Studies in Computational Intelligence 424

Ronaldo Menezes
Alexandre Evsukoff
Marta C. González (Eds.)

Complex Networks

Springer

Studies in Computational Intelligence 424

For further volumes:
http://www.springer.com/series/7092

Ronaldo Menezes, Alexandre Evsukoff,
and Marta C. González (Eds.)

Complex Networks

Springer

Editors
Ronaldo Menezes
Florida Institute of Technology
Melbourne, FL
USA

Marta C. González
Massachusetts Institute of Technology
Cambridge, MA
USA

Alexandre Evsukoff
COPPE/Federal University of Rio de Janeiro
Rio de Janeiro, RJ
Brazil

ISSN 1860-949X
ISBN 978-3-642-30286-2
DOI 10.1007/978-3-642-30287-9
Springer Heidelberg New York Dordrecht London

e-ISSN 1860-9503
e-ISBN 978-3-642-30287-9

Library of Congress Control Number: 2012939520

Preface

The International Workshop on Complex Networks series – CompleNet (www.complenet.org) was initially proposed in 2008 with the first workshop taking place in 2009. The initiative was the result of efforts from researchers from the *BioComplex Laboratory in the Department of Computer Sciences at Florida Institute of Technology, USA*, and the *Dipartimento di Ingegneria Informatica e delle Telecomunicazioni, Universita' di Catania, Italia*. CompleNet aims at bringing together researchers and practitioners working on areas related to complex networks. In the past two decades we have been witnessing an exponential increase on the number of publications in this field. From biological systems to computer science, from economics to social systems, complex networks are becoming pervasive in many fields of science. It is this interdisciplinary nature of complex networks that CompleNet aims at addressing. CompleNet 2012 was the third event in the series and was hosted by the *BioComplex Laboratory, Department of Computer Sciences at the Florida Institute of Technology*, USA from March 7–9, 2012.

This book includes the peer-reviewed list of works presented at CompleNet 2012. We received 98 submissions from 22 countries. Each submission was reviewed by at least 3 members of the Program Committee. Acceptance was judged based on the relevance to the workshop themes, clarity of presentation, originality and accuracy of results, and proposed solutions. After the review process, 9 papers and 18 short papers were selected to be included in this book.

The 27 contributions in this book address many topics related to complex networks and have been organized in seven major groups: (1) Network Measures and Models, (2) Agents, Communication and Mobility, (3) Communities, Clusters and Partitions, (4) Emergence in Networks, (5) Social Structures and Networks, (6) Networks in Biology and Medicine, and (7) Applications of Networks.

We would like to thank to the Program Committee members for their work in promoting the event and refereeing submissions. We deeply appreciate the efforts of our keynote speakers: Albert-László Barabási (Northeastern University), Sinan Aral (New York University), and Robert Bonneau (Air Force Office of Scientific Research); their presentation is one of the reasons CompleNet 2012 was such a success. We are grateful to our invited speakers who enriched CompleNet 2012 with

their presentations and insights in the field of Complex Networks (in alphabetical order): Julia Poncela Casasnovas (Northwestern University), Gourab Ghoshal (Northeastern University), Neil Johnson (University of Miami), Sune Lehmann (Technical University of Denmark), Nathalie "Henry" Riche (Microsoft Research), and My Thai (University of Florida).

Special thanks also go to Marco Carvalho, Eraldo Ribeiro, Ryan Stansifer and William Shoaff from the Florida Institute of Technology for their help in organizing CompleNet 2012. The next edition of CompleNet will be hosted by the Freie Universität Berlin, Germany, from March 13-15, 2013.

March 2012
Melbourne, Florida

Ronaldo Menezes
Alexandre Evsukoff
Marta C. González

Contents

Network Measures and Models

Agents, Communication and Mobility

Communities, Clusters and Partitions

Emergence in Networks

Social Structures and Networks

Networks in Biology and Medicine

Applications of Networks

Hybrid Centrality Measures for Binary and Weighted Networks

Alireza Abbasi and Liaquat Hossain

Abstract. Existing centrality measures for social network analysis suggest the importance of an actor and give consideration to actor's given structural position in a network. These existing measures suggest specific attribute of an actor (i.e., popularity, accessibility, and brokerage behavior). In this study, we propose new hybrid centrality measures (i.e., *Degree-Degree, Degree-Closeness and Degree-Betweenness*), by combining existing measures (i.e., *degree, closeness and betweenness*) with a proposition to better understand the importance of actors in a given network. Generalized set of measures are also proposed for weighted networks. Our analysis of co-authorship networks dataset suggests significant correlation of our proposed new centrality measures (especially weighted networks) than traditional centrality measures with performance of the scholars. Thus, they are useful measures which can be used instead of traditional measures to show prominence of the actors in a network.

1 Introduction

Social network analysis (SNA) is the mapping and measuring of relationships and flows between nodes of the social network. SNA provides both a visual and a mathematical analysis of human-influenced relationships. The social environment can be expressed as patterns or regularities in relationships among interacting units [1]. Each social network can be represented as a graph made of nodes or actors (e.g. individuals, organizations, information) that are tied by one or more specific types of relations (e.g., financial exchange, trade, friends, and Web links). A link between any two nodes exists, if a relationship between those nodes exists. If the nodes represent people, a link means that those two people know each other in some way.

Alireza Abbasi · Liaquat Hossain
Centre for Complex Systems Research, Faculty of Engineering and IT,
University of Sydney, NSW 2006, Australia
e-mail: alireza.abbasi@sydney.edu.au

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 1–7.
springerlink.com

Measures of SNA, such as network centrality, have the potential to unfold existing informal network patterns and behavior that are not noticed before [2]. A method used to understand networks and their participants is to evaluate the location of actors within the network. These measures help determine the importance of a node in the network. Bavelas [3] was the pioneer who initially investigates formal properties of centrality as a relation between structural centrality and influence in group process. To quantify the importance of an actor in a social network, various centrality measures have been proposed over the years [4]. Freeman [5] defined centrality in terms of node degree centrality, betweenness centrality, and closeness, each having important implications on outcomes and processes.

While these defined measures are widely used to investigate the role and importance of networks but each one is useful based on especial cases, as discussed below:

(i) Degree centrality is simply the number of other nodes connected directly to a node. It is an indicator of an actor's communication activity and shows popularity of an actor;

(ii) Closeness centrality is the inverse of the sum of distances of a node to others ('farness'). A node in the nearest position to all others can most efficiently obtain information;

(iii) Betweenness centrality of a node is defined as the portion of the number of shortest paths that pass through the given node divided by the number of shortest path between any pair of nodes (regardless of passing through the given node) [6]. This indicates a node's potential control of communication within the network and highlights brokerage behavior of a node;

(iv) Eigenvector centrality is a measure of the importance of a node in a network. It assigns relative scores to all nodes in the network based on the principle that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. Bonacich [7] defines the centrality of a node as positive multiple of the sum of adjacent centralities.

For detail explanations and equations for the centrality measures please refer to [8].

In this study, we propose new centrality measures (i.e., *Degree-Degree, Degree-Closeness and Degree-Betweenness*), which combines existing measures (i.e., *degree, closeness and betweenness*) for improving our understanding of the importance of actors in a network. To show the significance of proposed new measure in evaluating actors' importance in the network, we first compare our proposed measures with a sample simple network and then we test it with a real co-authorship network having performance measure of nodes (scholars).

2 Hybrid Centrality Measures

To investigate the role and importance of nodes in a network, the traditional (popular) centrality measures could be applied in especial cases. By developing

hybrid (combined) centrality measures, we are expecting to have a better understanding of importance of actors (nodes) in a network which can assist in exploring different characteristics and role of the actors in the network.

The proposed new measures work in combining (at least) two of the most popular and basic existing centrality measures of each actor. Thus, to achieve our goal, we propose three measures with an emphasis on degree, closeness and betweenness centralities of the direct neighbors of an actor. This will support in identifying the nodes which are central themselves and also connected to direct central nodes, which demonstrates strategic positions for controlling the network.

To define new hybrid centrality measures, we consider a network having centrality measures of each node as the attribute of the node. Then, we define hybrid centrality measures of a node as sum of centrality measure of all directly connected nodes. Thus, the *Degree-Degree (DD), Degree-Closeness (DC) and Degree-Betweenness (DB)* centralities of node *a* is given by:

$$DD(a) = \sum_{i=1}^{n} C_D(i), \qquad DC(a) = \sum_{i=1}^{n} C_C(i), \qquad DB(a) = \sum_{i=1}^{n} C_B(i)$$

Where n is the number of direct neighbors of node a (degree of node a) and $C_D(i)$ is the degree centrality measure, $C_C(i)$ is the closeness centrality measure and $C_B(i)$ is the betweenness centrality measure of node i (as a representation of direct neighbors of node a).

To have generalized measures, considering weighted networks which their links have different strengths, we can extend definitions by considering the weight of the links. Thus, the general hybrid centrality measures of node a are given by:

$$DD_w(a) = \sum_{i=1}^{n} [w(a,i) * C_D(i)] \quad , \quad DC_w(a) = \sum_{i=1}^{n} [w(a,i) * C_C(i)] \quad , \quad DB_w(a) = \sum_{i=1}^{n} [w(a,i) * C_B(i)]$$

Where n is the number of direct neighbors of node a and $w(a,i)$ is the weight of the link between node a and its neighbors i.

Degree-Degree (DD) centrality indicates the actors who are connected better to more actors. It reflects the theory that connecting to more powerful actors will give you more power. So, it indicates the popularity of an actor based on popularity of its direct neighbors. Degree-Closeness (DC) centrality indicates not only an actors' power and influence on transmitting and controlling information but also efficiency for communication with others or efficiency in spreading information within the network. It indicates popularity and accessibility of an actor simultaneously. Also, Degree-Betweenness (DB) centrality indicates not only an actors' power and influence on transmitting and controlling information but also potential control of communication and information flow within the network. It shows popularity and brokerage attitude of an actor in the network simultaneously.

4 Applicability of New Measures for Analyzing Nodes in Networks

4.1 Simple Examples

To compare our new proposed centrality measures and traditional centrality measures, we consider a simple network (Figure 1) and calculate nodes centrality measures (Table 1) and show the different ranks of the nodes based on each centrality measures in Table 2.

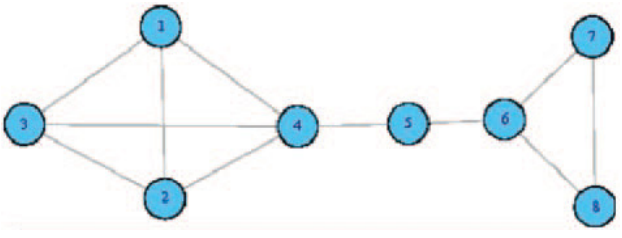


Fig. 1. An example simple network for comparing traditional and new centrality measures

Table 1. Nodes' centrality measures for example network in Figure 1

No	C_D	C_C	C_B	C_E	DD	DC	DB
1	.429	.438	0	.671	1.429	1.458	0.571
2	.429	.438	0	.671	1.429	1.458	0.571
3	.429	.438	0	.671	1.429	1.458	0.571
4	.571	.583	.571	.739	1.571	1.896	0.571
5	.286	.583	.571	.280	1.000	1.083	1.048
6	.429	.500	.476	.130	0.857	1.320	0.571
7	.286	.368	0	.062	0.714	0.868	0.476
8	.286	.368	0	.062	0.714	0.868	0.476

Table 2. Ranking nodes based on different centrality measures for network in Figure 1

Rank	C_D	C_C	C_B	C_E	DD	DC	DB
1	4	4,5	4,5	4	4	4	5
2	1,2,3,6	6	6	1,2,3	1,2,3	1,2,3	1,2,3,4,6
3	5,7,8	1,2,3	1,2,3,7,8	5	5	6	7,8
4		7,8		6	6	5	
5				7,8	7,8	7,8	

As we expect the results and even ranks between traditional centrality measures are different except for eigenvector centrality (C_E, DD and almost DC). That is because the hybrid centralities can be considered as variants of eigenvector centrality.

4.2 A Real Co-authorship Network

Several studies have been shown the applicability of centrality measures for co-authorship networks for demonstrating how centrality measures are useful to reflect the performance of scholars (i.e., scholars' position within their co-authorship network) [8-10]. Here, also in another attempt, to assert the applicability of new hybrid centrality measures, we study a real co-authorship network having performance measure of actors (scholars) and their centrality measures, and test the correlation between centrality measures and performance measures.

4.2.1 Data

We analyzed the dataset which has been used in [8-9], publication list of five information schools: University of Pittsburgh, UC Berkeley, University of Maryland, University of Michigan, and Syracuse University. The data sources used are the school reports, which include the list of publications of researchers, DBLP, Google Scholar, and ACM portal. Citation data has been taken from Google Scholar and ACM Portal. Our data covered a period of five years (2001 to 2005), except for the University of Maryland iSchool, which had no data for the year 2002 in their report. We followed Google Scholars approach and did not differentiate between the different types of publications. After the cleansing of the publication data of the five iSchools, 2139 publications, 1806 authors, and 5310 co-authorships were finally available for our analysis.

4.2.2 Measuring Scholars' Performance

To assess the performance of scholars, many studies suggest quantifying scholars' publication activities (mainly citations count) as a good measure for the performance of scholars. Hirsch [11] introduced the h-index as a simple measure that combines in a simple way the quantity of publications and the quality of publications (i.e., number of citations). The h-index is defined as follows: "A scientist has an h-index of h, if h of her Np papers have at least h citations each, and the other (Np - h) papers have at most h citations each" [11]. In other words, a scholar with an index of h has published h papers, which have been cited at least h times.

4.2.3 Results

The result of Spearman correlation rank test between centrality measures and scholars' performance (e.g., sum of citations and h-index) has been shown in Table 3. As it shows all traditional and new centrality measures are significantly correlated to performance measure except for eigenvector centrality and closeness which have weak or not significant correlations.

Table 3. Spearman correlation rank test between scholars' network centrality measures and their performance

Centrality Measures (N=1806)	Scholars Performance	
	Sum_Cit.	h-index
C_D	.332 **	.311 **
C_C	- .012	.052 *
C_B	.388 **	.501 **
C_E	.060 *	.041
DD	.296 **	.261 **
DC	.303 **	.295 **
DB	.203 **	.255 **
DD_W	.394 **	.426 **
DC_W	.385 **	.432 **
DB_W	.304 **	.503 **

*. Correlation is significant at the .05 level (2-tailed).

**. Correlation is significant at the .01 level (2-tailed).

All new hybrid centrality measures of scholars have high positive significant association with their performance rather than traditional centrality measures. That is because the new measures combined two centrality measures' attributes and highlights the importance of the nodes in the network more than traditional ones. The new centrality measures considering weighted links have higher correlation coefficients. This is due to taking into account scholar's repeated collaborations.

