graph wherein the anonymity requirement is satisfied by injecting uncertainty on edges rather than just adding and removing them (Boldi et al. 2012).



Anonymization and De-anonymization of Social Network Data, Fig. 3 Attribute-diversifying 2-degree-anonymization of Fig. 1. Now the label range for degree-1 vertices is [0.05, 0.20]

Attribute Disclosure

In another type of attack, the adversary Dean is not necessarily interested in identifying his target, but merely inferring her label. Such an attack is called *attribute disclosure*. Consider again the 2-degree-anonymization in Fig. 2. Despite knowing that her degree is one, Dean is unable to ascertain which vertex represents Alice and which represents Bob. He *can* infer, however, that Alice's contribution is 0.05, because the label is the same for both vertices.

The 2-degree-anonymization given in Fig. 3 achieves the same level of identity anonymization with the same number of additional edges as the anonymization in Fig. 2. This time, however, Dean's knowledge of Alice's degree can only reveal Alice's contribution to be within the range [0.05, 0.20], because Alice is now in the same equivalence class as Gwen, not Bob. The new anonymization also expands the label range for the degree-2 vertices from [0.10, 0.40] to [0.05, 0.40]. So, Fig. 3 offers an anonymization that better protects the sensitive information about everyone.

If Dean can infer which vertex is Alice or if Alice's equivalence class has a small label range (like in Fig. 2), then attribute disclosure will occur. So, k-anonymity is necessary, *but in addition to that*, some attribute concealment condition must also be met. The graph is *l-diverse* (an adaptation from the similar idea in table literature (Machanavajjhala et al. 2007)) if each equivalence class contains at least l different labels (Zhou and Pei 2008). A graph could also be made α-*proximal* (an adaptation for graphs of *t-closeness* (Li et al. 2007)) if the distribution of labels in each vertex's neighborhood is within α of the distribution across the entire network (Chester and Srivastava 2011).

With distortions to a social network that sufficiently diversify attribute labels among equivalence classes (defined by P) that are sufficiently large, Dean's local knowledge about graph structure can, in fact, be rendered uninformative. But what if Dean's knowledge goes beyond that?

De-anonymization Beyond Local Knowledge

A common approach in anonymization of social networks is *naive identity anonymization*: Before the data is released, any sensitive information associated with individual vertices of the social network graph is suppressed, and a sanitized graph that only reveals edge relationships between users is released for data mining purposes. Does this method work well in practice? There is now sufficient evidence that it *does not*. It has been shown that de-anonymization attacks can be used to extract sensitive information about certain users from such an anonymized graph by an adversary whose knowledge is *global* in nature.

Backstrom et al. (2007) showed how active and passive attacks can be used to reveal true identities of specific users easily by an adversary whose only knowledge is an identity-anonymized version of the social network graph. An active adversary can create a small number of dummy nodes with a special edge pattern among themselves and with edges to users whose privacy it wishes to violate. Later, it easily finds this edge pattern to locate the dummy nodes in the released network and hence reidentify other users in the network. They also describe passive attacks in which a group of users can collude to discover their location in the anonymized graph using the knowledge of the edge structure among themselves. This information is in turn used to violate privacy of their immediate neighbors. It was pointed out by Narayanan and Shmatikov (2009) that this approach has some limitations in practice. For example, active attacks involving a large number of nodes may not be feasible

in many real-world social networks such as a phone-call network. Furthermore, the lack of incoming edges to the dummy nodes in a directed graph could make the network operator suspect and identify an active attack.

Another notable work on de-anonymization is by Narayanan and Shmatikov (2009), who show that the nodes in a fully identity-anonymized social network graph (targets) can be identified quite effectively when the adversary has available another (auxiliary) social graph that has a significant overlap with the target graph. Their experiments with a crawled Twitter graph as a target graph, and a Flicker graph as auxiliary graph showed that the Twitter nodes could be recognized (de-anonymized) with a low error rate. The method used is based on first discovering the mappings of a small set of nodes in the auxiliary graph, the "seeds," to corresponding nodes in the target graph. Then, these mappings are propagated to other nodes in the neighborhoods of the seeds, and the propagation continues similarly to neighborhoods of the nodes discovered so far, until no more nodes can be discovered any further. The mapping exploration crucially depends on matching the degrees of the nodes in the auxiliary graph to degrees of the nodes in the target graph. Despite the success of the Narayanan's and Shmatikov's method, what remains to be investigated is the amount of disruption that can be caused on its effectiveness when the target graph is degree-anonymized as opposed to only identity-anonymized.

A more recent work by Srivatsa and Hicks (2012) used a method similar in spirit to Narayanan and Shmatikov's to de-anonymize mobility traces. Location-based services that release anonymized data about location traces of various users gathered from smartphones and GPS-sensor data have become very popular. They show that such mobility traces can be de-anonymized if the adversary has auxiliary information in the form of a social network involving the participating users. For example, they were able to de-anonymize bluetooth contact traces of a set of conference attendees using their DBLP coauthorship graph as auxiliary information.

Differential Privacy

A rather different approach to anonymization is *differential privacy*, which does not require the release of data. Differential privacy provides a model for privacy-preserving analysis of statistical databases, which are collections of records or datasets, which contain statistical information about individuals. It is characterized by a property of algorithms operating on the data, typically computing some statistical function (query) of the data. In particular, a randomized algorithm K is differentially private if for all datasets D, D′ which are *close* (i.e., one may be obtained from the other by the deletion of exactly one record,) and all $S \subseteq \mathrm{Rng}(K)$,

$$\Pr[K(D) \in S] \leq e^{\varepsilon} \cdot \Pr[K(D') \in S].$$

This definition captures the intuitive requirement that the distribution of the output of a statistical function should not be significantly influenced by the participation of a particular individual. A natural concern here is the trade-off between utility and privacy, in particular, whether it is possible to compute functions which are statistically useful while maintaining privacy. A natural approach to devising such functions is output perturbation, that is, the addition of some form of noise to the output of the statistical function. This must be done with care, for example, to avoid noise cancellation over a sequence of queries, but techniques based on the addition of Laplacian and other forms of noise have been proposed which provide differential privacy and lead to useful mechanisms for various problems in statistics (e.g., contingency table release) and learning theory. A further discussion of techniques and results in differential privacy is beyond the scope of this article; we refer the reader to the survey by Dwork (2008) for a detailed presentation.

In the setting of graphs, two versions of differential privacy are immediately apparent, namely, node differential privacy and edge differential privacy. The definitions of both will follow the pattern for database privacy, differing only on the notion of what it means for two graphs to be close. Graphs G and G′ are close in the edge setting if one may be obtained from the other by

the deletion of exactly one edge, and, in the node setting, if one may be obtained from the other by the deletion of exactly one node and its adjacent edges. Edge differential privacy is introduced by Nissim et al. (2007), where it is shown how to compute differentially private approximations of minimum spanning tree cost and number of triangles. In subsequent work (Hay et al. 2009; Karwa et al. 2011) refined techniques that are used to obtain further results, including differentially private approximations of the degree sequence. A recent paper Kasiviswanathan et al. (2013) considers node differential privacy for problems including edge counting, small subgraph counting, and degree distribution.

Future Directions

The field of social network anonymization and the opposing field of social network de-anonymization are both quite young and rapidly expanding. Section "Alternative Formulations of k-Anonymity" shows some ways in which the original notion of k-anonymity for graphs is being challenged, and assessing the merits of and extending these approaches needs to be done. Many schemes and techniques do exist, but there is still little secondary literature reviewing these. Finally, one cannot necessarily release social network data and be fully confident that nobody can attack it. Methods for preventing the global attacks described in section "De-anonymization Beyond Local Knowledge" must first be developed.