Another outcome of this result is that new centrality measure are different from eigenvector centrality and to support this we also applied non-parametric independent t-test (Mann-Whitney U test) to compare the distribution of eigenvector centrality measure between two groups (lower than mean of h-index and above mean) and it was not significant while the t-test was significant for new centrality measures. So, this also supports that new centrality measures are different from eigenvector centrality.

5 Conclusions

In this paper, we proposed a new class of hybrid centrality measures (i.e., DD, DC, DB). We illustrated similarities and dissimilarities with respect to the traditional (standard) measures considering a sample network and a real co-authorship network. Our analysis showed that they are good indicators of the importance of an actor in a social network by combing traditional centrality measures: degree of each node with degree, closeness and betweenness of its direct contacts for Degree-Degree, Degree-Closeness, Degree-Betweenness measures respectively. As each of them combines two different attributes (characteristics) of traditional measures, they could be a good extension of traditional centrality measures.

To demonstrate that the new measures are useful in practice to evaluate actors' importance in the network, we test it with having performance measures (e.g., sum of citations, h-index) of scholars. The results highlighted that Degree-Degree (DD), Degree-Closeness (DC) and Degree-Betweenness (DB) centralities have

significant correlation with performance of the actors. Based on the results, we suggest that DD, DC and DB centralities of an actor are good measures to demonstrate the importance of an actor (e.g., performance, power, social influence) in a network.

It has been shown that in complex networks, Betweenness centrality of an existing node is a significantly better predictor of preferential attachment by new entrants than degree or closeness centrality [12]. We expect that the new proposed measure may be a better driver of attachment of new added nodes to the existing ones during the evolution of the network.

The computational complexity for calculating the proposed measure can be considered as one of the limitations of these new proposed measures which needs more research in future works. Also to generalize the applicability of the new hybrid measures, it is needed to apply them in different (complex) networks in future works.

Acknowledgments. We appreciate Dr. Kenneth Chung's feedback on the earlier version of this work.

References

1. Wasserman, S., Faust, K.: Social network analysis: Methods and applications. Cambridge Univ. Press (1994)
2. Brandes, U., Fleischer, D.: Centrality measures based on current flow. Springer (2005)
3. Bavelas, A.: Communication patterns in task-oriented groups. Journal of the Acoustical Society of America 22, 725–730 (1950)
4. Scott, J.: Social network analysis: a handbook. Sage (1991)
5. Freeman, L.C.: Centrality in social networks conceptual clarification. Social Networks 1(3), 215–239 (1979)
6. Borgatti, S.: Centrality and AIDS. Connections 18(1), 112–114 (1995)
7. Bonacich, P.: Factoring and weighting approaches to status scores and clique identification. Journal of Mathematical Sociology 2(1), 113–120 (1972)
8. Abbasi, A., Altmann, J., Hossain, L.: Identifying the Effects of Co-Authorship Networks on the Performance of Scholars: A Correlation and Regression Analysis of Performance Measures and Social Network Analysis Measures. Journal of Informetrics 5(4), 594–607 (2011)
9. Abbasi, A., Altmann, J.: On the Correlation between Research Performance and Social Network Analysis Measures Applied to Research Collaboration Networks. In: Proceedings of the 44th Annual Hawaii International Conference on System Science. IEEE, Waikoloa (2011)
10. Yan, E., Ding, Y.: Applying centrality measures to impact analysis: A coauthorship network analysis. Journal of the American Society for Information Science and Technology 60(10), 2107–2118 (2009)
11. Hirsch, J.: An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences 102(46), 16569 (2005)
12. Abbasi, A., Hossain, L., Leydesdorff, L.: Betweenness Centrality as a Driver of Preferential Attachment in the Evolution of Research Collaboration Networks. Journal of Informetrics 6(2) (2012)

A Growing Model for Scale–Free Networks Embedded in Hyperbolic Metric Spaces*

Giuseppe Mangioni and Antonio Lima

Abstract. Some results by Krioukov et al. show how real world networks are produced by hidden metric spaces. Specifically, scale-free networks can be obtained from hyperbolic metric spaces. While the model proposed by Krioukov can produce a static scale-free network, all nodes are created at one time and none can be later added. In this work we propose a growing model which leverages the same concepts and allows to gradually add nodes to a scale-free network, obtained from a discretised hyperbolic model. We also show how nodes are correctly positioned relying on local information and how greedy routing builds optimal paths in the network.

1 Introduction

Complex networks provide a natural abstraction for many processes which happen everyday [1, 9, 4], in all areas of human life and knowledge, making it possible to study them more easily and more deeply, sometimes opening the path to new groundbreaking discoveries. For this reason, it is fundamental to fully understand their structure, their dynamic behaviour and all the underlying mechanics.

Many complex networks observed from real world exhibit a well-defined property, known as scale-free topology [3]. Scale–free topology is characterised

Giuseppe Mangioni
DIEEI, University of Catania, Viale Andrea Doria, 6 - I-95125 Catania (Italy)
e-mail: giuseppe.mangioni@dieei.unict.it

Antonio Lima
School of Computer Science, University of Birmingham, Edgbaston,
Birmingham B15 2TT United Kingdom
e-mail: axl162@cs.bham.ac.uk

* This work was carried out when Antonio Lima was at the DIEEI, University of Catania.

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 9–17.
springerlink.com

by two aspects: the node degree distribution $P(k)$ follows a power law distribution $P(k) \sim k^{-\gamma}$ and the network has a high clustering coefficient, which means you can find many triangles in the graph. One of the most important functions of a complex network is the transport function, which for computer networks represents the information transmission, for transportation networks the movement of people or goods, for social networks the spreading of news or gossips, and so on. It is interesting to analyse how the network builds a path from the source to the target, or in other words how it solves the routing request, without having a global view of the network, but having only local informations to take its routing decision at each step. Surprisingly, real world networks can solve this problem very efficiently, due to their topological properties [5, 6].

Krioukov et al. [7] have proposed to consider complex networks on a different prospective, stating that they exist in some so-called hidden metric spaces, which directly influence the topology of the related complex networks. More in detail, the hidden metric space defines a distance function between two entities in the hidden space, and in turn the distance influences the probability that the two nodes related to the entities will be connected in the resulting network. Krioukov et al. also show how hyperbolic spaces ([2]) naturally form networks with scale-free topology and, as a consequence of this, greedy routing can be used on these and it achieves very high efficiency.

While this result is extremely important, the Krioukov model is a static model and for this reason it cannot be used as-it-is for real-world applications. In this paper we propose a growing model which leads to a scale-free network by using Krioukov model in a discretised fashion. The paper is structured as follows: we will first introduce Krioukov model, then we will describe our discretised version and at last we will show our results and draw our conclusions.

2 The Model of Krioukov et al [7]

Krioukov et al. [7] propose a model of networks embedded in hyperbolic spaces showing that it gives naturally a network with a power–law degree distribution. The idea behind this model is to build a network whose nodes are more likely connected if they are near in the sense of the distance metric defined in the hyperbolic space. So, the first thing to do is the choice of the hyperbolic space. The second step is the decision about the nodes distribution function. The third is about the choice of the connection probability as function of the hyperbolic distance between nodes.

In the model they propose, the following choices have been made: 1) to use the hyperbolic plane, 2) to distribute non-uniformly N nodes over a disc of radius R and 3) to use the step function on $[0, R]$ as connection probability function. In particular, given a target number of nodes N and average degree $\overline{k}$, they generate a network as follows:

- Set the radius R of the hyperbolic disc according to $N = \kappa e^{R/2}$ (k is a parameter used to tune $\overline{k}$).
- Assign to each node an angular coordinate θ uniformly distributed in $[0, 2\pi)$.
- Assign to each node a radial coordinate $r \in [0, R]$ with a probability $\rho(r) = \alpha e^{\alpha r}(e^{\alpha R} - 1)^{-1}$, $\alpha \in [1/2, 1]$.
- Connect every pair of nodes whenever the hyperbolic distance between them is smaller than R. Given two nodes with polar coordinates (r, θ) and (r', θ'), the hyperbolic distance x between them is defined as:

$$cosh(x) = cosh(r)cosh(r') - sinh(r)sinh(r')cos(\Delta\theta) \tag{1}$$

where $\Delta\theta = min(|\theta - \theta'|, 2\pi - |\theta - \theta'|)$.

Using such a model, the generated network has a power–law degree distribution that, for large k, can be approximately given by $P(k) \sim k^{-\gamma}$, where:

$$\gamma = \begin{cases} 2\alpha + 1 & \text{if } \alpha \geq \frac{1}{2} \\ 2 & \text{if } \alpha \leq \frac{1}{2} \end{cases} \tag{2}$$

Moreover, in [7] it is shown that on such a kind of networks a greedy routing algorithm achieves both 100% reachability and optimal path lengths. In other words, starting from a node of the network it is possible to find any other node using a greedy routing strategy based only on local information. More specifically, a node selects as the next hop the neighbour that is closest (in terms of hyperbolic distance) to the destination in the hyperbolic space. This is a very interesting feature that can be exploited in order to perform optimal message routing without needing a, sometimes heavy, routing protocol.

3 A Model of Networks that Grow over an Hyperbolic Space

Our idea is to use the results presented in [7] to build a synthetic network which topology is congruent with an hidden hyperbolic space. For doing that, the model presented in [7] is not immediately applicable, since it generates a whole network at once. Instead, what we need is a model of networks that grow over an hyperbolic space.

As proposed in [10], a natural way to deal with network growing is to consider a model where the radius of the hyperbolic disk grows as nodes join the network. In this case the disk radius should grow with a rate given by: $R = 2ln(N/\kappa)$. Unfortunately, such a kind of model requires that a node joining the network knows the number of nodes already present in it. Even if in [10] an algorithm to find the number of nodes of the network is proposed, in general, in a distributed environment the computation of such a kind of information is costly. Our work is devoted to develop a distributed network growing model that can be efficiently used in a variety of applications, such as to build overlay networks [8] in distributed computing environment.

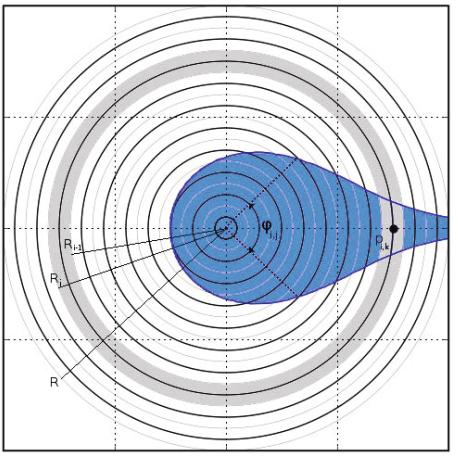


Fig. 1. Hyperbolic disk of radius R. The grey shaded area is the annulus i. The blue shaded shape is the area containing all the nodes whose hyperbolic distance from the node p is less than or equal to R.

In the model we propose, nodes are mapped into a hyperbolic disc having a radius $R = 2ln(N_{max}/\kappa)$, where N_{max} is the maximum number of nodes of the network. Moreover, nodes can be placed on the hyperbolic disc only at certain fixed distances from the disc centre (see figure 1). To explain this, we suppose to divide the hyperbolic disc into N_L annuli (or levels) having outer and inner radii respectively of R_i and R_{i-i}, where $i \in \{1, 2, \dots N_L\}$, and defined as:

$$R_i = \frac{R}{N_L} i \tag{3}$$

The number of nodes N_i placed on each level i, is the expected number of nodes in the interval $[R_{i-1}, R_i]$ (i.e. the expected number of nodes in the annulus i), computed using the distribution function $\rho(r)$. It is given by:

$$N_i = N_{max} \frac{1 - e^{-\alpha \frac{R}{N_L}}}{e^{\alpha R} - 1} e^{\alpha \frac{R}{N_L} i} \tag{4}$$

In our model the N_i nodes of the level i are positioned only along the circumference placed in the middle of the annulus i and having radius $\tilde{R}_i$. Therefore, the hyperbolic polar coordinates of a generic node p placed at the level i are $p(r_i, \theta_k)$ where:

$$r_i = \tilde{R}_i = \frac{R}{N_L} \left(i - \frac{1}{2} \right) \qquad i \in \{1, \dots N_L\} \tag{5}$$

$$\theta_k = \frac{2\pi}{N_i} k \qquad k \in \{0, \dots N_i - 1\} \tag{6}$$

In the following, a node p with coordinates $p(r_i, \theta_k)$ will be referred to as $p_{i,k}$. As can be seen from equations 5 and 6, both the radial coordinate r and the angular coordinate θ of a node can assume only a given number of discrete values or slot, so our model can be considered as a sort of discretisation of the model proposed in [7].

Once a node is placed in the hyperbolic disc, it is connected to those nodes having a hyperbolic distance from it smaller than R, i.e. all those nodes within the blue shaded shape in figure 1. Given the nodes placing strategy implemented in our model, it is quite easy to implement an algorithm able to find the neighbourhood of a given node, as will be explained later in detail.

4 Model Analysis

To study the topological characteristics of our model, firstly we calculate the average degree $k(i)$ of nodes located at level i. To do so, let consider a node $p_{i,k}$ placed at the level i and whose angular coordinate is, for the sake of simplicity, 0 (i.e. $k = 0$). In figure 1 the blue shaded area is the region containing all the points whose hyperbolic distance from the point $p_{i,k}$ is less than or equal to R. The average degree of the node $p_{i,k}$ is given by:

$$k(i) = \sum_{j=1}^{N_L} f_j N_j \tag{7}$$

where N_j is the number of nodes placed at level j, computed using equation 4 and f_j is the fraction of the N_j nodes whose hyperbolic distance from $p_{i,k}$ is less than or equal to R (i.e. those nodes that are within the blue shaded area in figure 1). A simple formulation of f_j can be derived noting that such a number is equal to the fraction of circular angle within the blue shaded area; it is given by:

$$f_j = \frac{\varphi_{i,j}}{2\pi} \tag{8}$$

Looking at the figure 1 it is possible to note that $\varphi_{i,j}$ is equal to 2π for those j for which $i+j \leq N_L+1$ (i.e. all nodes at level j at hyperbolic distance from $p_{i,k}$ smaller than R)[1]; in the other cases it is given by the following equation:

$$cosh(R) = cosh(\tilde{R}_j)cosh(\tilde{R}_i) - sinh(\tilde{R}_j)sinh(\tilde{R}_i)cos(\varphi_{i,j}/2) \tag{9}$$

Therefore:

$$\varphi_{i,j} = \begin{cases} 2\pi & i+j \leq N_L+1 \\ 2acos\left(\frac{cosh(\tilde{R}_j)cosh(\tilde{R}_i)-cosh(R)}{sinh(\tilde{R}_j)sinh(\tilde{R}_i)}\right) & i+j > N_L+1 \end{cases} \tag{10}$$

Substituting equations 4 and 8 in equation 7 we obtain:

$$k(i) = \frac{N_{max}}{2\pi}\frac{1-e^{-\alpha\frac{R}{N_L}}}{e^{\alpha R}-1}\sum_{j=1}^{N_L}\varphi_{i,j}e^{\alpha\frac{R}{N_L}j} \tag{11}$$

Putting in evidence that $\varphi_{i,j}$ is equal to 2π for each i, j such that $i+j \leq N_L+1$, we can rewrite equation 11 as:

[1] From now on, we will refer to these levels as *fully contained levels*.

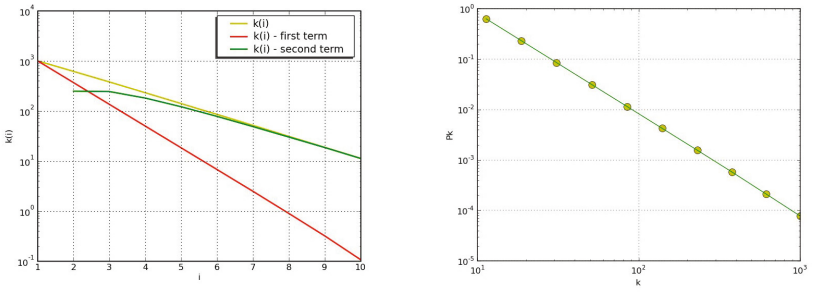
$$k(i) = \frac{N_{max}}{2\pi} \frac{1 - e^{-\alpha \frac{R}{N_L}}}{e^{\alpha R} - 1} \left[2\pi \sum_{j=1}^{N_L+1-i} e^{\alpha \frac{R}{N_L} j} + \sum_{j=N_L-i+2}^{N_L} \varphi_{i,j} e^{\alpha \frac{R}{N_L} j} \right] \tag{12}$$

The first sum in the equation 12 is a finite power series, then it can be rewritten as:

$$\sum_{j=1}^{N_L+1-i} e^{\alpha \frac{R}{N_L} j} = \frac{e^{\alpha R} e^{-\alpha \frac{R}{N_L}(i-1)} - 1}{1 - e^{-\alpha \frac{R}{N_L}}} \tag{13}$$

Equation 13 shows an exponential decreasing trend, meaning that the contribution on this term in equation 12 decreases as j approaches the value of N_L, corresponding to the last circumference. This behaviour is easy to explain since that term is related to the fully contained levels, and its contribution tends to 0 as the target node moves to the level N_L.

In figure 2(a) we show the average degree $k(i)$ as a function of the level i and the contribution of the first and second term of equation 12 for a network of 1000 nodes. It is possible to note that the average degree decreases exponentially with the distance from the disc centre.



(a) (b)

Fig. 2. For a network of 1000 nodes and 10 levels, (a) average degree at level i (in semi-log plot) and (b) degree distribution

By definition the degree distribution P_k of a network is the fraction of nodes in the network with degree k. Therefore, in order to derive the degree distribution P_k in our model, we need to get two values for every level i: 1) the fraction of nodes inside level i and 2) their degree. The former is given by the ratio N_i/N_{max} (N_i is computed by using equation 4), whereas the latter is computed by equation 12. We omit here the analytic expression for P_k, but it follows a power-law, as shown in figure 2(b) for a network with 1000 nodes. Such an analytic result matches simulations as detailed in the next section.

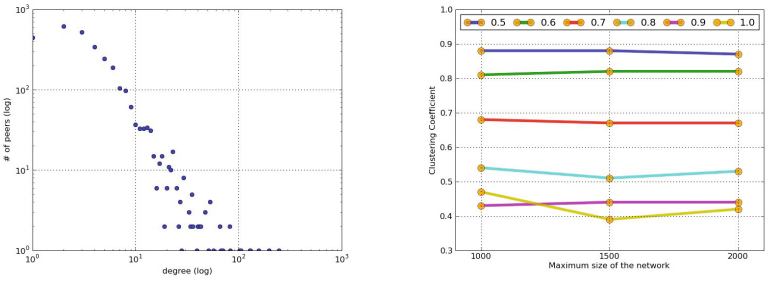
5 Growing Model and Results

Now that we have assessed the structure of the network, i.e. the positions nodes can be placed in, let's discuss the growing model or, in other words, how the position of every single node joining the network is determined. Our final aim is to maintain a degree distribution consistent with the model, throughout network dynamical changes. Also, we want to rely only on local information for deciding where a next node can be placed. Since in the hyperbolic model we have shown that the position of each node is directly correlated with its degree, the task of maintaining a given degree distribution can be translated in terms of maintaining a given spatial node distribution over the hyperbolic space. If we assure this condition, the working model continues to be valid despite the changes in the network. The joining phase consists of four steps.

1. The node randomly chooses, according to a uniform distribution, an angle $\varphi \in [0, 2\pi]$. This angle will define an ideal point $P = (R, \varphi)$ with maximum radius and random angle. Ideally, the node wants to place itself in the free slot that has the minimum distance to this P.
2. The node queries the network (through an arbitrarily-chosen node, which doesn't need to be close to the ideal point). Such query drives the newcomer node to the closest real node to the ideal point.
3. The closest node has total knowledge about the nodes (and the free slots) which have hyperbolic distance smaller than R, so he can effectively assign the right slot for the newcomer.
4. The newcomer places itself in the assigned slot and notifies all its neighbours (i.e. nodes who are at distances smaller than R from its position) the slot is now occupied.

It may be useful to remark that the fourth step is fundamental in order to keep the network up to date and to allow every node to have total knowledge of his local neighbours (i.e. which nodes and which free slots are present within hyperbolic distance R from it), as it is required by the joining phase itself, in step three. Since every joining phase goes through the fourth step, every node in the network is kept up to date with local knowledge of its neighbours. We developed a network simulator which grows according to the proposed model, in order to analyse its structure and verify whether it conserves the same scale-free property of its counter part. Figure 3(a) shows the degree distribution for a full network of 5000 peers in 20 levels. Figure 3(b) shows average clustering coefficient for networks with different sizes and values of α. Both graphs are compatible with the scale-free model.

We also analysed the average path length (APL), while varying the maximum size of the network, for different values of α, from 0.5 to 1.0. To assess this measurement we saturated completely the networks, this means they were almost at full capacity. APL does not primarily depend upon the size of the network, but mainly upon the parameter α, that affects directly γ,



(a) (b)

Fig. 3. (a) Degree distribution and average clustering coefficient (b) for a full network of 5000 peers, 20 levels, $\alpha = 0.8 \Rightarrow \gamma = 2.6$

as shown before. The unique APL variation is for $\alpha = 1.0$, but it should be considered an outlier, since the network behaves strangely at that limit value (that corresponds to $\gamma = 3$). Another limit-case situation is network obtained for $\alpha = 0.5$: each node can be reached in just 2 hops because all nodes are connected to a central hub. The model structure allows to reach each node in a relatively low number of hops, regardless of the real network size. In addiction, we analysed the paths used by greedy routing and we checked them against the shortest paths. All paths were coincident to the shortest path, confirming that greedy routing in the model always chooses the best path.

6 Conclusion

We have presented a growing model for a hyperbolic graph, which naturally produces a scale-free network. Starting from a static model which relies on a continuous hyperbolic space, we have designed a discretised version which conserves the desirable properties of the static model (scale-free topology and extremely efficient and cheap greedy routing), while allowing the network to grow based only on local information and local greedy routing algorithm.

Future works will be focused on deeper discovery of static and dynamic properties of the network and analysis of network robustness against failure and targeted attack.

References

1. Albert, R., Barabási, A.-L.: Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 47–97 (2002)
2. Anderson, J.W.: Hyperbolic Geometry. Springer, London (2005)
3. Barabasi, A.-L.: Scale-free networks: A decade and beyond. Science 325(5939), 412–413 (2009)

4. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.-U.: Complex networks: Structure and dynamics. Physics Reports 424(4-5), 175–308 (2006)
5. Kleinberg, J.M.: Navigation in a small world. Nature 406(6798) (August 2000)
6. Kleinberg, J.M.: Complex networks and decentralized search algorithms. In: ICM (2006)
7. Krioukov, D., Papadopoulos, F., Boguñá, M., Vahdat, A.: Efficient navigation in scale-free networks. Embedded in Hyperbolic Metric Spaces (2008) arXiv:0805.1266v1 [cond-mat.stat-mech]
8. Lua, E.K., Crowcroft, J., Pias, M., Sharma, R., Lim, S.: A survey and comparison of peer-to-peer overlay network schemes. IEEE Communications Surveys and Tutorials 7, 72–93 (2005)
9. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45, 167 (2003)
10. Papadopoulos, F., Krioukov, D., Boguñá, M., Vahdat, A.: Greedy forwarding in dynamic scale-free networks embedded in hyperbolic metric spaces. In: Proceedings of INFOCOM 2010. IEEE Press, USA (2010)

The Robustness of Balanced Boolean Networks

Ming Liu and Elena Dubrova

Abstract. One of the characteristic features of genetic regulatory networks is their inherent robustness, that is, their ability to retain functionality in spite of the introduction of random errors. In this paper, we focus on the robustness of *Balanced Boolean Networks (BBNs)*, which is a special kind of Boolean Network model of genetic regulatory networks. Our goal is to formalize and analyse the robustness of *BBNs*. Based on these results, applications using Boolean network model can be improved and optimized to be more robust.

We formalize *BBNs* and introduce a method to construct *BBNs* for 2-singleton attractors Boolean networks. The experiment results show that *BBNs* have a good performance on tolerating the single stuck-at faults on every edge. Our method improves the robustness of Boolean networks by at least 13% in average, and in some special case, up to 61%.

1 Introduction

A living cell could be considered as a molecular digital computer that configures itself as a part of the execution of its code. The core of a cell is the *DNA*, which represents the information for building the basic components of cells as well as encodes the entire process of assembling complex components.

The most attractive feature of living organisms is the robustness [2], which is always an important topic of biology research. Especially after Kauffman modelled the genetic process using Boolean Networks [11] in 1969, the research of robustness has been highly promoted. Recently, a lot of researches on the robustness of *genetic regulatory networks (GRNs)* focus on the effect of shape and size of basin of certain attractor [17, 8]. In a view of biologist, almost all cells in one living organism have

Ming Liu · Elena Dubrova
School of Information and Communication Technology,
Royal Institute of Technology(KTH), Stockholm, Sweden
e-mail: {lming,dubrova}@kth.se

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 19–30.
springerlink.com

the same copy of *DNAs* or the same *GRNs*; and the different functions of these cells right represent the different attractors [10, 13, 12]. The more basin states of an attractor means the higher probability for a cell to perform in this function mode; so, many researchers defined their robustness as the probability of a *GRN* settled down into one attractor in a random environment [14]. However, we believe that robustness is one important characteristic of *GRNs*, and it's necessary to treat the robustness of a *GRN* as a whole. In this paper, our research is about the robustness of *Balanced Boolean Networks(BBNs)*, a special kind of Boolean Network model of *GRNs*.

A real *GRN* can be modelled by a Boolean network [1, 5, 9]. *State Transition Graphs (STGs)* are used to illustrate the dynamic behaviour of these Boolean networks, and it is a common view that almost every attractor has a large basin, no matter if it is a singleton attractor or a cycle attractor. A large basin makes the Boolean network more robust and stable [14]. However, when we construct Boolean networks for our own applications, the models are always not so *natural* and are greatly unstable. One easy way to solve this problem is to construct Balanced Boolean Networks by making every attractor to have the same number of basin states.

In this paper, we start with the synchronous 2-singleton-attractors Boolean network model. In the next section, we formalize *BBNs*, and introduce a method for constructing *BBNs*. Then in *Sect*. 3, we present a definition of robustness for *stuck-at faults* on every edge in Boolean network. In *Sect*. 4, experiment is performed. The results show that *BBNs* have good performance on tolerating the single stuck-at faults on every edge, and the robustness of *BBNs* built in our method improve by at least 13% in average, and in some special case, up to 60%. *Section* 5 concludes the paper and discusses open problems.

2 Balanced Boolean Networks

In this section, we formalize Balanced Boolean Networks and introduce a method to construct them.

2.1 Definition

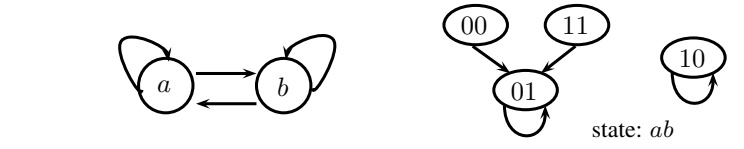
A *Balanced Boolean Network (BBN)* is a special genetic regulatory network, which is defined as $G(V,F)$ with a set of nodes $V = \{x_1,\ldots,x_n\}, x_i \in \{0,1\}$, and a set of Boolean functions $F = \{f_1,\ldots,f_n\}, f_i : \{0,1\}^{k_i} \to \{0,1\}$. Each node x_i represents the expression state of the gene x_i, where $x_i = 0$ means that the gene is *OFF*, and $x_i = 1$ means it is *ON*. Each Boolean function $f_i(x_{i_1},\ldots,x_{i_{k_i}})$ with k_i specific input nodes is assigned to node x_i and is used to update its value. Under the synchronous updating scheme, all genes are updated simultaneously according to their corresponding update functions. The *state* of a network is a vector of values of its state variables $(x_1,x_2,\ldots,x_n)$. Time is viewed as proceeding in discrete steps. For the *synchronous* type update, at every time step, the next state of a network, $(x_1^+,x_2^+,\ldots,x_n^+)$,

is determined from the current state, $(x_1, x_2, \ldots, x_n)$, by updating the values of the state variables of all nodes simultaneously to the values of the corresponding f_is:

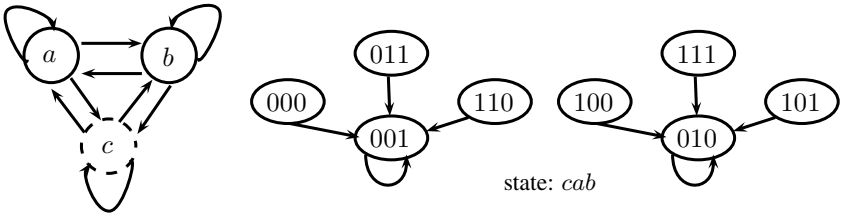
$$x_i^+ = f_i(x_{i_1}, x_{i_2}, \ldots, x_{i_{k_i}}) \tag{1}$$

where $x_{i_1}, x_{i_2}, \ldots, x_{i_{k_i}}$ are the state variables associated to the predecessors of node i.