Cross-References

▸ Consequences of Publishing Real Personal Information in Online Social Networks
▸ Dark Sides of Social Networking
▸ Ethics of Social Networks and Mining
▸ Privacy in Social Networks, Current and Future Research Trends on
▸ Statistical Research in Networks – Looking Forward
▸ Transforming and Integrating Social Networks and Social Media Data

References

Aggarwal G, Feder T, Kenthapadi K, Motwani R, Panigrahy R, Thomas D, Zhu A (2005) Anonymizing tables. In: Proceedings of the ICDT, Edinburgh, pp 246–258

Backstrom L, Dwork C, Kleinberg JM (2007) Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In: Proceedings of the WWW, Banff, pp 181–190

Boldi P, Bonchi F, Gionis A, Tassa T (2012) Injecting uncertainty in graphs for identity obfuscation. PVLDB 5(11):1376–1387

Bonizzoni P, Vedova GD, Dondi R (2009) The k-anonymity problem is hard. In: Proceedings of the FCT, Wroclaw, pp 26–37

Casas-Roma J, Herrera-Joancomart J, Torra V (2012) Comparing random-based and k-anonymity-based algorithms for graph anonymization. In: Proceedings of the MDAI, Girona. Springer, pp 197–209

Cheng J, Fu AW-C, Liu J (2010) K-isomorphism: privacy preserving network publication against structural attacks. In: Proceedings of the SIGMOD, Indianapolis, pp 459–470

Chester S, Srivastava G (2011) Social network privacy for attribute disclosure attacks. In: Proceedings of the ASONAM, Kaohsiung, pp 445–449

Chester S, Gaertner J, Stege U, Venkatesh S (2012a) Anonymizing subsets of social networks with degree constrained subgraphs. In: Proceedings of the ASONAM, Istanbul, pp 418–422

Chester S, Kapron BM, Ramesh G, Srivastava G, Thomo A, Venkatesh S (2012b) Why Waldo befriended the dummy? k-anonymization of social networks with pseudo-nodes. Soc Netw Anal Min, 3(3):381–399

Chester S, Kapron BM, Srivastava G, Venkatesh S (2012c) Complexity of social network anonymization. Soc Netw Anal Min, 3(2):151–166

Dwork C (2008) Differential privacy: a survey of results. In: Proceedings of the TAMC, Xi'an, pp 1–19

Erdos P, Gallai T (1960) Gráfok elöírt fokszámú pontokkal. Matematikai Lapok 11:264–274

Hay M, Miklau G, Jensen D, Weis P, Srivastava S (2007) Anonymizing social networks. Amherst technical report, University of Massachusetts

Hay M, Li C, Miklau G, Jensen D (2009) Accurate estimation of the degree distribution of private networks. In: Proceedings of the ICDM 2009, Miami, pp 169–178

Karwa V, Raskhodnikova S, Smith A, Yaroslavtsev G (2011) Private analysis of graph structure. PVLDB 4(11):1146–1157

Kasiviswanathan SP, Nissim K, Raskhodnikova S, Smith A (2013) Analyzing graphs with node differential privacy. In: Proceedings of the TCC, Tokyo, pp 457–476
Li N, Li T, Venkatasubramanian S (2007) t-closeness: privacy beyond k-anonymity and l-diversity. In: Proceedings of the ICDE, Istanbul, pp 106–115
Lui K, Terzi E (2008) Towards identity anonymization on graphs. In: Proceedings of the SIGMOD, Vancouver, pp 93–106
Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M (2007) L-diversity: privacy beyond k-anonymity. TKDD 1(1):52
Meyerson A, Williams R (2004) On the complexity of optimal K-anonymity. In: Proceedings of the PODS, Paris, pp 223–228
Narayanan A, Shmatikov V (2009) De-anonymizing social networks. In: Proceedings of IEEE symposium on security and privacy, Oakland, pp 173–187
Nissim K, Raskhodnikova S, Smith A (2007) Smooth sensitivity and sampling in private data analysis. In: Proceedings of the STOC, San Diego, pp 75–84
Srivatsa M, Hicks M (2012) Deanonymizing mobility traces: using social network as a side-channel. In: Proceedings of the ACM conference on computer and communications security, Raleigh, pp 628–637
Sweeney L (2002) k-anonymity: a model for protecting privacy. Int J Uncertain Fuzziness Knowl Based Syst 10(5):557–570
Thompson B, Yao D (2009) The union-split algorithm and cluster-based anonymization of social networks. In: Proceedings of the ASIACCS, Sydney, pp 218–227
Wu W, Xiao Y, Wang W, He Z, Wang Z (2010) k-symmetry model for identity anonymization in social networks. In: Proceedings of the EDBT, Lausanne, pp 111–122
Ying X, Pan K, Wu X, Guo L (2009) Comparisons of randomization and K-degree anonymization schemes for privacy preserving social network publishing. In: Proceedings of the SNA-KDD, Paris. Article #10, 10 pages
Zhou B, Pei J (2008) Preserving privacy in social networks against neighborhood attacks. In: Proceedings of the ICDE, Cancun, pp 506–515

Anonymization of Data

▸ Ethics of Social Networks and Mining

Applications

▸ NodeXL: Simple Network Analysis for Social Media

Archives

▸ Sources of Network Data

Art About Networks

▸ Arts and Humanities, Complex Network Analysis of

Arts and Humanities, Complex Network Analysis of

Isabel Meirelles[1], Maximilian Schich[2], and Roger Malina[2]
[1]Department of Art + Design, College of Arts, Media and Design, Northeastern University, Boston, MA, USA
[2]Arts & Technology, School of Arts and Humanities, The University of Texas at Dallas, Richardson, TX, USA

Synonyms

Art about networks; Networks in art; Networks in culture; Networks in the humanities; Research in network visualization

Glossary

Artificial Life (Sometimes A-Life) Term originally coined by Christopher Langton to study "life as it could be" based on the use of simulations, computer models, robotics, and more recently synthetic biology. A-Life Art are artworks which use A-Life science and technologies
Humanistic Inquiry Methodologies employed in traditional humanities disciplines
Scientometrics The science of measuring and analyzing science

Definition

Complex network-related research and creative practice in the arts and humanities examine emerging structure, dynamics, and evolution of connections between concepts, objects, individuals, locations, and events. To this end, discipline-specific perspectives, ranging from vigorous humanistic inquiry to free artistic expression, are facilitated with scientific approaches from relevant areas including graph theory in mathematics, the physics of complex networks, complexity science, computer science, and information visualization.

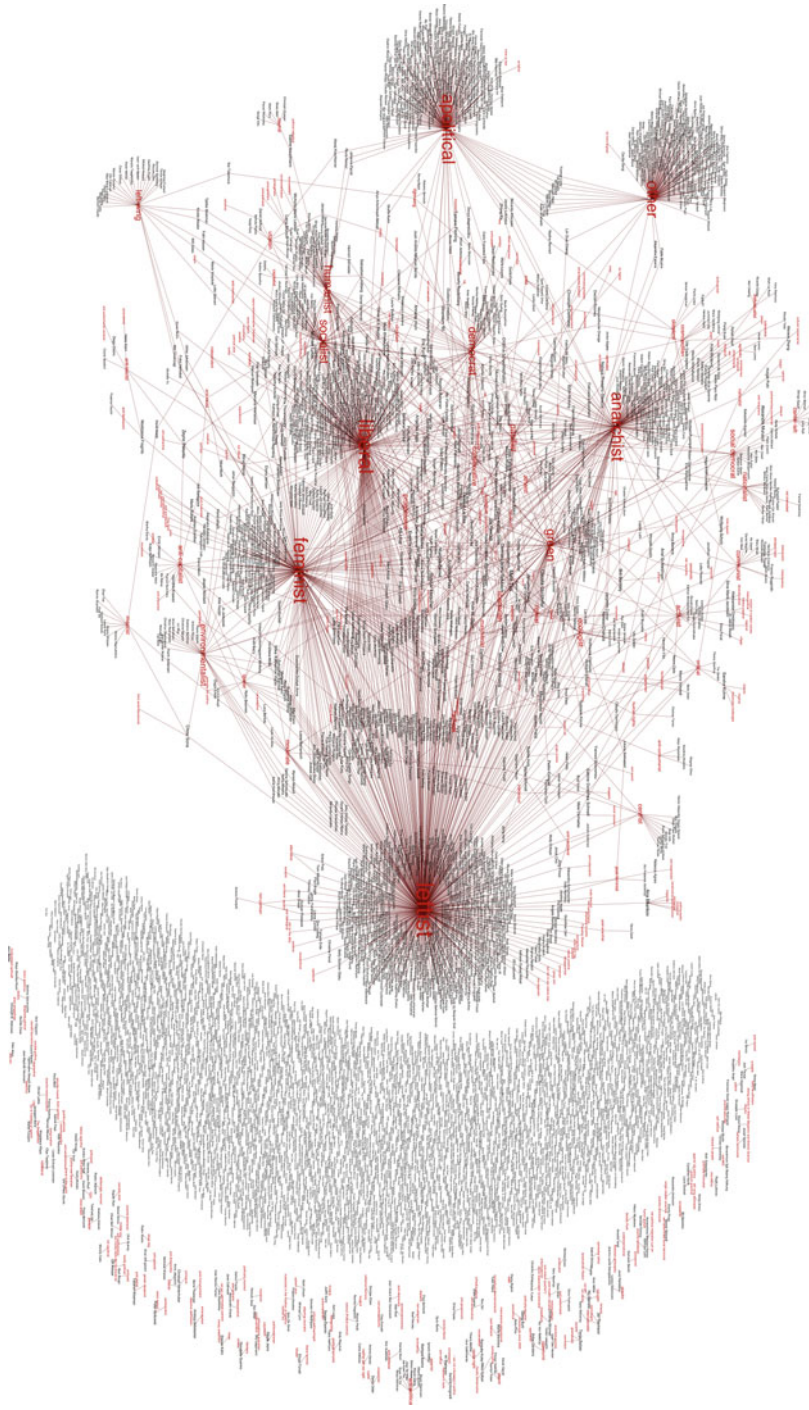
Previous work in the area of arts, humanities, and complex networks can be organized into three general groups with significant overlap: (a) networks in art, the humanities, and culture; (b) art about networks; and (c) research in network visualization.