Since a synchronous *Balanced Boolean Network* is deterministic and finite, any sequence of its consecutive states eventually converges to either a singleton state, or a cycle of states, called *attractor*. The *basin* of attractor A, denoted by $B(A)$, is the set of all states from which A can be reached. When we draw all these states using circles and directed arrows, we get the *state transition graphs*. An example of a 2-nodes Boolean network and its *BBN* is shown in *Fig.* 1.



(a) A 2-nodes Boolean network



(b) A *BBN* for Boolean network in (a)

Fig. 1. A 2-nodes Boolean networks and its *BBN*

These two networks both have 2-singleton attractors. *Figure* 1.(a) shows the network structure of a 2-nodes Boolean network and the state transition graphs. *Figure* 1.(b) shows one *BBN* for (a). The left side of *Fig.* 1 shows the network structures, and the dashed circle in (b) shows the additional node inserted to make the original 2-nodes Boolean networks balance. The right side of *Fig.* 1 shows the *STGs*. The attractors in *BBN*, which have the same number of basin states, are balanced, while the 2-nodes Boolean network is not.

2.2 Properties

Suppose that we have a *Balanced Boolean Network* G with n vertices $v_1, \ldots, v_n$ and m-single attractors $A_1, A_2, \ldots, A_m$ such that the attractor A_i is given by the state $(a_{i1}, a_{i2}, \ldots, a_{in}) \in \{0,1\}^n, i \in \{1,2,\ldots,m\}$. As all the m attractors are single attractors and have the same number of basin states, the number of basin states for

every attractor will be $\frac{2^n-m}{m}$. The basin $B_i(A)$ partition the Boolean space $\{0,1\}^n$ into m connected components via a dynamic process. We assume that the set of all points of the Boolean space corresponding to the states in the basin of attraction of A_i is mapped into the state of the attractor A_i. Then, G defines a set of n Boolean functions of type $g_j : \{0,1\}^n \to \{0,1\}$of variables $x_1,\ldots,x_n$, where the variable x_i corresponds to the variable associated to the vertex i. More formally:

Definition 1. *A* Balanced Boolean Network *with n vertices and m single attractors* $A_1,A_2,\ldots,A_m$*, such that the attractor* A_i *is given by the state* $(a_{i1},a_{i2},\ldots,a_{in}) \in \{0,1\}^n, i \in \{1,2,\ldots,m\}$*, represents a set of n Boolean function of type* $g_j : \{0,1\}^n \to \{0,1\}$ *, which are defined as follows:*

$$g_j(s_1,\ldots,s_n) = a_{ij}, \textit{ if and only if } (s_1,\ldots,s_n) \in B(A_i),$$

for all $(s_1,\ldots,s_n) \in \{0,1\}^n$ *, and all* $i,j \in \{0,1,\ldots,m-1\}$

2.3 Construct BBNs

As it described above in *Sect*. 2.2, we need only to make the attractors have the same amount of basin states to build a *BBN*. It's a good idea to find all possible *BBNs*, but it is not feasible for large networks. One real question is finding a best *BBN* from the point of view of robustness. It's a *NP-hard* problem and it is impossible to search in the entire space of all *BBNs* for large Boolean networks. Even for small Boolean networks, the computation will be extremely heavy.

Here, we introduce a simple method, in which we can build *BBNs* from any 2-singleton-attractors Boolean networks. Experiment results in *Sect*. 4.3 show that this method is useful for finding the best *BBNs*, and is practical to avoid huge computation for large Boolean networks. Considering the Boolean network in *Fig*. 1, we present the Boolean functions of the 2-nodes Boolean network and its *BBN* in *Fig*. 2 using the *Definition* 1.

x_a	x_b	g_a	g_b
0	0	0	1
0	1	0	1
1	0	1	0
1	1	0	1

(a)

x_c	x_a	x_b	$g_c^\star$	$g_a^\star$	$g_b^\star$
0	0	0	0	0	1
0	0	1	0	0	1
0	1	0	0	1	0
0	1	1	0	0	1
1	0	0	0	1	0
1	0	1	0	1	0
1	1	0	0	0	1
1	1	1	0	1	0

(b)

Fig. 2. The Boolean functions for networks in *Fig*. 1

The 2-nodes Boolean network represents the Boolean function g_a, g_b specified by the table in *Fig.* 2.(*a*), and the the *BBN* represents the Boolean functions $g_c^\star, g_a^\star, g_b^\star$ specified by the table in *Fig.* 2.(*b*). With some knowledge of logic synthesis [7], these Boolean functions can be synthesised as follows:

$$\begin{cases} g_a = ab' \\ g_b = a' + b \end{cases}, \text{and } \begin{cases} g_a^\star = c \oplus (ab') = c \oplus g_c^\star \oplus g_a \\ g_b^\star = c \oplus (a' + b) = c \oplus g_c^\star \oplus g_b \\ g_c^\star = 0 \end{cases}$$

In this format, the *BBN* can be viewed as adding an additional *XOR*-gate to every node. Suppose the Boolean functions of a Boolean network $G(V,F)$ are $g_i, i \in \{1,2,\ldots,n\}$, and the Boolean function of the additional node c is $g_c^\star$. Then we can get the Boolean functions of a *BBN* in equation,

$$g_i^\star = \begin{cases} g_i & \text{, if } g_i = 0 \text{ , or } 1 \\ c \oplus g_c^\star \oplus g_a & \text{, otherwise.} \end{cases} \tag{2}$$

And for different $g_c^\star$, there will be many different *BBNs*. In this paper, four different $g_c^\star$ are found and used in our experiment in *Sect.* 4, which are $\{0, 1, c \oplus f(x_1,\ldots,x_n), c' \oplus f(x_1,\ldots,x_n)\}$.

2.4 Basin Type

There are many different basin types for Boolean networks [19], such as *star-type*, *string-type*, ... As we use *Definition* 1 to describe Boolean networks and construct *BBNs*, the most proper basin type is *star-type*. The *STGs* for a *star-type* basin shows in *Fig.* 3. We see that all the basin states of attractor A_i connect directly to A_i, which just look like stars around the attractor state.

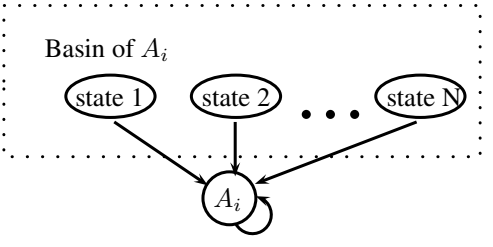


Fig. 3. The *STGs* for star-type basin

3 Robustness of BBNs

Living organisms can sustain a wide variety of genetic changes. Gene regulatory networks and metabolic pathways self-organize and re-accommodate to make the

organism continue performing under many point mutations, gene duplications and gene deletions [18]. The amazing robustness capability of surviving and keeping the species stability under certain changes in the environment is always an important subject of *fault-tolerance*, which has been desired by the electronic industry for a long time.

A single *stuck-at fault* in a circuit can cause serious results. Although we already have some methods to check it out, it becomes harder and harder to check every bit as the system becomes more complex than we can image. However, *BBNs* show a good performance on these faults in *Sect*. 4, of which the whole state space can be split into several small equal components by attractors as showing in *Fig*. 1.(*b*). In the following parts of this section, we first define a yardstick for robustness, and then compare the real *GRNs* with random Boolean network models to insure our idea of constructing *BBNs* to improve the fault-tolerance ability for *single stuck-at faults*.

3.1 Definition

Suppose that we have a Boolean network $G(V,F)$, in which V represents a set of nodes $\{v_1,\ldots,v_n\}, v_i \in \{0,1\}$ and F represents a set of Boolean functions $\{f_1,\ldots,f_n\}, f_i : \{0,1\}^{k_i} \to \{0,1\}$. Each node v_i represents the state of the vertex v_i in the Boolean network. Each Boolean function $f_i(v_{i_1},\ldots,v_{i_{k_i}})$ with k_i specific input nodes are assigned to node v_i and used to update its value. If we draw a directed graph for G, there will be totally n vertices and $\sum_1^n k_i$ edges.

When a *stuck-at fault* happens in Boolean network G, there will be two different cases: *fault on node*, and *fault on edge*. If this *fault* takes place on vertex v_i, then the state of v_i will be stuck at 0 or 1. The network model will degenerate to a smaller Boolean network by removing node v_i. However, this will be inconsistent with our objective – tolerating this fault. If the *fault* takes place on edge e_{ij}, which points from vertex j to vertex i, then the change will only influence vertex i and other vertices still work well. So all the *stuck-at faults* discussed for Boolean networks in this paper take place on the edge.

Definition 2. *A Boolean network* $G(V,F)$ *is constructed with a set of nodes* $V = \{v_1,\ldots,v_n\}, v_i \in \{0,1\}$ *and a set of Boolean functions* $F = \{f_1,\ldots,f_n\}, f_i : \{0,1\}^{k_i} \to \{0,1\}$. *And* tolerating a fault *is defined as when a fault happens in the Boolean network, the attractors still keep the same as the normal case. Then* robustness R *can be defined as the ratio of* tolerating faults *and the total faults, which can be expressed as:*

$$R = \frac{Num_{FT}}{Num_{fault}}, \tag{3}$$

for Num_{FT} *represents the number of faults can be tolerated by Boolean network G, and* Num_{fault} *represents the total number of all possible faults.*

According to *Definition* 2, when we consider only the single *stuck-at faults*, the robustness of a Boolean network R_{ST} can be expressed as:

$$R_{ST} = \frac{Num_{FT_{ST}}}{Num_{fault}} = \frac{Num_{FT_{ST}}}{2\sum_1^n k_i} \tag{4}$$

3.2 Comparisons

In order to show an intuitive understanding of robustness, we compare the robustness of real *GRNs* with our random models. The results are shown in *Table* 1.

Table 1. Robustness for real *GRNs* and $(16,4)$-random Boolean networks

Real *GRNs*				$(16,4)$-random models			
Name	nodes	faults	R_{ST}	Name	nodes	faults	R_{ST}
Ap-1	10	32	0.656	Rdm1	16	128	0.016
Arabidopsis	15	88	0.352	Rdm2	16	128	0.016
MammalianCell	10	78	0.372	Rdm3	16	128	0
BuddingYeast2009	18	120	0.375	Rdm4	16	128	0.078
BuddingYeast2004	12	74	0.257	Rdm5	16	128	0.016
BY2004Modified	11	58	0.603	Rdm6	16	128	0.023
BuddingYeast2008	9	38	0.263	Rdm7	16	128	0
DrosophilaCellCycle	14	84	0.655	Rdm8	16	128	0.336
ERBB2	20	102	0.784	Rdm9	16	128	0
FissionYeast	10	54	0.167	Rdm10	16	128	0
T-cellReceptor	10	78	0.372	Rdm11	16	128	0
ThBoolean	40	116	0.379	Rdm12	16	128	0.25

All the 12 real *GRNs* are taken from [3], and the $(16,4)$-random Boolean networks are produced by our random Boolean network generating program. We can see that the robustness of real *GRNs* is much higher than the one of random Boolean networks. *ERBB2* [16] is the best model in the table with the highest robustness of 78%, which means *ERBB2* can tolerate as high as 78% single stuck-at faults happen on every edge. As a contrast, random models performe so poor that there are only two models with the robustness higher than 20%. These comparisons truly reflecte the objective facts that real *GRNs* and cells are the results of natural evolution, and the most robustness system is cell itself. The only way for the robust design is learning from nature.

4 Experiment Results

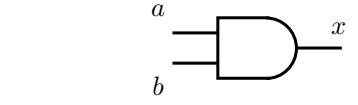
Experiment is designed to evaluate the robustness of *balanced Boolean networks* and the performance of our *BBNs* constructing method. All the *BBN* models are built using the model generating program for the two logic gates – *AND* and *XOR*.

They are not only simple and useful for every logic circiut, but also very convenient for constructing the Boolean function $g_c^\star$ in $Eq.2$. The SAT-based attractor computing program *BNS* [4] is used to compute the attractors, and also the synchronous hardware synthesis and verification program *abc* is used to calculate the cost of 2-inputs *AND* gates [15] in the implementation.

4.1 BBN Models

Choosing proper models are always important for the success of any experiment. To study and evaluate the robustness R_{ST} of *BBNs*, we need lots of *BBN* models and consider all the *stuck-at faults* for them. However, these requirements are conflicting with the poor performance of computers for solving this *NP-hard* problem. For these reasons, the 2–5 inputs *AND* and *XOR* gates are chosen in our experiment.

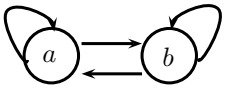
Taking the 2-inputs *AND* gate as an example, we know that this gate has only *two* output values – 0 and 1. When we use different singleton attractors to represent them, a 2-nodes Boolean network and the *BBNs* can be constructed via $Eq.2$. The whole process of constructing a *BBN* is shown in $Fig.4$.



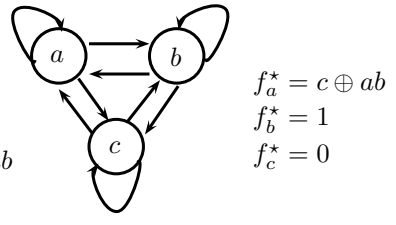
(a) 2-inputs *AND* gate

a	b	x	state
0	0	0	01
0	1	0	01
1	0	0 $\Rightarrow$	01
1	1	1	11

(b) Truth table for *AND*



(c) A 2-nodes Boolean network



(d) A *BBN* for 2-inputs *AND*

Fig. 4. The whole process for constructing a *BBN*

In $Fig.4.(b)$, we using attractors 01 and 11 to represent 0 and 1 in the truth table, a 2-nodes Boolean network can be built as it shows in $Fig.4.(c)$. So we can easily get a balanced Boolean network as it shows in $Fig.4.(d)$. The robustness R_{ST} of this *BBN* is 33.3%, a little higher than the 2-nodes Boolean network, which is only 25%.

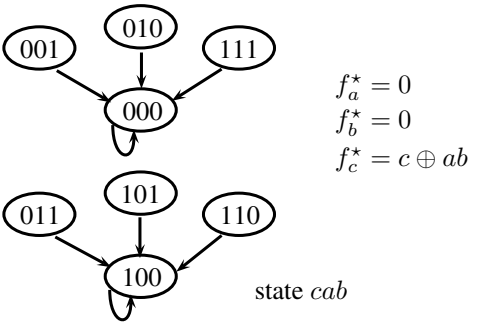


Fig. 5. The Best *BBN* for 2-inputs *AND*

A *BBN* is uniquely determined by the original Boolean network and $g_c^\star$. When we use different combinations of attractors and $g_c^\star$, a lot of *BBNs* can be generated. In the way, the best *BBN* for 2-inputs *AND* can be found. *Figure* 5 shows the *STGs* and Boolean functions of the best *BBN*, which has the highest R_{ST} for 66.7%.

4.2 Results for AND *and* XOR *Gates*

With the *BBNs* built in *Sect*. 4.1, we get the results of robustness R_{ST} and cost for 2–5-inputs *AND* and *XOR* gates, which are separately listed in *Table* 2 and *Table* 3. From these tables, *BBNs* for *AND* gates perform much better than *XOR* gates. All the maximum R_{ST} of *BBNs* for *AND* gates are larger than 0.65, but they are all below 0.45 for *XOR*. Also, there is a very interesting phenomenon in *Table* 3, that the results for *BBNs* are always the same with the Boolean networks with one more node. One explanation is the method used for constructing *BBNs*. We use *XOR* in *Eq*. 2 for *BBNs*.

Table 2. Experimental results for *AND* gate

	Boolean Networks		*BBNs*			
Name	R_{ST_max}	cost	R_{ST_max}	cost	$Rise_{R_{ST}}$	$Rise_{cost}$
2-inputs *AND*	0.25	1	0.67	4	1.67	3.00
3-inputs *AND*	0.50	2	0.75	6	0.50	2.00
4-inputs *AND*	0.50	3	0.80	6	0.60	1.00
5-inputs *AND*	0.50	4	0.83	7	0.67	0.75

Column $Rise_{R_{ST}}$ shows that *BBNs* are much more robust than the original Boolean networks. However, this improvement of robustness is made by the high cost of redundancy in the network structure [6]. While the increasing of nodes in Boolean networks, the increase of $Rise_{R_{ST}}$ for *AND* is much more obvious than *XOR*. However, the decrease of column $Rise_{cost}$ is much more attractive. This trend in $Rise_{cost}$

Table 3. Experimental results for *XOR* gate

Name	Boolean Networks		*BBNs*			
	R_{ST_max}	cost	R_{ST_max}	cost	$Rise_{R_{ST}}$	$Rise_{cost}$
2-inputs *XOR*	0.25	3	0.33	9	0.33	2.00
3-inputs *XOR*	0.33	9	0.38	15	0.13	0.67
4-inputs *XOR*	0.38	15	0.40	21	0.07	0.40
5-inputs *XOR*	0.40	21	0.42	27	0.04	0.29

means that in large Boolean networks, there are more common components, which can be used to construct *BBNs*. This result will be useful for improving the robustness of large Boolean networks.

4.3 Advantages of Our Method

In this subsection we present results demonstrating advantages of our method for constructing *BBNs*. *Table* 4 shows the results of the average robustness during the experiment in *Sect*. 4.2. The first row shows the average robustness for all the original Boolean networks; and the second row is for *BBNs*. From the values in the third row, we see that the robustness R_{ST} of *BBNs* improve at least 13% in average.

Table 4. Comparisons of average R_{ST}

Number of inputs	*AND* gate				*XOR* gate			
	2	3	4	5	2	3	4	5
$R_{ST\_average}$	0.25	0.38	0.44	0.47	0.25	0.17	0.19	0.20
$R_{ST\_average}$ for *BBNs*	0.39	0.47	0.51	0.53	0.33	0.27	0.26	0.25
$Rise_{R_{ST\_average}}$	56%	25%	17%	13%	33%	61%	38%	27%

Table 5 shows the results of the best *BBNs*, all of which have the same value of robustness R_{ST}, no matter which method is used. The first row shows the number of best *BBNs* found by using *Eq*. 2, and the second row shows the result found by the easy method mentioned in *Sect*. 2.3 – construct all possible *BBNs* and search for the best. Although we do not find all the best *BBNs* using *Eq*. 2, this result is still as good as we expect.

Table 5. Comparisons for the best BBNs

Number of inputs	*AND* gate				*XOR* gate			
	2	3	4	5	2	3	4	5
Best *BBNs* by *Eq*. 2	1	4	11	26	12	32	80	192
Best *BBNs* of all	1	10	67	406	12	32	80	192

5 Conclusion and Open Problems

The paper formalizes *BBNs* and introduces a method for constructing *BBNs* for 2-singleton attractors Boolean networks. Using *BBNs* to improve the robustness of Boolean networks is a new idea. The experiment results show that *BBNs* are capable to tolerate the single stuck-at faults on every edge, and the robustness for the *BBNs* constructed in our method improves at least 13% in average. Although this increase is not so high, it's still useful for large models to avoid huge computations. However, the research of *BBNs* is just in its beginning. Much more issues should be considered to make it complete, such as string basin, mixed basins, and cycle attractors. The effect of these factors demands a further study.

References

1. Albert, R., Othmer, H.G.: The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in drosophila melanogaster. Journal of Theoretical Biology 223, 1–18 (2003)
2. Alberts, B., Bray, D., Lewis, J., Ra, M., Roberts, K., Watson, J.D.: Molecular Biology of the Cell. Garland Publishing (1994)
3. Dubrova, E., Liu, M., Teslenko, M.: Finding attractors in synchronous multiple-valued networks using SAT-based bounded model checking. In: Proceedings of the 2010 40th IEEE International Symposium on Multiple-Valued Logic (2011)
4. Dubrova, E., Teslenko, M.: A SAT-based algorithm for computing attractors in synchronous Boolean networks. IEEE/ACM Trans. Comput. Biol. Bioinformatics 8, 1393–1399 (2011)
5. Espinosa-Soto, C., Padilla-Longoria, P., Alvarez-Buylla, E.R.: A gene regulatory network model for cell-fate determination during arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. The Plant Cell 16, 2923–2939 (2004)
6. Gershenson, C., Kauffman, S.A., Shmulevich, I.: The role of redundancy in the robustness of random boolean networks. In: Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems (2006)
7. Hassoun, S., Sasao, T. (eds.): Logic Synthesis and Verification. Kluwer Academic Publishers, USA (2002)
8. He, C., Ren, Q.: Robustness during network evolution. In: International Conference on Complex, Intelligent and Software Intensive Systems, CISIS 2009, pp. 1240–1244 (2009)
9. Helikar, T., Kochi, N., Konvalina, J., Rogers, J.A.: Boolean modeling of biochemical networks. The Open Bioinformatics Journal 5, 16–25 (2011)
10. Huang, S., Ingber, D.E.: Shape-dependent control of cell growth, differentiation, and apoptosis: Switching between attractors in cell regulatory networks. Experimental Cell Research 261, 91–103 (2000)
11. Kauffman, S.A.: Metabolic stability and epigenesis in randomly constructed genetic nets. Journal of Theoretical Biology 22, 437–467 (1969)
12. Kauffman, S.A.: The Origins of Order: Self-Organization and Selection of Evolution. Oxford University Press (1993)
13. Kauffman, S.A.: At Home in the Universe: The Search for the Laws of Self-Organization and Complexity. Oxford University Press (1996)

14. Li, F., Long, T., Lu, Y., Ouyang, Q., Tang, C.: The yeast cell-cycle network is robustly designed. Proceedings of the National Academy of Sciences of the USA (PNAS) (2004)
15. Mishchenko, A., Chatterjee, S., Brayton, R.: Dag-aware aig rewriting: a fresh look at combinational logic synthesis. In: Proceedings of the 43rd Annual Design Automation Conference (2006)
16. Sahin, O., Fröhlich, H., Löbke, C., Korf, U., Burmester, S., Majety, M., Mattern, J., Schupp, I., Chaouiya, C., Thieffry, D., Poustka, A., Wiemann, S., Beißbarth, T., Arlt, D.: Modeling ERBB receptor-regulated G1/S transition to find novel targets for de novo trastuzumab resistance. BMC Systems Biology 3, 1–20 (2009)
17. Schmal, C., Peixoto, T.P., Drossel, B.: Boolean networks with robust and reliable trajectories. New Journal of Physics 12, 113, 054 (2010)
18. Wagner, A.: Robustness and evolvability in living systems. Oxford University Press (2007)
19. Wuensche, A.: Basins of Attraction in Network Dynamics: A Conceptual Framework for Biomolecular Networks. Chicago University Press (2004)

Structural Evolution in Knowledge Transfer Network: An Agent-Based Model

Haoxiang Xia, Yanyan Du, and Zhaoguo Xuan

Abstract. We use an agent-based model to study the effect of knowledge transfer on the structural evolution of a social network. In the proposed model, the agents exchange knowledge with their network neighbors; and simultaneously they adjust their neighbors by edge-rewiring in order seek better chance for knowledge transfer. This gives rise to the coevolution of the population's knowledge state and the network topology. Through computational simulations, interesting phenomena are observed, most notably the disassembly and reassembly of the network connectivity and the emergence of the small-world structure that is self-organized from the initial random network. The underlying mechanisms are partly analyzed.

1 Introduction

With the growing research interests on complex networks, the collective dynamics on complex social networks have also been extensively studied [Castellano et al. 2009]. One noticeable sub-area is the dynamics of knowledge transfer, which can essentially be traced back to the studies on the diffusion of innovations and technologies in economics and management science [Coleman et al. 1957], and in parallel on social cognition and social learning in social psychology [Fiske & Taylor 1991]. The attentions on knowledge transfer have become more widespread with the prominence of knowledge management since the 1990s [Argote & Ingram 2000]. In the context of social networks, the studies on the dynamics of knowledge transfer are mostly focused on the effect of the network properties on the performance of knowledge transfer [Reagans & McEvily 2003]. As the network topology is one key network property that has great impact on knowledge transfer, knowledge transfer on different types of network structures has also been extensively studied [Cowan & Jonard 2004, Kim & Park 2009].

Haoxiang Xia · Yanyan Du · Zhaoguo Xuan
Institute of Systems Engineering, Dalian University of Technology
Dalian 116024 China

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424,, pp. 31–38.
springerlink.com

These researches are doubtlessly valuable, as they greatly contribute to improve our understandings on how the network properties influence the performance of knowledge transfer. However, there is another side of the coin that the network can in turn be influenced by the collective action of knowledge transfer. Generally, this reverse problem of the effect of the collective action on the structural evolution of the network has partially been tackled in some recent work on the "adaptive coevolutionary networks" [Gross & Blasius 2008]. In the community of knowledge transfer, nevertheless, this issue is less well-addressed. Although some authors have conceptually and empirically discussed the coevolution of the social and knowledge networks [Palazzolo et al. 2006, Roth & Cointet 2010], in those contributions there was an insufficiency of in-depth analysis about the underlying mechanisms that boost such coevolution. In particular, it was not well-explained how the collective knowledge activities affect the structural evolution of the social network.

Based on the above observations, we in this paper give a primitive attempt, by using an agent-based computational model, to explore the coevolutionary dynamics of social network and knowledge, especially to study the structural evolution of the social network that is affected by the knowledge transfer activities between the participating actors. We hope our virtual experiments in the computer world may give implications to understand the underlying mechanisms that govern the structural evolution of the knowledge transfer networks in the real world.

2 Model Description

The proposed model is about a set of agents that interact with one another to exchange knowledge. Each agent contains "knowledge". The knowledge level of agent i is denoted as a real value v_i~U[0.0,10.0]. The agents are then interconnected with one another to form a social network, in which the vertices are the participating agents and the edges are the social relations between the agents. The overall executive procedure of the proposed model can be described as follows:

1. Configure the initial network as a random network with given N vertices and M edges; and initialize each vertex (i.e. agent) and its knowledge vector.
2. Arbitrarily select an agent from the network as the focal agent;
3. The focal agent either exchanges knowledge with a neighbor or adjusts its neighborhood in the network. With probability p, the agent exchanges knowledge with one of the neighbors using a Knowledge Transfer or KT rule; otherwise, a Neighborhood Adjustment or NA rule is applied so that the focal agent rewires one existing link to a new agent.
4. Update the knowledge status of each agent and the entire network structure;
5. Repeat steps 2), 3), and 4) until the count of the iterations reaches the prespecified upper-limit T_{max}.

In this procedure, the KT and NA rules need to be specified in more detail.

The key idea behind the Knowledge Transfer or KT rule is that the transfer of knowledge is most effective when the "knowledge diversity" between the two interacting agents is neither too large nor too small [Scholl 1996]. The great diversity between communicating partners would lead to ineffective learning due to the communication difficulties. On the contrary, the amount of transferred knowledge would also be small in the case of small diversity due to the small potential in knowledge transfer. Following this view, we define a "knowledge-exchange threshold", which is denoted as d, as a measure for the upper limit of knowledge diversity for the success of knowledge transfer. When the knowledge diversity is beyond this threshold, we predict no knowledge is transferred. Based on this threshold, the knowledge transfer between two interacting agents i and j can then be specified. Suppose agent i is with lower knowledge level and the difference in their knowledge levels is $d_{ij}=v_j-v_i$. Then, the amount of knowledge that agent i can learn from agent j is determined by the following equation:

$$k_{ij} = \begin{cases} \min\{\max\{\alpha * (v_j - v_i), ks\}, d_{ij} < d \\ 0, d_{ij} \geq d \end{cases} \quad (1)$$

In equation (1), α is the knowledge transfer rate, and ks is the upper-bound of knowledge that can be transferred within one interaction. The knowledge level of agent i updates accordingly, namely $v_i(t+1)=v_i(t)+k_{ij}(t)$, whereas that of agent j keeping unchanged.

We then turn to the Neighborhood Adjustment or NA rule. The basic idea is that the agents tend to remove the links that are disadvantageous for knowledge transfer so as to seek better knowledge transfer possibilities in the population. The neighborhood adjustment process is then accomplished by the following steps.

- With probability w, agent i is rewired to a randomly-selected neighbor of its original neighbors, or otherwise (i.e. with probability 1-w) to a random one selected from the whole population except for its current neighbors.
- If agent i has been successfully rewired to a new neighbor, one of its original links is to be removed. If there is at least one neighbor whose knowledge diversity to agent i is larger than the threshold d, remove the link to the neighbor with the largest knowledge diversity; otherwise, remove the link to the neighbor with the smallest knowledge diversity with agent i.

To sum up, in the proposed model, three parameters, namely d, p, and w, are the key factors that influence the structural evolution of the network as well as the performance of knowledge transfer on the network. In the next section, we report the computational simulations that examine the effects of these three parameters.

3 Simulation Results and Analyses

With the previous model, we conduct computational simulations to investigate the coevolutionary dynamics of network and knowledge, in particular to examine how the knowledge status of the agents and the actions of knowledge transfer affect the

network structure. In our simulations, we set the agent population N=500 and the edge account M=5,000; the upper-limit of amount of knowledge to be transferred in one interaction, ks, is set to 0.4. The knowledge transfer rate α is set to 1.0. Owing to the upper limit of knowledge level is 10, we let the knowledge exchange threshold d range from 0 to 7.