Networks in art, the humanities, and culture are a rapidly growing topic of research, whose roots go back well into the twentieth century, similar to the genesis of social network analysis. The recent explosive availability of data and quantification that goes along with it have propelled a qualitative change in research practice, a phenomenon shared with other disciplines from social to natural sciences. Complex networks, while still being a conceptual construct, are a tangible dimension of data that is subject to measurement, visualization, modeling, simulation, and ultimately prediction. As a consequence, complex networks allow for a vigorous integrated approach to art, the humanities, and culture that combines quantitative and qualitative research practice into a single process.

Social, natural, and computer scientists play a pivotal role in this process, as their methods and tools are adopted more widely in subfields of arts and humanities that were dominated by theoretical debate throughout the last decades. Furthermore, art, the humanities, and culture lie at the crossroads of many lines of thought in network science, as most relevant datasets contain a number of node and link types that pertain to existing specializations of network research (e.g., Schich 2010). The resulting diversity of research is striking as it involves collaborations of archaeologists; (art) historians; biologists; cognitive scientists; computer scientists; economists; educators; entrepreneurs; information designers; musicologists; researchers in language, literature, and scientometrics; social scientists; physicists; and many more (see, e.g., the contributions in Schich et al. 2012).

Art about networks refers to creative work by artists including collaboration with philosophers, programmers, and scientists. In recent years, artists have invested a number of fields of science and research as artistic domains. Steve Wilson in his book Information Arts (2003) develops at lengths the aims and methodologies of these artists in fields ranging from molecular biology to astrophysics. Artists were very early adopters of the field of complexity science and Artificial Life, as made visible by a number of conferences and competitions. In Artificial Life, of particular note is VIDA Artificial Life started in 1999 (http://www.fundacion.telefonica.com/en/que_hacemos/conocimiento/exposiciones/actuales/vida_1999-2012.htm (accessed 20 March 2013)); see also http://malina.diatrope.com/2012/01/03/artificial-life-in-art-and-science-25-years-later-new-worlds-and-virtual-humans/ (accessed 20 March 2013).

Artists were very early adopters not only of the Internet in general but also of data mining, complexity science, and complex network science. See, for instance, the work of George Legrady and the Experimental Visualization Lab (http://www.georgelegrady.com/ (accessed 20 March 2013)) and his essay in Schich et al. (2012). A number of artists have exploited large art databases and their structures. Spanish artist Muntadas has an extensive work The File Room (http://www.thefileroom.org/ (accessed 20 March 2013)) which in particular seeks to understand the structure of censorship. Burak Arikan (http://burak-arikan.com/ (accessed 20 March 2013)) has used complex network visualizations to both create visualizations of cultural networks and display such visualizations as artworks in themselves (Fig. 1). In the field of



Arts and Humanities, Complex Network Analysis of, Fig. 1 Network map of artists and their political inclinations, 7th Berlin Biennale, 2012, by Burak Arikan

music, there are many examples. A pioneering group of computer musicians, the Hub, was formed in 1986 (http://hub.artifactrecordings.com/ (accessed 20 March 2013)) to develop computer network music that exploits and makes audible the properties of network structures.

Research in network visualization is an intrinsic component of complex network research. Visual representations and analytical tools have the potential to augment our reasoning capacities by facilitating perceptual inference, discover patterns, and expand working memory (e.g., Meirelles 2013). Research in network visualization is rooted in the arts and humanities and has emerged as a crucial component in other scientific domains due to the growing complexity and amount of data available. Development of visualization tools and techniques can be roughly divided into three areas:

- **Novel visualization techniques**: involves the work of computer scientists, information designers, and usability specialists collaborating with scholars in developing custom-made tools. Tackling problems specific to arts and humanities data, these techniques have the potential of becoming part of a broader visual analytics tool-kit. An example is the work of Nathalie Riche at Microsoft Research who has developed techniques for visualizing heterogeneous networks (with multiple node and link types) and networks evolving over time (http://research.microsoft.com/en-us/um/people/nath/ (accessed 20 March 2013)).
- **Programming languages**: involves the development of environments that can be used by other researchers to build visual displays of their data by means of writing code. For example, D3.js is a JavaScript library for visualizing data and manipulating the document object model. Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer at Stanford University initially developed it in 2011 (http://d3js.org/ (accessed 20 March 2013)). Processing is another widely used programming environment created by Ben Fry and Casey Reas in 2001 while at the MIT Media Lab (http://processing.org (accessed 20 March 2013)).
- **Visualization applications**: involves the development of visual tools that enable nonprogrammer researchers to visualize their data without the need to code. The first and most successful of such applications is the website Many Eyes by IBM, originally developed by Fernanda Viégas and Martin Wattenberg in 2008 (http://www-958.ibm.com/software/data/cognos/manyeyes/ (accessed 20 March 2013)).

In terms of method, the common approach towards arts, humanities, and complex networks can be characterized as a process, where typical goals are the observation and anticipation of patterns and regularities, in contrast to the natural and social sciences where the focus lies on experimentation, modeling, simulation, and prediction. In short, the details of a typical task sequence include a feedback cycle of data acquisition, data cleaning, measurement, visualization, and effective communication. **Data acquisition** typically includes negotiation with data providers and the use of a variety of query languages. **Data cleaning** includes data normalization towards a network format (e.g., Segaran 2009), data filtering and cloaking techniques, as well as further processing using applications (e.g., Excel, OpenRefine), command line tools (e.g., Grep, Awk), and programming languages (e.g., R, Python). **Extraction of inherent order** (i.e., measurement and visualization) involves standard measures of network science such as degree distributions, clustering, and component structure (the growing list of which is easily found in the literature, e.g., Barabási et al. 2006; Newman 2010; Easley and Kleinberg 2010). Particularly relevant areas include network communities (Porter et al. 2009), temporal structure in networks (Holme and Saramäki 2012), and networks with multiple node and link types (Schich 2010). **Visualization** makes extensive use of node-link diagrams, adjacency matrices, and statistical diagrams, using applications (e.g., Cytoscape, Gephi) and/or coding environments (e.g., D3,

Processing). Dissemination focuses less on traditional monographs and more on scientific journals, conference papers, posters, and video presentations.