The initial network is set to be a random network, the connectivity of which is basically guaranteed since the mean edges per vertex are enormously greater than ln(N) [Watts & Strogatz 1998]. In the proposed model, the structural evolution of the network is through an edge-rewiring process. In general, the network would still be random if the edge-rewiring process is fully-random. In this case, the process is trivial since the network structure keeps statistically unchanged in a high-entropy state. However, as the rewiring process is not fully-random in the the proposed model, we may anticipate the emergence of some nontrivial structure during the evolution of the network.

First, we can examine the dynamics of the network connectivity. It is obvious that the value of parameter w is negatively related the network connectivity, as local clusters are easier to form through neighborhood adjustment when w is high. The entire population is then prone to split into multiple local clusters that are isolated with one another. However, the influences of parameters d and p are not straightforward. Hence, we test the influences of these two parameters by keeping parameter w fixed in our simulations. Fig.1. plots the network connectivity, which is measured by the fraction G of agents that belong to the largest connected subgraph of the whole graph, and the average knowledge level V of all the agents, within the d-p parameter space. Parameter w is fixed to 0.3, and the iteration time T is 100,000. The plotted data is based on the averaging of 10 simulation results under the same initial conditions.

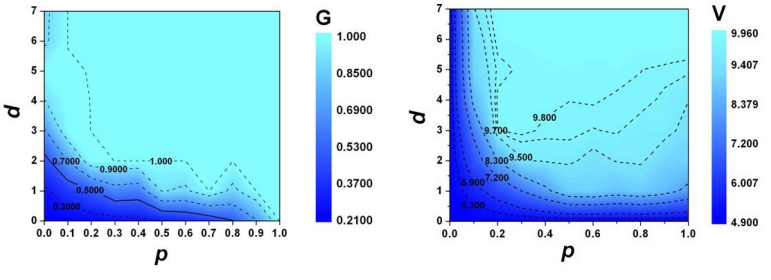


Fig. 1. Parameters d and p influence network connectivity (left) and the average knowledge level (right) of the entire population (w=0.3, iteration time T=100,000)

From the left part of Fig. 1, it can be observed that parameter p is positively related to the connectivity of the entire network except for some small areas in the d-p space. This phenomenon is reasonable since the low p value indicates high probability of edge-rewiring; and this increases the risk for the network to separate as there is the local-grouping factor in the rewiring mechanism.

It can also be observed in Fig. 1. that parameter *d* is another positive factor for network connectivity. This is not so straightforward but we can partly explain this phenomenon by examining the NA rule. A bifurcation structure is implicitly contained in the NA rule. When *d* is high, the agent-pairs with low knowledge diversity are more likely to be disconnected; consequently, the strongly-connected local cliques tend to dissolve and the long-range links that connect agents with diverse knowledge levels have more chance to survive. On the contrary, for small *d*, the agent-pairs with high knowledge diversity are more likely to be disconnected. The strongly-connected local clusters are more likely to survive; and the entire population is more likely to be separated into multiple small clusters.

Comparing the left and right parts of Fig.1, it can be seen that the average knowledge level *V* also increases as parameters *d* and *p* increase. This indicates that the increase of the network connectivity may augment the transfer of knowledge and vice versa. Another interesting observation is that in the connected regime in the *d*-*p* space (i.e. the area of G=1.0 in the left part), the final average knowledge levels are not identical. This indicates in the network structure and the knowledge transfer patterns may diverge in the connected regime. Subsequently we attempt to give further analysis on the network dynamics in this regime.

Our experiments show that, when $d>=4$ and $p>=0.7$, the network connectivity is guaranteed in the whole process of simulation, regardless of the value of *w*. To the other extreme, when $d<=2$ and $p<=0.2$, the network becomes unconnected through edge-rewiring even when $w=0$. Between the two extremes, the value of *w* has significant impact on the network connectivity. One interesting phenomenon is the disassembly and reassembly of the connectivity in various *d*-*p* combinations. Fig. 2 illustrates this process when $d=3$ and $p=0.2$.

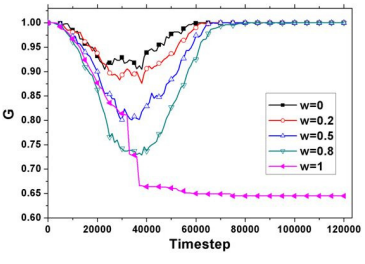


Fig. 2. The fraction G as a function of time for $d=3$, $p=0.2$ and different values of *w*

As shown in Fig.2, in our simulations the network connectivity is descending at the early stage of simulation, for each $w \in [0,1]$. However, an inflexion point occurs at around a time-step between 30,000 and 40,000, when *G* stops decreasing and turn to increase for any *w* that is less than 1. The connectivity of the entire network is reestablished at around the 80,000th time-step. The exception is the case of $w=1$, where the network connectivity does not reestablish; instead, the declining of *G* rapidly slows down and a steady-state at around G=0.65 is eventually reached, indicating that the whole network is formed by a giant connected sub-

group and various smaller groups. This process of disassembly and reassembly shows that in some areas in the *d*-*p* parameter space, although the final state of the network is connected, the connectivity does not persist in the whole process of the structural evolution. When the values of parameters *d* and *p* are small, the network is likely to form local clusters of low knowledge diversity; and the entire network tends to be fragmented at the early stage of simulation when the overall knowledge diversity is high. With the simulation continues, the overall knowledge diversity diminishes and the reestablishment of the links between the previously-isolated clusters becomes feasible. The chance of global or remote linking is ensured by the condition $w<1$. If $w=1.0$, the isolated local clusters are stabilized by the strong social cohesion, i.e., the rewiring only takes place in local groups.

What's more, we can also observe the emergence of a "small-world" during network evolution in the regime of $G=1$. When the parameters *d* and *p* are greater than the situation as Fig.2 shows, it is common that the network self-organize to form a small-world for a period of time during the whole process of network evolution. This phenomenon is illustrated in Fig.3.

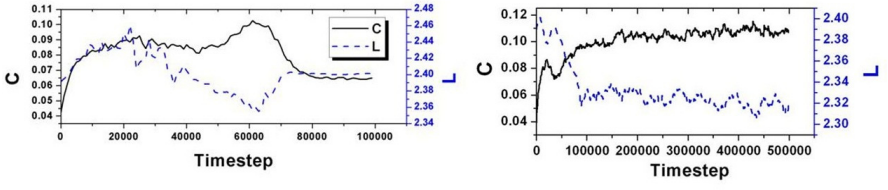


Fig. 3. Emergence of a "small-world" in the proposed model (d=6, p=0.5, w=0.5), with no knowledge creation (left), and with repetitive knowledge creation (right)

The left part of Fig.3 show the time-evolution of the network structure in terms of the cliquishness and the characteristic path length. It can be seen that a "small-world" gradually shapes in the first 60,000 time-steps, since it is with relatively high cliquishness measured by *C* (the clustering coefficient) and simultaneously short path lengths measure by *L* (the characteristic path length). After around 80,000 time-steps, nonetheless, the small-world diminishes and the network becomes a random network once more. In comparison, our computational experiments also show that such "small-world" phenomenon is not observed if setting *p* to zero. This means that the edge-rewiring itself does not generates the small world from a random network, without the transfer of transfer.

What's more, when we add a knowledge creation mechanism into our model, by letting a fixed small set of agents increase their knowledge level every 1,000 steps and proportionally increasing the knowledge exchange threshold according to the average knowledge level of the entire population, we find that the small-world phenomenon sustains, as shown in the right part of Fig.3. The major difference between the situations represented in the left and right parts of Fig. 3 lies in that the overall knowledge diversity is maintained by the repetitive addition of new knowledge. It is then natural to conjecture that the maintenance of knowledge diversity is critical for the sustaining of the small-world.

Putting the prior results together, we can draw a rough picture for the overall evolutionary dynamics of the network. In particular, the influence of the key parameter *d* can partly be analyzed. In the proposed model, the knowledge diversity within local groups and at the global level plays a vital role for network evolution. The parameter *d* leverages the agents' tendency of "homophily" and "heterophily" [Rogers & Bhowmik 1970] for neighbor selection. For low *d*, the homophily mechanism dominates the neighborhood adjustment and local clusters that are comprised of agents with low knowledge diversity are likely to form; for high *d*, by contrast, the influence of heterophily becomes more prominent so that the edges between diverse agents are more stable. Thus, in the early stage of a simulation, when the average knowledge diversity between agents is high, the network structure may evolve through different routes in accordance with the value of parameter *d*. For high *d*, the overall connectivity can easily retain and the network is generally random. When *d* is low, the network tends to split into isolated cliques due to the dominance of homophily, this is a fragmentation state of the network. Between the connected random network state and the fragmentation state, a small-world network state may emerge when *d* is at a middle level. In this "small-world" state, the agents are generally clustered into local groups; but the local groups are not extremely homogenous in terms of knowledge so that long-range edges that connect different cliques persist. With the continuing transfer of knowledge, all the agents may reach a high knowledge level (up to 10.0 in the model) and the overall he knowledge diversity diminishes, correspondingly there is a reestablishment of the random network as the rewiring becomes fully arbitrary. If the network becomes highly fragmented, the isolation between the local clusters hinders further knowledge transfer and the connectivity of the whole network is not able to reassemble. If the knowledge diversity is maintained by continual addition of new knowledge, the network would not return to the state of the connected random network and the small world persists, as shown in the right part of Fig.3.

4 Concluding Remarks

In this paper we develop an agent-based model for testing the effects of knowledge transfer on the structural evolution of the social network. Some interesting phenomena can be observed in our computational experiments. Most notably, starting from a connected random network, we observe the network can for a period of time in the evolutionary process be fragmented into isolated cliques or become a "small-world", depending on the parameter setting. The small world sustains if the overall knowledge diversity is maintained by repetitive creation of new knowledge. The results obtained in this work may imply more general mechanisms for the dynamics of social networks. We are now underway to generalize our model and to give more thorough analysis, in order to examine how the mechanisms of homophily-and-heterophily and clustering-and-randomization work in combination to shape the dynamics of social network. In particular, we hope our work will improve the understandings of the small-world dynamics, by enriching Watts and Strogatzs's [1998] classic model on this subject.

Acknowledgments. This work is partly supported by the Natural Science Foundation of China under Grant Nos. 70871016 and 71031002, respectively. The authors are grateful for the constructive suggestions and comments by the anonymous reviewers.

References

Argote, L., Ingram, P.: Knowledge transfer: a basis for competitive advantage in firms. Organizational Behavior and Human Decision Processes 82(1), 150–169 (2000)

Castellano, C., Fortunato, S., Loreto, V.: Statistical physics of social dynamics. Reviews of Modern Physics 81(2), 591–646 (2009)

Coleman, J., Katz, E., Menzel, H.: The diffusion of an innovation among physicians. Sociometry 20(4), 253–270 (1957)

Cowan, R., Jonard, N.: Network structure and the diffusion of knowledge. Journal of Economic Dynamics and Control 28, 1557–1575 (2004)

Fiske, S., Taylor, S.E.: Social cognition, 2nd edn. McGraw-Hill, New York (1991)

Gross, T., Blasius, B.: Adaptive coevolutionary networks: a review. Journal of the Royal Society Interface 5(20), 259–271 (2008)

Kim, H., Park, Y.: Structural effects of R&D collaboration network on knowledge diffusion performance. Expert Systems with Applications 36(5), 8986–8992 (2009)

Palazzolo, E., Serb, D., She, Y., Su, C., Contractor, N.: Co-evolution of communication and knowledge networks as transactive memory systems. Communication Theory 16(2), 223–250 (2006)

Reagans, R., McEvily, B.: Network structure and knowledge transfer: The effects of cohesion and range. Administrative Science Quarterly 48, 240–267 (2003)

Rogers, E.M., Bhowmik, D.K.: Homophily-heterophily: relational concepts for communication research. Public Opinion Quarterly 34(4), 523–538 (1970)

Roth, C., Cointet, J.-P.: Social and semantic coevolution in knowledge networks. Social Networks 32(1), 16–29 (2010)

Scholl, W.: Effective teamwork: A theoretical model and a test in the field. In: Witte, J., Davis, J. (eds.) Understanding Group Behavior, vol. 2, pp. 127–146. Erlbaum, Hillsdale (1996)

Watts, D.J., Strogatz, S.H.: Collective dynamics of "small-world" networks. Nature 393(6684), 440–442 (1998)

Using Network Science to Define a Dynamic Communication Topology for Particle Swarm Optimizers

Marcos A.C. Oliveira Junior, Carmelo J.A. Bastos Filho, and Ronaldo Menezes

Abstract. We propose here to use network sciences, specifically an approach based on the Barabási-Albert model, to define a dynamic communication topology for Particle Swarm Optimizers. We compared our proposal to previous approaches, including a simpler Barabási-Albert-based approach and other most used approaches, and we obtained better results in average for well known benchmark functions.

1 Introduction

Particle Swarm Optimization (PSO) is a swarm intelligence technique that has been widely used to solve optimization problems in hyper-dimensional search spaces with continuous variables. PSO was first proposed by Kennedy and Eberhart in 1995 [15] and it was inspired by the social behavior of flocks of birds working together to find food. In the PSO paradigm, each particle in the swarm represents a candidate solution in the fitness function domain. During the algorithm execution, each particle adjusts its velocity and position based on the current position, the current velocity, the best position achieved by itself during the search process so far and the best position obtained by the particles among a pre-determined neighborhood during the search process so far.

There are a few important issues that influence on the convergence velocity and on the quality of the final solution returned at the end of the algorithm execution. Among them are: the equation used to update the velocities of the particles, the mechanisms deployed to avoid explosion states, the quality of the Pseudo Random Number Generator (PRNGs) and the communication scheme adopted to exchange

Marcos A.C. Oliveira Junior · Carmelo J.A. Bastos Filho
University of Pernambuco, Brazil
e-mail: carmelofilho@ieee.org

Ronaldo Menezes
Florida Institute of Technology, USA
e-mail: rmenzes@cs.fit.edu

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 39–47.
springerlink.com

information among the particles. There are several works that tackle the three former issues [11, 8, 4]. The latter has been been widely discussed since it defines the neighborhood of the particles and, as a consequence, determines how the information flows through the whole swarm [6, 16].

Previous works have shown that less connected topologies slow down the information flow, since the information about the convergence is transmitted indirectly through intermediary particles [16]. On the other hand, highly connected topologies diminish the average distance between any pair of individuals. As a consequence, there is a tendency for the whole swarm to move quickly toward the first local optimum found by any particle of the swarm when the average distance between nodes is too short (*e.g.* a small-world topology). Unfortunately, in simple and static communication schema, fast convergence generally means premature convergence to a local optimum, specially in multimodal search spaces [6].

Recently, many efforts have been made to analyze how to link components in complex systems [17]. Some examples are social networks, World Wide Web, power grids [20] and biochemical networks [18]. In all these systems, there are several aspects that can be analyzed, such as the way these components can interact with themselves, or the pattern of connections between the components, which is in general highly correlated with the system behavior.

Until the last decades, perhaps due to the lack of deeper analysis or because of the limited processing capacity of computer, real-world networks were usually seen as a result of a completely random process [2]. Indeed, the study of real networks has gained relevance since they present many interesting features, such as fast spread of information through the network compounds, robustness, reliabilty [9, 10, 7].

Barabási and Albert showed that large real networks follow a scale-free power law distribution. They pointed out that this feature was a consequence of two underlying mechanism: (*i*) networks expand continuously by addition of new vertices; and (*ii*) new vertices usually attach to nodes that are already well connected [3]. Thus, they proposed a model, known as Barabási-Albert model (BA model), consisting of an algorithm for generating random scale-free networks using a preferential attachment mechanism [1]. A variation of the BA model, called Bianconi-Barabási model, that the probability of a node to connect to one another is given by a term that depends on the fitness of the involved node [5].

The idea of preferential attachment and complex networks was already proposed to define the PSO communication topology, as in the work of Godoy and von Zuben [13]. In this approach, the PSO starts with a scale-free topology generated by the BA model, and then, the particles are connected or disconnected along the iterations depending on the fitness of the particles. One may notice that there are some undesired outcomes from this approach: (*i*) since the swarm is initiated with a small-world topology, probably the swarm will present a high probability to be stuck in a local minima, specially in multimodal search spaces; (*ii*) the algorithm is not quick for connecting and disconnecting particles, and this behavior is not a desired feature for dynamic or multimodal problems; and (*iii*) the mechanism used to reconnect the particles does not take into account the past of the particle, it solely depends on the current fitness of the particle.

In this paper, we propose a novel approach to define the dynamic topology based on preferential attachment. The proposal aims to balance information flow in the swarm. The topology is initiated as a local topology and evolves to allow the particles to increase the communication capability when it is necessary. Besides, it also considers if the particles are improving or not their solutions. The paper is organized as follows: we briefly review the Particle Swarm Optimization in the next section. In Section 3, we present our proposal to define a dynamic communication topology. The simulation setup and results are given in Section 4. Finally, we present our conclusions and suggest some future works in the last section.

2 Particle Swarm Optimization

Particle Swarm Optimization (PSO) is composed by a swarm of particles, where each particle has a position within the search space $\mathbf{x_i}(t)$ and each position represents a possible solution for the optimization problem. The particles fly through the search space of the problem searching for the best solution. Each particle updates its position according to the current velocity $\mathbf{v}_i(t)$, the best position found by the particle itself $[\mathbf{P}_{best_i}(t)]$ and the best position found by the neighborhood of the particle i during the search so far $[\mathbf{N}_{best_i}(t)]$.

Therefore, the velocity and the position of every particle are updated iteratively by applying the following update equations for each particle in each dimension d:

$$\mathbf{v_i}(t+1) = \mathbf{v_i}(t) + r_1 c_1[\mathbf{P}_{best_i}(t) - \mathbf{x_i}(t)] + r_2 c_2[\mathbf{N}_{best_i}(t) - \mathbf{x_i}(t)], \quad (1)$$

$$\mathbf{x_i}(t+1) = \mathbf{x_i}(t) + \mathbf{v_i}(t+1), \quad (2)$$

where r_1 and r_2 are numbers randomly generated by an uniform distribution in the interval $[0,1]$. c_1 and c_2 are the cognitive and the social acceleration constants, respectively. The original PSO updates the velocities of the particles considering the current value for the velocity of the particles, as presented in equation (1). Clerc [8] performed a study on the dynamic of the particles and stated a parameter known as the constriction factor (χ) that avoids the explosion state. χ is defined in equation (3). The velocity update equation is depicted in equation (4).

$$\chi = \frac{2}{|2 - \varphi - \sqrt{\varphi^2 - 4\varphi}|}, \ \varphi = c_1 + c_2, \quad (3)$$

$$\mathbf{v_i}(t+1) = \chi \cdot \{\mathbf{v_i}(t) + r_1 c_1[\mathbf{P}_{best_i}(t) - \mathbf{x_i}(t)] + r_2 c_2[\mathbf{N}_{best_i}(t) - \mathbf{x_i}(t)]\}. \quad (4)$$

The way the information flows through the particles is determined by the communication topology used by the swarm [12]. The topology of the swarm defines the neighborhood of each particle, that is the subset of particles which the particle is able to communicate with [6]. In the context of social networks, there are many factors that influence the flow of information between nodes [19, 20]. These aspects include the degree of connectivity among the nodes, the average number of neighbors in common per node and the average shortest distance between nodes.

Kennedy and Mendes analyzed these factors on the particle swarm optimization algorithm [16]. It has been shown that the presence of intermediaries slows the information flow down. On the other hand, the information moves faster if more pairs of individuals are connected. Thus, when the average distance between nodes are too short, there is a tendency for the population to move quickly toward the best solution found in earlier iterations. For simple unimodal problem, it usually implies in a faster convergence to the global optimum. However, this fast convergence might means a premature convergence to a local optimum, specially in multi-modal problems [6]. In this case, communication topologies with intermediaries, *i.e.* with a lower number of connections, could help to reach better results.

A first communication model proposed by [14] to model the natural behavior of flocks of birds presented a dynamic topology based on the distance between the particles. However, due to the high computational cost, it was discarded, albeit the similar behavior of flocks of birds [6]. The global topology, which is often known as $\mathbf{G}_{best}$, is a static topology proposed in the PSO white paper [15]. In the $\mathbf{G}_{best}$, all the particles of the swarm are neighbors of each particle of the swarm. This means that the social memory of the particles is shared by the entire swarm. This topology leads to a fast convergence, since the information spreads quickly. On the other hand, in less connected topologies, each particle only shares information with a subset of the swarm. Thus, the social memory is not the same for the whole swarm. The most used local topology is called $\mathbf{L}_{best}$. In the $\mathbf{L}_{best}$ approach, each particle has two neighbors and the neighbor is based on the index. For example, the neighbors of particle #2 are particles #1 and #3. The $\mathbf{L}_{best}$ helps to avoid a premature attraction of all particles to a single spot of the search space, once the information is spread slowly and the swarm has more chances to explore different regions of the search space. Nevertheless, it presents a slower convergence. The two extreme behaviors of the $\mathbf{G}_{best}$ and $\mathbf{L}_{best}$ topologies have encouraged efforts to propose approaches that can present fast convergence while avoiding local minima. Indeed, many other topologies were already proposed, such as *von Neumann*, *Focal*, *Four Clusters*, *Clan PSO*, among others.

Godoy and von Zuben proposed to use a scale-free based topology, called Complex Neighborhood based Particle Swarm Optimization (*CNPSO*)[13]. The evolution of the topology is based on the Barabási-Albert model and it tries to maintain the scale-free characteristic of the topology, while the optimization is being performed. In the *CNPSO*, the swarm topology starts with a scale-free topology generated by the BA-model and it does not take into account any particle information. Thus, it is possible to have a bad particle as a hub in the swarm. Moreover, the initial topology has a small mean-shortest path length. This feature is not desirable in the initial stages of the algorithm because it can attract the swarm to a local optimum in earlier iterations, since the information flows fast. The *CNPSO* reconnecting mechanism also does not take into account the fitness information through the iterations. For example, it does not matter if the particle has stagnated or not in a local optimum. After *times* number of iterations, random particles will have its connections mutated even if they are having success or not. Therefore, this approach is not

dynamic in the sense that its mechanism is not based on the swarm condition but rather it is based on random particles in any state.

3 Our Proposal

We aim to create a dynamic topology which can balance the search behavior of the swarm. It begins with the swarm being less connected. As a consequence, the swarm will present a high capacity to explore along the entire search space. Besides, it is desirable to change the communication scheme of the particles as they reach a stagnation state. Thus, in order to state when a particle k is stagnated, a new attribute, named $P_k failures$, is included to the particles. If the particle k does not improve its position in the current iteration, $P_k failures$ is incremented, otherwise $P_k failures$ is set to zero.

As each particle tries to find better particles to be connected with, there is a preferential attachment connecting mechanism based on the particles fitness. Therefore, to have this mechanism, we used a roulette wheel based on a rank that depends on the fitness of the particles. The best particles have more chances to be chosen for new connections. The proposed algorithm is shown in Algorithm 1.

Algorithm 1. Pseudocode of our proposal

```
1  Generate the neighborhood of particles with a ring topology
2  Initialize position, velocity and personal best position of the N particles
3  while stop criterion is not satisfied do
4      for k = 1 to N do
5          Update Particle k
6          if Particle k improved its position then
7              Update p_k best position vector
8              p_k failures ← 0
9          else
10             p_k failures ← p_k failures + 1
11         if p_k failures > failures_threshold then
12             for n = 1 to N do
13                 A particle r is chosen by using a roulette wheel based on the rank of
                   the particles
14                 if n = r and p_n is better than p_k then
15                     Connect Particle n to Particle k
16                 else
17                     Disconnect Particle n and Particle k
18         Update Particle k
```

The algorithm begins with a ring topology with N particles. For all PSO iterations, each particle k has $P_k failures$ updated according to the fitness evolution. When the threshold of failures ($failures\_threshold$) is reached, the particle searches

for better particles to follow. The selection of new neighbors is based on a roulette wheel with a fitness-based rank.

The threshold of failures is crucial to the algorithm performance. If it has a low value, the particles will easily try to reconnect. Otherwise, particles will maintain the previous behavior for a long time.

4 Simulation Setup and Results

We used four well-known benchmark functions to evaluate our proposal and compare it to the previous approaches [6]. The functions are used for minimization problems. Two of them are unimodals, Rosenbrock and Ackley, and two are multimodal, Rastrigin and Griewank. The global optimum of all of them is at $(0,...,0)$.

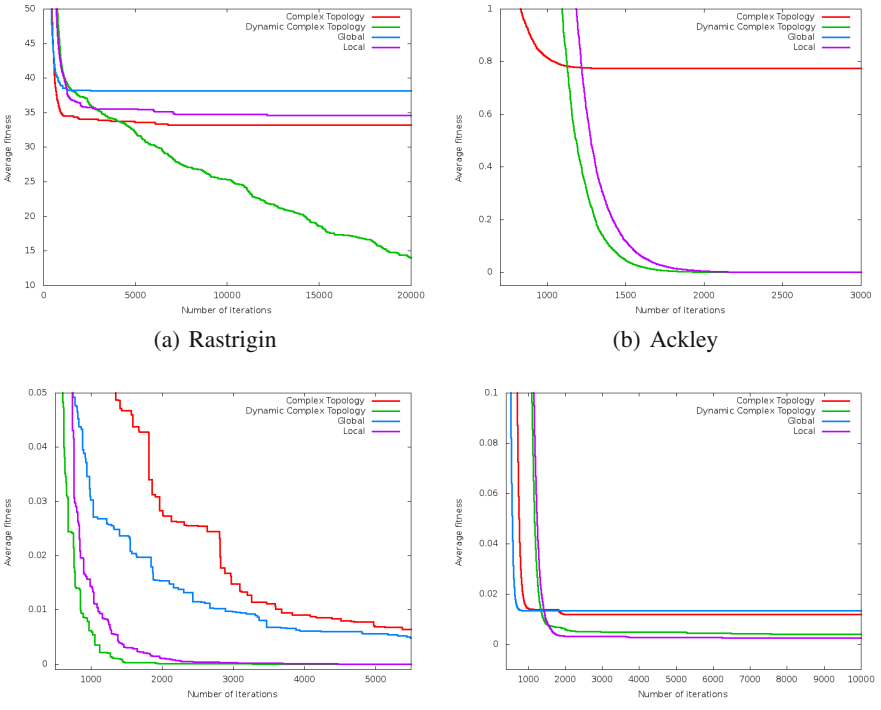
In all experiments, all functions were implemented in 30 dimensions. We have executed the PSO algorithm 30 times with $3,000$ iterations in all functions. The threshold of failures for the particles was set to 100. The particles were updated according to the Equation 4. We used $c_1 = 2.05$ and $c_2 = 2.05$.

Table 1 presents the mean value and the (standard deviation) of the best fitness found for each function by each tested topology. One can observe that the results achieved by our proposal are similar to the Local topology for the functions Ackley, Rosenbrock and Griewank, but we obtained the best performance for the Rastrigin function. One can also notice that we far outperformed the Global topology and the *CNPSO* approach (static complex topology).

Table 1. Mean value and (standard deviation) of the best fitness found for each function

PSO Topology	Rastrigin	Ackley	Rosenbrock	Griewank
Global topology	38.1401	7.4857	0.0011	0.0134
	(9.2908)	(9.3576)	(0.0015)	(0.0189)
Local topology	34.5914	**0.0000**	$\mathbf{6.2587}\times10^{-8}$	**0.0025**
	(9.0085)	**(0.0000)**	$\mathbf{(1.6376}\times10^{-7}\mathbf{)}$	**(0.0052)**
Static complex topology	33.1985	0.7740	0.0017	0.0119
	(8.6007)	(2.2623)	(0.0023)	(0.0148)
Our proposal	**14.0476**	**0.0000**	1.7766×10^{-7}	0.0037
(Dyn. Complex Topology)	**(5.2370)**	**(0.0000)**	(2.7928×10^{-7})	(0.0063)

The average values of the best fitness achieved along the iterations by the PSO algorithm using the four different topologies for the functions Rastrigin, Ackley, Rosenbrock and Griewank are shown in Figure 2. As can be seen, our proposal converges faster for Rosenbrock and Ackley functions. Besides, our approach does not get stuck in local minima in the Rastrigin function, while all other tested approaches quickly stagnate.



(a) Rastrigin (b) Ackley

(c) Rosenbrock (d) Griewank

Fig. 1. The average value of the best fitness achieved in each function through the iterations by the PSO algorithm using different topologies

5 Conclusions and Future Works

In this paper, a novel dynamic communication topology based on the Barabási-Albert model for the Particle Swarm Optimization is proposed. In this approach, the particles explore the search space at the beginning and, as the particles get stagnated, they try to seek for better particles to follow. This search for new neighbors is based on the preferential attachment of the Barabási-Albert model.

The simulation results showed that the proposed approach is in average better than other well known topologies and outperforms a simpler previously proposed topology based on the Barabási-Albert model.

For the future, we intend to test this approach in dynamic problems. We also intend to investigate the impact of the failures threshold of the particles in the optimization process and the impact of the initial topology as well.