Cross-References

- ▸ Analysis and Planning of Urban Networks
- ▸ Analysis and Visualization of Dynamic Networks
- ▸ Clustering Algorithms
- ▸ Collection and Analysis of Relational Data from Digital Archives
- ▸ Combining Online Maps with Text Analysis
- ▸ Community Detection, Current and Future Research Trends
- ▸ Community Evolution
- ▸ Community Identification in Dynamic and Complex Networks
- ▸ Data Mining
- ▸ Ethical Issues Surrounding Data Collection in Online Social Networks
- ▸ Flickr and Twitter Data Analysis
- ▸ Gephi
- ▸ GUESS
- ▸ Linked Open Data
- ▸ Microtext Processing
- ▸ Modeling and Analysis of Spatiotemporal Social Networks
- ▸ Multiplex Networks
- ▸ Network Representations of Complex Data
- ▸ Networks in Geography
- ▸ NodeXL: Simple Network Analysis for Social Media
- ▸ RDF
- ▸ Scholarly Networks Analysis
- ▸ Semantic Social Networks Analysis
- ▸ Sentiment Analysis in Social Media
- ▸ Sources of Network Data
- ▸ SPARQL
- ▸ Spatial Networks
- ▸ Temporal Networks
- ▸ UCINET
- ▸ Visual Methods and Tools for Social Network Analysis
- ▸ Visualization of Large Networks
- ▸ Web Science

References

Barabási AL, Newman MEJ, Watts DJ (eds) (2006) The structure and dynamics of networks. Princeton University Press, Princeton

Easley D, Kleinberg J (2010) Networks, crowds, and markets. Cambridge University Press, New York

Holme P, Saramäki J (2012) Temporal networks. Physics Report, 519(3):97–125

Meirelles I (2013) Design for information. An introduction to the histories, theories, and best practices behind effective information visualizations. Rockport Publishers, Beverly

Newman MEJ (2010) Networks. An introduction. Oxford University Press, New York

Porter M, Onnela JP, Mucha PJ (2009) Communities in 278 networks. Notices of the American Mathematical Society 56(9):1082–1197/1164–1166

Schich M (2010) Revealing matrices. In: Steele J, Iliinsky N (eds) Beautiful visualization. O'Reilly, Sebastopol

Schich M, Meirelles I, Malina R (eds) (2012) Arts, humanities, and complex networks [Kindle edn.]. MIT, Cambridge

Segaran T (2009) Connecting data. In: Segaran T, Hammerbacher J (eds) Beautiful data. O'Reilly, Sebastopol

Wilson S (2003) Information arts: intersections of art, science, and technology. MIT, Cambridge

ASP

- ▸ Server-Side Scripting Languages

ASP.NET

- ▸ Server-Side Scripting Languages

Assortative Behavior

- ▸ Demographic, Ethnic, and Socioeconomic Community Structure in Social Networks

Auction Fraud

▸ Fraud Detection Using Social Network Analysis, a Case Study

Augmented Matrix

▸ Matrix Algebra, Basics of

Authority

▸ Time- and Event-Driven Modeling of Blogger Influence

Automatic Document Topic Identification

▸ Automatic Document Topic Identification Using Social Knowledge Network

Automatic Document Topic Identification Using Social Knowledge Network

Mostafa M. Hassan, Fakhreddine Karray, and Mohamed S. Kamel
Department of Electrical and Computer Engineering, Centre for Pattern Analysis and Machine Intelligence (CPAMI), University of Waterloo, Waterloo, ON, Canada

Synonyms

Automatic document topic identification; Clustering; Ontology; Social knowledge network; Wikipedia

Glossary

ADTI Stands for automatic document topic identification

Ontology "A model for describing the world, that consists of a set of types (concepts), properties, and relationship types" (Garshol 2004)

SKN Stands for social knowledge network

WHO Stands for Wikipedia Hierarchical Ontology

TF-IDF A term weighting methodology that is commonly used in text mining and in information retrieval. It stands for term frequency-inverse document frequency

hi5 An online social networking website

RDF Stands for Resource Description Framework. It is a method of representing information to facilitate the data interchange on the Web

ASR Stands for automatic speech recognition

NMI Stands for normalized mutual information. It is a well-known document clustering performance measure

NMF Stands for nonnegative matrix factorization. Nonnegative matrix factorization is a family of algorithms that tries to factor a matrix X into two matrices Y and Z, with the property that all three matrices have no negative elements

Clustering Is the process of assigning each input pattern to a group (cluster), such that each group contains similar patterns

Taxonomy Is the division of concepts or topics into ordered groups or categories

Definition

Document topic identification or indexing is usually used to refer to the task of finding relevant topics for a set of input documents (Coursey and Mihalcea 2009; Medelyan et al. 2008). It is used in many different real applications, such as improving retrieval of library documents pertaining to a specific topic. It could also be used to improve the relevancy of search engine results, by categorizing the search results according to

their general topic and giving users the ability to choose the domain which is more relevant to their needs.

Overview

Nowadays, social networks are being frequently used by 1.73 billion people or more in 2013 (European Travel Commission 2013). Different social networks with different levels of complexity and popularity have been developed recently providing rich media and knowledge sources. These sources include, but are not limited to, media sharing sources like YouTube and Flickr, micro-blogging like Twitter, general media such as Facebook and Google+, content tools such as Hi5, blog and journal such as Blogger, connection tools such as LinkedIn, and authoritative sources (Korfiatis et al. 2006) such as Wikipedia. Wikipedia can be seen as a social knowledge network (SKN), where the users share their knowledge together collaboratively to build some sort of knowledge repository. It is a free online encyclopedia whose contents are written collaboratively by a large number of voluntary contributors around the world. Wikipedia webpages can be edited freely by any Internet user. This type of SKN leads to a rapid increase of its good-quality contents, as any potential mistakes are quickly corrected within the collaborative environment. Wikipedia coverage of topics has become as comprehensive as other well-known encyclopedias such as Britannica (Giles 2005), with reasonable accuracy.

This paper introduces a novel approach for identifying document topics using social knowledge network. In this approach, human background knowledge in the form of SKN is utilized to help in automatically finding the best matching topic for input documents. There are several applications for automatic document topic identification (ADTI). For example, ADTI can be used to improve the relevancy of search engine results by categorizing the search results according to their general topic. It can also give users the ability to choose the domain which is most relevant to their needs. The proposed ADTI technique extracts background knowledge from a human knowledge source, in the form of a SKN, and stores it in a well-structured and organized form, namely, an ontology. This ontology encompasses both ontological concepts and the relations between these concepts and is used to infer the semantic similarity between documents, as well as to identify their topics.

Document topics are among this valuable information that needs to be extracted for several applications. Recently, many approaches in the literature employ the use of background knowledge to improve performance of document topic identification. Coursey and Mihalcea (2009) and Coursey et al. (2009) proposed an unsupervised method based on a biased graph centrality algorithm, applied to a large knowledge graph built from Wikipedia. They mapped the input documents to Wikipedia articles based on the similarity between them, and then they used their proposed biased graph centrality algorithm to find the matching topics. Similar is the work presented by Schönhofen in (2009), but instead of using the Wikipedia full articles' contents, they used the articles' titles to match the input documents to the Wikipedia categories. Huynh et al. (2009) suggested an update to the work proposed by Schönhofen in (2009): they added the use of the hyperlinks in Wikipedia in articles' titles to improve topic identification.

Janik and Kochut (2008a, b) have used the Wikipedia RDF (which is defined in Auer and Lehmann (2007)) to create their ontology. They then transfer the document text into a graph structure, employing entity matching and relationship identification. The categorization is based on measuring the semantic similarity between the created graph and the categories defined in their ontology.

Basic Methodology

Extracting an Ontology from a Social Knowledge Network

This section introduces the approach to building a Wikipedia Hierarchical Ontology (WHO)

from the Wikipedia knowledge repository. We use this ontology to utilize the knowledge stored in Wikipedia for document representation. We assume that each Wikipedia category represents a unique topic. These topics are considered to be the basic building blocks of the ontology; we refer to them as concepts. Wikipedia categories are organized in a hierarchical manner, so that the root concepts represent abstract ideas. Reciprocally, the leaf concepts represent very specific ideas. This reflects the world knowledge in different domains with different level of granularity. Each category (concept) is associated with a collection of Wikipedia articles that describe and present different ideas related to this concept. Using these articles, we can extract the set of terms that represent each concept. Furthermore, we associate a weight with each of these terms, which expresses how that term contributes to the meaning of that concept. These weights are calculated based on the frequency of occurrence of these terms in the articles under that concept. We construct the concept-term mapping matrix M as follows:

$$M = \begin{bmatrix} \text{tf-icf}_{1,1} & \dots & \text{tf-icf}_{1,j} & \dots & \text{tf-icf}_{1,l} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \text{tf-icf}_{i,1} & \dots & \text{tf-icf}_{i,j} & \dots & \text{tf-icf}_{i,l} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \text{tf-icf}_{n,1} & \dots & \text{tf-icf}_{n,j} & \dots & \text{tf-icf}_{n,l} \end{bmatrix} \quad (1)$$

where $\text{tf-icf}_{i,j}$ is the weight of a term i in the concept j, n is the total number of concepts that we have extracted, and l is the total number of terms that we have found for all the extracted concepts. This describes the basic idea of the creation process of Wikipedia Hierarchical Ontology. The details of the algorithm have been omitted due to space limitations. For more details, we refer to our work in Hassan (2013).

Automatic Document Topic Identification Using WHO

There are some different tasks in text mining that fall under document indexing, including document tagging and keyphrase extraction. Medelyan (2009) has classified these tasks according to two aspects: first, the source of the terminology that the topics are extracted from and, second, the number of topics that can be assigned to the documents. She has set three different values for the first aspect, which are vocabulary-restricted, document-restricted, and no restriction. In vocabulary-restricted, the source of the topic is usually some sort of background knowledge such as a thesaurus or a structured glossary. In document-restricted, the source of the topics is the input documents themselves, where we try to select the most representative terms from these documents. In no restriction, the topics that are assigned to the documents are selected freely with no restriction to a knowledge source.

The number of the topics aspect has three different values, which are very few (main topics), detailed topics, and all possible topics. In main topics, the number of topics is limited to a small set of topics (usually less than a 100). In detailed topics, more specific topics are included, which makes the number of topics much bigger (usually from hundreds to thousands), and usually more than one topic is assigned to each document. In all possible topics, the number of topics is limited to all the terms found in the input documents. This type is usually called full-text indexing.

In order to complete this classification of tasks, a third aspect would be added to these two aspects, which is the learning paradigm. As is well known, learning paradigms can be classified into three types: supervised learning, semi-supervised learning, and unsupervised learning. Table 1 shows a list of possible topic indexing tasks classified based on these three aspects.

Cluster labeling is a complementary task to document clustering, in order to be a complete topic indexing task. It is the process of selecting a representative label (topic) for each cluster obtained from the document clustering process (Popescul and Ungar 2000). The source of these labels is usually from the terms and/or the phrases which the input documents are indexed with. One of the approaches used for cluster labeling is to use a feature selection technique, such as mutual information and chi-squared feature selection, to differentiate cluster labeling.

Automatic Document Topic Identification Using Social Knowledge Network, Table 1 Topic indexing tasks

Task	Source of terminology	Number of topics	Learning paradigm
Document classification	Vocabulary-restricted	Main topics only	Supervised
Document clustering with cluster labeling	Document-restricted	Main topics only	Unsupervised
Term assignment	Vocabulary-restricted	Detailed topics	Supervised/unsupervised
Keyphrase extraction	Document-restricted	Detailed topics	Supervised/unsupervised
Document tagging	Unrestricted	Detailed topics	Supervised/unsupervised

Term assignment, or subject indexing, is the process of finding the best representative topics for each document. The source of terminology is usually extracted from an external thesaurus, unlike the keyphrase extraction task where the main goal is to extract the most distinct phrases appearing in the documents. Lastly in document tagging, or in short tagging, tags can be chosen freely without any formal guideline. Usually the last three tasks, term assignment, keyphrase extraction, and tagging, are referred to as document topic identification. Although they differ in the source of terminology, they more or less do the same task, which is assigning each document a set of representative terms/phrases/tags. Also, these three tasks have implementations for both learning paradigms.

Automatic document topic identification (ADTI) can be seen as an optimal assignment problem where given a set of topic labels, $\mathcal{L} = \{b_1, b_2, \ldots, b_p\}$, which has been marked as being "of interest," and a set of input documents, $\mathcal{D} = \{d_1, d_2, \ldots, d_m\}$, it is required to assign each input document d_i to one of these topics b_j.

The word "identification" in ADTI means finding the best match between the input document set and the input topic list. In contrast to other approaches, the list of topics is a known entity in our approach, which means that there is no need to predict them (Hassan 2013). To better understand the difference, let us consider the following example: assume that we are interested in the topics economics, politics, and sports and we have a document declaring "Barack Obama is the new President of the United States." Usually other topic indexing approaches will use some form of background knowledge to predict the topic of this document, and they might give the following topics as the best matching topics: "US presidential election," "Presidents of the USA," and so on. These topics usually will be very specific to this document. In contrast, our approach will try to identify the most relevant topic from the list of given topics and will identify it as "politics."

ADTI can be used in many different real applications, for example, improving retrieval of library documents pertaining to a certain topic. It can also be used to improve the relevancy of search engine results, by categorizing search results according to their general topic and giving users the ability to choose the domain which is more relevant to their needs. It is also needed for an organization like a news publisher or news aggregators, where they want to automatically assign each news article to one of the predefined news main topics. Similarly, it can be applied for digital libraries to assign each new article to one of the predefined lists of topics. ADTI can also be used to improve the output of automatic speech recognition (ASR) system by selecting the language model that is most relevant to the topic of the speech input.

Automatic Document Topic Identification Methodology

The idea of ADTI is to map topics and input documents to the same space and then find the closest topic to each input document. Here the common space between input documents and topics is the term space. This mapping process can be split into three different steps. The first step is to extract the representative concept vector for each input topic from the constructed ontology. The second step is to use the concept taxonomy in WHO to enrich the topics' representations.

The last step is to use extracted topic vectors to identify documents' topics. The following subsection discusses the different ways to extract the representative concept vector examined in the proposed approach.

Extracting Representative Concepts for Input Topics

As mentioned earlier, one of the inputs of the ADTI application is a list of topics of interest, to which we want to classify input documents. In this module, we try to find the matching list of concepts to these topics. Usually, this process is done by direct matching of topics' names and concepts' names. The problem is that sometimes there is no direct match between some topics and concepts or the direct matching is not so accurate. For example, sometimes the given topic of interest is "technology," but the best matching concept describing the input document set is "information technology." Or as an example of a topic with no direct matching concept, consider "economy." The best matching concept label for this topic is "economics" not "economy." To resolve this problem, applied manual matching is proposed. In this approach, we use the data set provider's experience about the input document set to find the matching concepts. Given the list of the ontology concept labels $\mathcal{L}$, for each given topic label, we start by searching that list for all available concept labels that contain this topic label. The output list for each topic label is sorted based on the orthographic similarity between the topic label and the list of concepts' labels. We pass on these lists of concepts' labels to the data set provider to select the best matching concept(s) for each topic, based on their experience with the input data set. The main drawback of this approach is that it makes the whole technique partially manual, as humans are still needed to select the matching concepts.

After identifying the matching concepts either manually or automatically, we construct the topic-concept map matrix P. Each row of this matrix represents a topic and each column represents a concept. In other words, if the number of the input topics is p and the total number of concepts extracted in WHO is n, then P size will be $p \times n$. Each element of this matrix, $P_{i,j}$, is equal to one when the concept j is considered a representative concept for the topic i, according to the list matching concepts extracted in the previous step, and is zero otherwise. Notice that the matrix P is too sparse and $n \gg p$.

Enhance Topic Representation by Utilizing Ontology Taxonomy

In some cases, this concept-term matrix P does not suffice to represent the topic well. This situation occurs often in abstract topics where the number of relevant articles in Wikipedia is too small; hence, the representing concept vector will have a very small list of representing terms. For example, the Wikipedia category "computer science" is only covered by only four articles. Consequently this will affect the identification of the topic. As mentioned previously, each concept in our extracted ontology, WHO, has a conceptual relationship to other concepts which are represented in the ontology taxonomy, in addition to its representative terms' vectors. We utilize the hierarchical structure of concepts to increase the amount of information that is associated with each topic; this also increases the generality of the topics. This is done by augmenting to the concept-term mapping vector of each topic of interest, the set of term mapping vectors associated with each concept under the hierarchy of that topic of interest. For example, if the topic of interest is "Biology," we add term mapping vectors that are associated with the concepts "Anatomy," "Botany," "Zoology," etc., to its associated concept-term mapping vector. This includes not only the directly connected concepts to this topic but also all the topics in the hierarchy down to a specific level l.

Although the previous augmentation of sub-concepts' information to the main concept increases the amount of information that is associated with the main topic, it also adds some noise. Noise here means a subset of the information that is related to the subtopic, but is not related to the main topic. Since the relatedness between the main concept and its sub-concepts decreases as they get farther from it, the quantity of noise increases as we go deeper

into the hierarchy of concepts. To resolve this problem, we introduced a penalty function Pen to penalize the information coming from the sub-concepts as follows:

$$\text{Pen} = e^{-L}$$

The penalty term is a function of the level of the subtopic, so that as one goes deeper into the hierarchy, the penalty value increases (Hassan 2013).

Identification Approach

The last step after selecting the list of concepts that will represent the given list of topics is utilizing these extracted concepts to identify the topics of input documents. The nearest centroid approach is proposed for identifying documents' topics. This approach is very similar to the nearest centroid classification approach. The idea of the nearest centroid classification is to create a prototype, a centroid in this case, for each class, to then use this prototype to classify the input documents by assigning each input document to the closest prototype. Similarly, we define a prototype for each topic, to use it in identifying input documents' topics. We use the constructed topic-concept map matrix P, defined previously, and the WHO concept-term mapping M, defined in (1), to extract the matrix Q that represents the topics of interest prototypes as follows:

$$Q = PM \tag{2}$$

The size of the output matrix Q is $p \times l$, where l is the total number of terms in WHO and p is the number of the input topics. Notice that each row in this matrix represents a topic which is a vector of summation of all its representative concepts' vectors. We can also notice that the size of the matrix Q is much smaller than the original matrix M as $p \gg n$, where n is the total number of concepts in WHO.

The next step is to construct the document-term matrix A from input documents according to the conventional VSM representation as shown in (3):

$$A = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m1} & w_{m2} & \dots & w_{mn} \end{bmatrix} \tag{3}$$

The size of the output matrix A is $m \times n$, where m is the total number of input documents and n is the total number of terms in the input document set. Then we remove all terms in A which are not defined in Q, as we consider them to be out-of-vocabulary terms and vice versa, and all terms found in Q and not defined in A are considered to be out-of-interest terms and are removed from Q. Then we normalize each vector of Q to be a length of 1. Hence, the new mapping matrix will be

$$\hat{Q} = L^{-1}Q \tag{4}$$

where L is a diagonal matrix whose elements are the lengths of the matrix Q. We can use the following equation to calculate the document-topic similarity S matrix as follows:

$$S = A\hat{Q}^T \tag{5}$$

where each row S_i,: in the matrix S represents the similarity between a document i and the list of the topics of interest. Then we can define the identification function $h(x)$ as follows:

$$h(x) = \underset{e \in 1:p}{\text{argmax}} S_{x,e} \tag{6}$$

where $h(x)$ is a function that takes an index of a document x and returns the index of the most similar topic, e.

Experimental Results

The following subsections describe the experimental results of applying ADTI using WHO.

Experiment Setup

Eight different benchmark data sets that fit ADTI requirements have been selected. These data sets are split into groups or classes, each of which explicitly represents a specific topic. Most of these data sets have been previously used by Zhao et al. (2005) to evaluate the performance of

Automatic Document Topic Identification Using Social Knowledge Network, Table 2 Summary of data sets used to evaluate the performance of ADTI: m is the number of documents, n is the total number of terms in all documents, n' is the number of used terms where the other terms that are not present in WHO are ignored, and k is the number of topics

ID	Source	m	n	n'	Topics	k
k1b	WebACE	2,340	21,839	19,021	Business, entertainment, health, politics, sports, tech	6
wap	WebACE	1,311	8,460	7,293	People, television, health, media, art, film, business, culture, music, politics, sports, entertainment, industry, multimedia	14
reviews	San Jose Mercury (TREC)	4,069	36,746	32,921	Food, movie, music, radio, restaurant	5
sports	San Jose Mercury (TREC)	8,580	27,673	24,115	Baseball, basketball, bicycle, boxing, football, golf, hockey	7
hitech	San Jose Mercury (TREC)	2,301	22,498	20,268	Computer, electronics, health, medical, research, technology	6
mm	San Jose Mercury (TREC)	2,521	29,973	27,262	Movie, music	2
bbc	BBC News	2,225	9,636	7,992	Business, entertainment, politics, sport, tech	5
bbc-sports	BBC Sport	737	4,613	3,804	Athletics, cricket, football, rugby, tennis	5

different document clustering algorithms. Table 2 summarizes the properties of these data sets.

In all data sets, terms which appear in only one document or do not appear in our ontology, WHO, are removed. Hence, we have n′ in Table 2 to represent the total number of terms after removing these terms. Then, the term weights inside documents are normalized according to the TF-IDF weighting. Lastly, all documents are normalized to represent unit vectors in different directions in term space.

Comparing ADTI to Document Clustering

In this experiment, ADTI is compared with four different standard and state-of-the-art clustering techniques:

- **K-means clustering (K-MEANS):** The spherical k-means version is used since the documents are represented as vectors, and the distance measure used is the cosine similarity. We used the MATLAB implementation of the Lloyd's algorithm (Lloyd 1982). As is well known, k-means clustering output depends on the initial step. Hence, the cluster assignments are changed for the different runs. So, we applied the k-means clustering 10 times. We report here the mean and standard deviation of these runs.
- **Hierarchical clustering:** Two different linkage methods are used in hierarchical clustering, average (***HIC-AVG***) and complete (***HIC-CMP***) linkage. We also used the MATLAB implementation for the hierarchical clustering. As the hierarchical clustering does not depend on any initial conditions, there is no need to apply it multiple times.
- **Spectral clustering (*SC*):** We have used the cosine similarity as the measure of similarity documents. Regarding the Laplacian matrix normalization, we have used the Ng et al. (2002) proposed approach: $L = I - D^{-1/2}SD^{-1/2}$. We have used the same k-means implementation for clustering. As this method depends on k-means approach, the cluster assignments are changed for the different runs. We applied this algorithm 10 times as for k-means, and we report the mean and standard deviation of these runs.
- **NMF clustering (*NMF*):** We have used the Xu et al. (2003) and Xu and Gong (2004) approach for NMF clustering. As factorization of matrices is generally nonunique,

NMF is considered a nondeterministic approach. Hence, we applied the NMF clustering algorithm 10 times and we report the mean and standard deviation of these runs.

After applying ADTI and the document clustering, the output labels are compared against ground-truth labels to evaluate each method. In the case of document clustering, we need to find the mapping between the clusters' labels and the provided ground-truth labels. The Hungarian algorithm is used to find this matching (Kuhn 2005). In the case of ADTI, there is no need to find the matching as ADTI not only partitions the data but also provides the topics of these partitions. Hence, we can match these labels with the provided ground-truth labels. Both approaches were applied to the eight data sets mentioned earlier.

Performance Measures

We have selected some of the most well-known document clustering performance measures. Generally, document clustering has two different sets of performance measures which are internal and external performance measures. It is meaningless to use the clustering internal performance measures with ADTI, as the partitioning between the documents is not based on the pairwise similarity between documents as in document clustering, and the aim of ADTI is to identify the topics of documents, neither to minimize the separation between classes nor to increase the compactness of each class.

We chose three external quality measures to evaluate the performance of the proposed approach: F-measure, purity, and NMI. F-measure evaluates the output accuracy with regard to a given ground truth. It is defined as the *harmonic mean* of the clustering precision and recall. Purity is a measure of the purity of the clusters generated. After mapping each cluster to a class, the purity is defined as that fraction of documents belonging to that class, over the total number of documents in the cluster. Normalized mutual information (NMI) is a well-known document clustering performance measure. It estimates the amount of shared information between the clusters' labels and the categories' labels. It measures the amount of information that can be obtained from the cluster labels by observing the category labels. The higher the value of these measures, the better the obtained output is. A detailed review of these measures can be found in Hassan (2013). The running time, T, of each technique is also measured to compare the efficiency of these approaches.

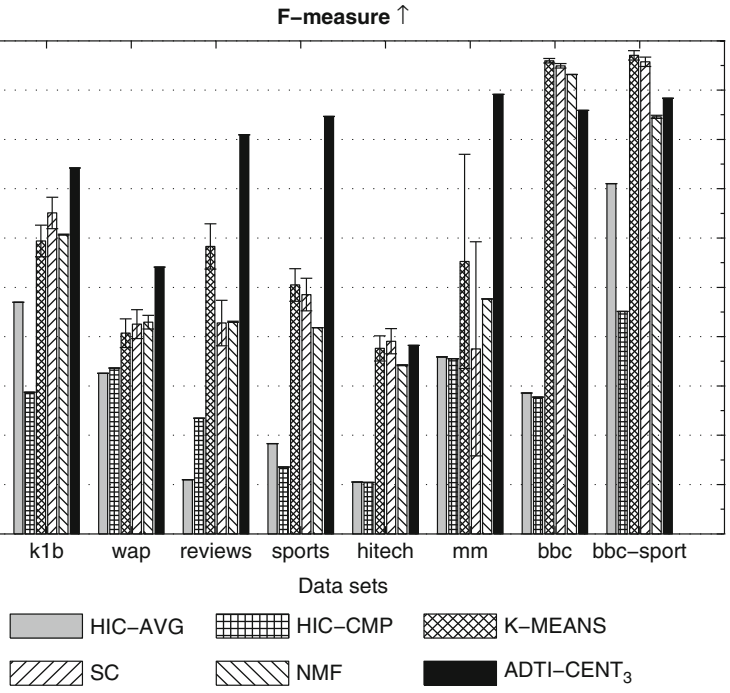
As shown in Fig. 1, ADTI outperforms both hierarchical clustering techniques for all data sets in terms of F-measures. The standard deviation is shown as error bars with each average value for the measure. Note that the hierarchical approaches and the proposed approach have no error bars as they have deterministic outputs.

We can also see that ADTI-$CENT_3$ outperforms the partitional clustering methods in five data sets. In the hitech data set, ADTI-$CENT_3$ has a competitive performance with partitional clustering methods. For bbc and bbc-sports data sets, we can see that ADTI-$CENT_3$ has a competitive performance with NMF and poorer performance than both spectral clustering and k-means. Experiments show that recall measure results are more or less the same performance as F-measure.

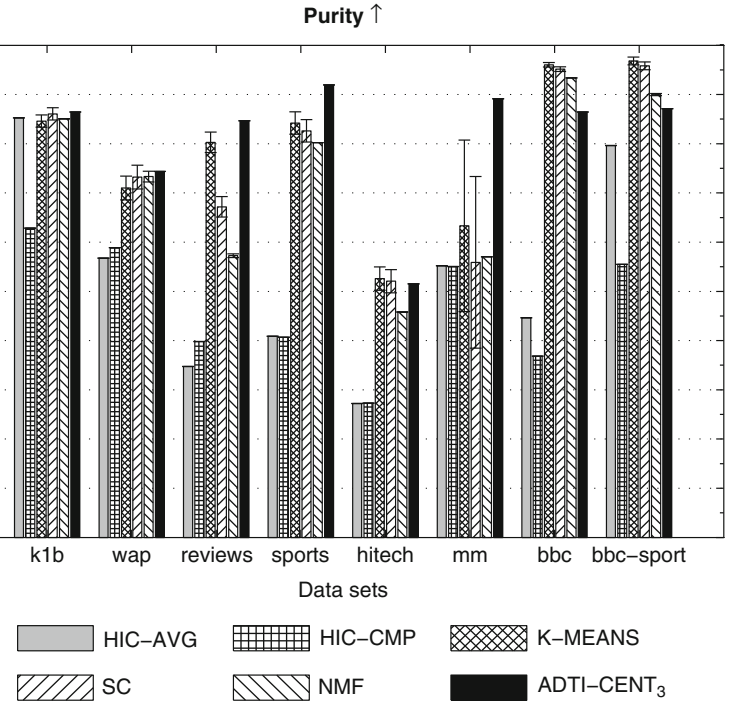
In terms of purity, Fig. 2 shows that the proposed ADTI approach has broadly similar performance to partitional clustering methods, except with the mm data set, where ADTI approach outperforms all clustering approaches. We can also see that the proposed ADTI approach outperforms both hierarchical clustering approaches in all data sets.

Figure 3 shows the output performance comparison with NMI measure. ADTI-$CENT_3$ outperforms the hierarchical clustering methods in all data sets and outperforms all different partitional clustering techniques in five data sets, while it has a very competitive performance with the partitional clustering techniques in second data set, but has poorer performance on the last two data sets.

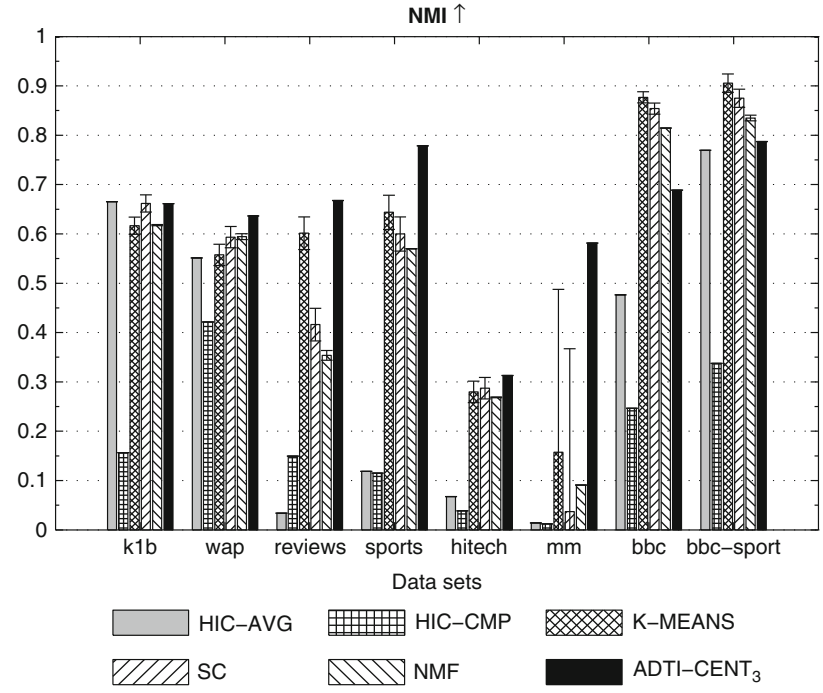
Figure 4 shows the running time comparison of these approaches. We can see that NMF is the slowest approach. As for ADTI approach, it has nearly the same running time to most clustering approaches in almost all data sets.



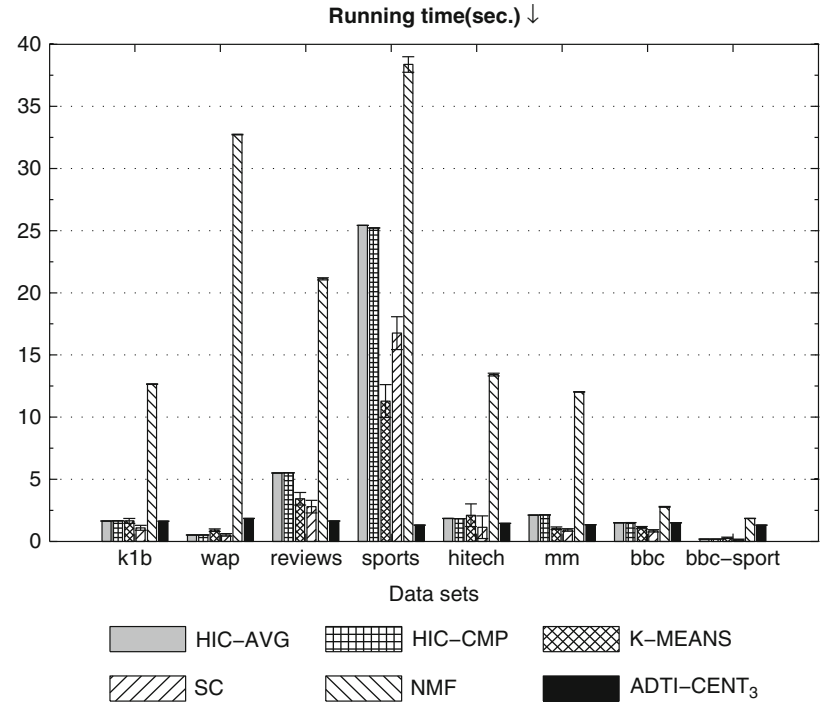
Automatic Document Topic Identification Using Social Knowledge Network, Fig. 1 The output F-measure of the different document clustering methods and ADTI approach for eight data sets



Automatic Document Topic Identification Using Social Knowledge Network, Fig. 2 The output purity of the different document clustering methods and ADTI approach for eight data sets



Automatic Document Topic Identification Using Social Knowledge Network, Fig. 3 The output NMI of the different document clustering methods and ADTI approach for eight data sets



Automatic Document Topic Identification Using Social Knowledge Network, Fig. 4 The output running time of the different document clustering methods and ADTI approach for eight data sets

Automatic Document Topic Identification Using Social Knowledge Network, Table 3 The overall relative performance measures for the different document clustering methods and ADTI approach with level 3

Method	F-measure	Purity	NMI	Time (sec.)
HIC-AVG	0.40 ± 0.21	0.61 ± 0.20	0.43 ± 0.37	0.54 ± 0.29
HIC-CMP	0.35 ± 0.13	0.55 ± 0.13	0.24 ± 0.19	0.54 ± 0.29
K-MEANS	0.79 ± 0.17	0.91 ± 0.10	0.78 ± 0.24	0.56 ± 0.23
SC	0.75 ± 0.23	0.88 ± 0.13	0.73 ± 0.30	0.83 ± 0.34
NMF	0.72 ± 0.19	0.84 ± 0.14	0.70 ± 0.27	0.09 ± 0.09
ADTI-CENT$_3$	0.95 ± 0.04	0.94 ± 0.05	0.88 ± 0.08	0.64 ± 0.32

Key Research Findings

Table 3 shows the overall relative performance measures and running time for the different document clustering methods and ADTI approach where relative performance means the ratio between the performance of a method and the one of the best performing method for a specific measure. The best output performance is shown in bold and the second best is underlined.

As shown in Table 3, the proposed ADTI method outperforms all the different document clustering techniques in overall relative performance measures. In terms of running time, ADTI-CENT$_3$ is better than all clustering except the spectral clustering.

The use of background knowledge to enhance the performance of text mining has been proposed and widely used in different applications. This paper presented a novel approach for automatic document identification (ADTI) approach utilizing the background knowledge in the form of an ontology. This approach encompasses two main modules. The first module is concerned with how to build an organized and structured form of knowledge, ontology, from a different format of knowledge repository. The second module details how to utilize this knowledge structure, an ontology, for a newly defined task.

Future Directions for Research

The proposed ADTI approach can be used in different applications. One of the potential applications we are planning to study is the applicability of the proposed approach to improve the output of an automatic speech recognition (ASR) system. In order to produce the recognized text, an ASR system usually needs to be supplied with a language model. The efficiency of the ASR system is very much dependent on the accuracy of the supplied language model. So if we could know the topic of the speech input, we could supply a more relevant language model. Most of the time we do not know this kind of information in advance; therefore, we provide a generic language model, which leads to a lower accuracy of the ASR system. To overcome this problem, we can provide the output of the ASR system using this kind of generic language model to the document identification system. The document identification system will provide in return the most relevant topic using the inaccurate output of the ASR. Consequently, a more relevant language model can be supplied to get a more accurate result.

Cross-References

▸ Analysis and Mining of Tags, (Micro)Blogs, and Virtual Communities
▸ Ontology Matching
▸ Topic Modeling in Online Social Media, User Features, and Social Networks for
▸ Web Ontology Language (OWL)
▸ Wikipedia Knowledge Community Modeling

References

Auer S, Lehmann J (2007) What have Innsbruck and Leipzig in common? Extracting semantics from Wiki content. In: Franconi E, Kifer M, May W (eds)

The semantic web: research and applications. Springer, Berlin/New York, pp 503–517
Coursey K, Mihalcea R (2009) Topic identification using Wikipedia graph centrality. In: Proceedings of human language technologies: the 2009 annual conference of the North American chapter of the Association for Computational Linguistics, companion volume: short papers, Boulder. Association for Computational Linguistics, pp 117–120
Coursey K, Mihalcea R, Moen W (2009) Using encyclopedic knowledge for automatic topic identification. In: Proceedings of the thirteenth conference on computational natural language learning, Boulder. Association for Computational Linguistics, pp 210–218
European Travel Commission (2013) Social networking and UGC. http://www.newmediatrendwatch.com/world-overview/137-social-networking-and-ugc, June 2013. (Online; Accessed 25 Oct 2013)
Garshol L (2004) Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all. J Inf Sci 30(4):378
Giles J (2005) Internet encyclopaedias go head to head. Nature 438(7070):900–901
Hassan M (2013) Automatic document topic identification using hierarchical ontology extracted from human background knowledge. Ph.D. dissertation, University of Waterloo
Huynh D, Cao T, Pham P, Hoang T (2009) Using hyperlink texts to improve quality of identifying document topics based on Wikipedia. In: International conference on knowledge and systems engineering, 2009 (KSE'09), Hanoi. IEEE, pp 249–254
Janik M, Kochut K (2008a) Training-less Ontology-based Text Categorization. In: workshop on exploiting semantic annotations in information retrieval (ESAIR 2008) at the 30th European Conference on Information Retrieval, ECIR
Janik M, Kochut K (2008b) Wikipedia in action: ontological knowledge in text categorization. In: IEEE international conference on semantic computing, 2008, Santa Clara. IEEE, pp 268–275
Korfiatis NT, Poulos M, Bokos G (2006) Evaluating authoritative sources using social networks: an insight from Wikipedia. Online Inf Rev 30(3):252–262
Kuhn HW (2005) The Hungarian method for the assignment problem. Nav Res Logist (NRL) 52(1): 7–21
Lloyd S (1982) Least squares quantization in PCM. IEEE Trans Inf Theory 28(2):129–137
Medelyan O (2009) Human-competitive automatic topic indexing. Ph.D. dissertation, The University of Waikato
Medelyan O, Witten I, Milne D (2008) Topic indexing with Wikipedia. In: Proceedings of AAAI workshop on Wikipedia and artificial intelligence: an evolving synergy, Chicago. AAAI, pp 19–24
Ng A, Jordan M, Weiss Y et al (2002) On spectral clustering: analysis and an algorithm. Adv Neural Inf Process Syst 2:849–856
Popescul A, Ungar LH (2000) Automatic labeling of document clusters. http://citeseer.ist.psu.edu/viewdoc/download?doi=10.1.1.33.141&rep=rep1&type=pdf
Schönhofen P (2009) Identifying document topics using the Wikipedia category network. Web Intell Agent Syst 7(2):195–207
Xu W, Gong Y (2004) Document clustering by concept factorization. In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, Sheffield. ACM, pp 202–209
Xu W, Liu X, Gong Y (2003) Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval, Toronto. ACM, pp 267–273
Zhao Y, Karypis G, Fayyad U (2005) Hierarchical clustering algorithms for document datasets. Data Min Knowl Discov 10(2):141–168

Avatar

▸ Gaming and Virtual Worlds

Awareness

▸ Online Privacy Paradox and Social Networks