References

1. Albert, R., Barabasi, A.L.: Statistical mechanics of complex networks. Reviews of Modern Physics 74, 47 (2002) doi:10.1103/RevModPhys.74.47
2. Barabasi, A.L.: Linked, 1st edn. Perseus Publishing (2002)
3. Barabasi, A.L., Albert, R.: Emergence of scaling in random networks. Science 286, 509–512 (1999)
4. Bastos-Filho, C., Andrade, J., Pita, M., Ramos, A.: Impact of the quality of random numbers generators on the performance of particle swarm optimization. In: IEEE International Conference on Systems, Man and Cybernetics, SMC 2009, pp. 4988–4993 (2009), doi:10.1109/ICSMC.2009.5346366
5. Bianconi, G., Barabasi, A.L.: Competition and multiscaling in evolving networks. EPL (Europhysics Letters) 54(4), 436–442 (2001), http://dx.doi.org/10.1209/epl/i2001-00260-6, doi:10.1209/epl/i2001-00260-6
6. Bratton, D., Kennedy, J.: Defining a standard for particle swarm optimization. In: IEEE Swarm Intelligence Symposium, SIS 2007, pp. 120–127 (2007), doi:10.1109/SIS.2007.368035
7. Callaway, D.S., Newman, M.E.J., Strogatz, S.H., Watts, D.J.: Network robustness and fragility: Percolation on random graphs. Phys. Rev. Lett. 85, 5468–5471 (2000), http://link.aps.org/doi/10.1103/PhysRevLett.85.5468, doi:10.1103/PhysRevLett.85.5468
8. Clerc, M., Kennedy, J.: The particle swarm - explosion, stability, and convergence in a multidimensional complex space. IEEE Transactions on Evolutionary Computation 6(1), 58–73 (2002), doi:10.1109/4235.985692
9. Cohen, R., Erez, K., Ben Avraham, D., Havlin, S.: Resilience of the internet to random breakdowns. Phys. Rev. Lett. 85, 4626–4628 (2000), http://link.aps.org/doi/10.1103/PhysRevLett.85.4626, doi:10.1103/PhysRevLett.85.4626
10. Cohen, R., Erez, K., Ben Avraham, D., Havlin, S.: Breakdown of the internet under intentional attack. Phys. Rev. Lett. 86, 3682–3685 (2001), http://link.aps.org/doi/10.1103/PhysRevLett.86.3682, doi:10.1103/PhysRevLett.86.3682
11. Eberhart, R., Shi, Y.: Comparing inertia weights and constriction factors in particle swarm optimization. In: Proceedings of the 2000 Congress on Evolutionary Computation, vol. 1, pp. 84–88 (2000), doi:10.1109/CEC.2000.870279
12. Ferreira de Carvalho, D., Bastos-Filho, C.J.A.: Clan particle swarm optimization. International Journal of Intelligent Computing and Cybernetics 2(2), 197–227 (2009), http://dx.doi.org/10.1108/17563780910959875, doi:10.1108/17563780910959875
13. Godoy, A., Von Zuben, F.: A complex neighborhood based particle swarm optimization. In: IEEE Congress on Evolutionary Computation, CEC 2009, pp. 720–727 (2009), doi:10.1109/CEC.2009.4983016
14. Heppner, F., Grenander, U.: A stochastic nonlinear model for coordinated bird flocks. In: Krasner, E. (ed.) The Ubiquity of Chaos, pp. 233–238. AAAS Publications (1990)
15. Kennedy, J., Eberhart, R.: Particle swarm optimization, vol. 4, pp. 1942–1948 (1995), http://dx.doi.org/10.1109/ICNN.1995.488968, doi:10.1109/ICNN.1995.488968
16. Kennedy, J., Mendes, R.: Population structure and particle swarm performance, pp. 1671–1676 (2002), doi:10.1109/CEC.2002.1004493

17. Newman, M.: Networks: An Introduction. Oxford University Press, Inc., New York (2010)
18. Shen-Orr, S.S., Milo, R., Mangan, S., Alon, U.: Network motifs in the transcriptional regulation network of Escherichia coli. Nature Genetics 31(1), 64–68 (2002), http://dx.doi.org/10.1038/ng881, doi:10.1038/ng881
19. Watts, D.J.: Small worlds: The dynamics of networks between order and randomness. Princeton University Press, Princeton (1999)
20. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393(6684), 440–442 (1998), http://dx.doi.org/10.1038/30918, doi:10.1038/30918

Weak Ties in Complex Wireless Communication Networks*

Amanda Leonel, Carlos H.C. Ribeiro, and Matthias R. Brust

Abstract. Hundreds of millions of devices—from book-sized notebooks to tiny hand-held mobile phones—are equipped with wireless communication adapters that are able to form a network among themselves. The spontaneous creation of this kind of network and the unpredictable joining and leaving of devices bring forward new challenges on network and topology organization. Network Science has proven to deliver a fruitful methodology to investigate systems such as complex communication networks, and new insights and solutions can be gained by understanding and imitating the function and structure of social networks. Following this line, this paper initially focuses on the development of models that reveal characteristics found to be inherent to social networks. In particular, we consider the finding that social networks can contain a diversity of links: we create clusters of friends, connected by strong links and, additionally, there are links to acquaintances, the so-called weak ties which, despite the name, have been hypothesized as essential for finding jobs or disseminating rumors when strong ties fail. As such links seem to be highly important to deal with the requirements of a complex network such as our own social network, we argue that bringing these structures to the design principles of complex communication networks may result in an increase of efficiency and robustness, and we describe the implementation of two algorithms for wireless communication networks using only local neighborhood information and producing features of complex social networks (weak ties in particular). The results imply that local removing promotes the emergence of weak ties, which we found by using a recently proposed link clustering algorithm for identifying link communities.

Amanda Leonel · Carlos H.C. Ribeiro
Computer Science Division, Technological Institute of Aeronautics,
São José dos Campos, Brazil
e-mail: {aln,carlos}@ita.br

Matthias R. Brust
Department of Electrical Engineering and Computer Science, University of Central Florida,
Orlando, USA
e-mail: mbrust@eecs.ucf.edu

* We are grateful to FAPESP, CNPq and CAPES for supporting the research reported in this paper.

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 49–56.
springerlink.com

1 Introduction

In the last years, it has become clear that the increasing number of wireless communication devices such as notebooks, hand-held mobile phones or even tiny sensors is generating an enormous impact in our daily lifes [6, 13]. The design of wireless networks such as an *ad hoc* or sensor network that consists of a diversity of a large number of devices is a hard task, since the paradigm of self-organization applies: these devices can join and leave the network unpredictably and form networks spontaneously.

These characteristics create new challenges on how to handle the emerging complex communication topologies that potentially can consist of thousands of devices. In order to deal with these challenges, we have to look for networks, which are used to naturally and inherently deal with the problems, and learn design principles by analogy. In fact, understanding the structure of our own social network might help finding answers of how to design a complex communication network and which patterns we have to evoke in a man-made communication network to deal with its own complexity [9].

This work focuses on the findings that our social network consists not of a single type of ties or links, rather it is built on a diversity of links. The human social network is actually a highly complex structure that is tied by different types of interdependency, such as histories, interests, trades, neighborhood, and communications. These ties or links are neither randomly nor uniformly distributed, and the characteristics of the links vary considerably. As a matter of fact, Granovetter [4] reports on the difference between friends and acquaintances, and points out that acquaintances are more useful for certain tasks such as finding a job and disseminating news or rumors. Granovetter calls the links between acquaintances as weak ties. The difference between a weak and a strong tie can be understood in different ways. For a wireless communication network, this can be interpreted by the fact that clusters should concentrate on processing information, while weak ties should dedicate mostly on information dissemination.

In this paper, we focus on the problem of evoking weak ties in ad hoc networks where devices communicate over a wireless medium without using any immediate router. This kind of wireless network belongs to the class of spatial graphs, where the links between nodes depend on the radio transmission range, which is a spatial relation between nodes [5, 8]. The main problem of emerging weak ties is that there is no formal definition available that could be used. Nevertheless, Kumpula *et al* [7] suggest a network model for emerging community-like structures, including strong and weak ties. Additionally and more restrictively, the introduction of new links is explicitly not allowed. This corresponds to the reality of self-organizing wireless networks since links can only be created if nodes are within their respectives transmission ranges. As in Kumpula *et al.* [7], our model also requires 2-hop neighborhood information for execution. On the other hand and in contrast to Kumpula *et al.* [7], our approach considers a localized topology control algorithm that does not rely on network evolution for the creation of weak ties.

Despite the significant limitations regarding link creation, we show that it is possible to build a distinction between strong and weak ties. This is accomplished for example with spatial graphs, but the algorithms work for relational graphs as well. We use (i) the clustering coefficient [12] and (ii) similarities between links [1] to control the topology, and promote the emergence of weak ties in a network. The objective is to create highly clustered regions with low average shortest path by removing superfluous links. To identify weak ties, we use a recently proposed link clustering algorithm for identifying link communities [1]. Link communities focuses on grouping links rather than nodes, and the algorithm incorporates overlap and reveals multiscale complexity in networks.

The remainder of this paper is organized as follows. Section 2 presents the system model. Section 3 describes topology control algorithms. Experiments and analysis of topological properties are in Section 4. Finally, Section 5 concludes this work.

2 System Model for a Wireless Network

We define an ad hoc sensor network consisting of a set of devices connected by wireless network links. We consider here that the initial network topology is a spatial graph such as a unit disk graphs [3]. The resulting wireless network can be represented as $G = (N, L)$ that is a graph with $|N|$ nodes and $|L|$ links. All nodes have the same transmission range r. Two nodes u and v can only form a link when they are in a spatial neighborhood, i.e. when their Euclidean distance d is smaller than the transmission range: $d(u, v) \leq r$. We abstract away the details of the MAC and network layer. Nodes are static, i.e. they keep their initial position and they are deployed uniformly at random in a squared simulation area with an edge length l. Thus, all possible links are already given from the initial configuration.

Furthermore, we assume that every node is aware of its current 2-hop neighbors, listed in a device neighbor list data type. We assume that, in practice, a neighbor discovery service on each device updates the neighbor list at particular time intervals, such that the neighbor list represents—with a minor delay—the current local topology of the network. Geographical positions of the nodes are not considered.

3 Topology Control: Clustering and Weak Ties

Kumpula *et al.* [7] shows a model where a sparse network evolves to a dense network. Since our system model does not allow such a procedure of link addition, we researched for a method that increases the clustering and keeps low the average shortest path by removing links. It turned out that the clustering coefficient indicates clustered nodes, and it can be increased by *removing* links. Our hypothesis is that the links that keep the clustered regions connected should then be weak ties. The clustering coefficient can be used for measuring the efficiency regarding the clustering behavior. Since the clustering coefficient is locally defined we counter the challenge to implement a solution that is localized [10].

The definition of the clustering coefficient might suggest that more links in each node neighborhood result in a higher clustering coefficient. Our approach, however, is based on the observation that this statement does not hold in general. Thus, we found that even the removal of dedicated links can increase the global clustering coefficient. Our algorithm is built on this observation and provides a generic approach. We argue that since the links can only be removed, weak ties have to appear naturally in the network topology.

Unfortunately, there is no formal definition in the literature that could be used to identify weak ties. Granovetter [4] produces an informal idea of the impact of weak ties on the network structure. We propose to use a link communities algorithm recently reported in the literature [1] to identify weak ties. Our final analysis consists of three steps:

- calculate the similarities between pairs of links (i.e. Jaccard index),
- cluster the ties, using a single-linkage hierarchical clustering [1], and then
- classify link communities as strong or weak ties.

In link communities, the Jaccard index can be used to calculate the similarity S between links from an undirected and unweighted network [1]. Link communities use single-linkage hierarchical clustering to find hierarchical community structures due to simplicity and efficiency, even on large-scale networks. Initially each link builds one community. The pairs of ties with higher similarity and common ties between them are grouped simultaneously. The algorithm ends when all links are clustered.

As the similarity S measures the strength of the merged community, we consider that weak ties appear, in the link cluster, as single communities, i.e., with low or no similarity to other link communities. Thus, the set of weak links in a network is represented by the union of these unitary link communities.

3.1 A Link Removal Algorithm Based on Clustering Coefficient (R_{cc})

The first proposed algorithm verifies if a link $e_{u,v}$ is inefficient in terms of the clustering coefficient, *i.e.* if its removal increases the clustering coefficient. In the case of inefficiency, the link $e_{u,v}$ is considered as a candidate for removal. It is not removed immediately because removal in this stage would be in accordance with the criterion of that particular node only. However, since removing a link affects the local clustering coefficients of the 2-hop neighborhood of the set u, v, an additional removal confirmation phase must be performed, when nodes exchange the removal candidate information with their corresponding neighbors. Connectivity is guaranteed by the fact that removing $e_{u,v}$ requires at least one neighbor of u to be connected to one neighbor of v. Thus, the resulting topology is connected, and the algorithm is therefore connectivity-preserving.

The algorithm requires 2-hop synchronization to remove a link, since 2-hop topological information is required in order to plan the action. Then again, this local link removal affects the 2-hop neighbors. For reasons of simplicity the algorithm has

been implemented in a synchronous network, and a desynchronization procedure is not detailed here. We notice, however, that any synchronous algorithm (*i.e.* an algorithm for synchronous networks) can be transformed in its asynchronous counterpart by using synchronizers [2].

3.2 A Link Removal Algorithm Based on Link Similarity (R_{simil})

The second proposed algorithm verifies if a link $e_{u,v}$ is inefficient in terms of similarity, *i.e.* if the link has a very high average similarity between its neighboring links. In this case, we propose that $e_{u,v}$ can be considered a strong or redundant link because its removal should not considerably affect the average shortest path. Otherwise, if the link $e_{u,v}$ has a low average similarity between its neighboring links, $e_{u,v}$ can be a weak tie, since removing this link may significantly increase the average shortest path.

An approach for the removing decision is based on a variable probability p that is proportional to the average similarity between the link $e_{u,v}$ and its neighboring links. This means that links with high average similarity are strong removal candidate with probability p. On the other hand, links with low average similarity, or weak ties, may be preserved in the network due to its low probability of removal.

As in the previous algorithm, the connectivity is guaranteed and the resulting topology is connected. In both algorithms, the stop condition is given by the choice of a percentage of links removed from the initial spatial network.

4 Simulation Study

The first experiments were run on a set of 200 nodes uniformly deployed at random in a square with edge length $l = 450$ units and transmission range $r = 60$ units. The initial topology was created using the unit disk graph model described in Section 3. An example is shown in Fig. 1(a). Table 1 shows statistics from 50 spatial networks. Fig. 1(b) shows an example of a resulting network using the link removal algorithm based on clustering coefficient. Each color represents different link communities. Fig. 1(c) shows an example of a resulting network using the link removal algorithm based on link similarities. Each dotted line represents an unitary link communities.

Fig. 2(a) reveals that the clustering coefficient increases approximately 30% using the algorithm based on clustering coefficient after removing 25% of links. Our experiments have indicated that this algorithm reaches its stop condition in 25% removed links. However, with the same percentage of removal, the clustering coefficient decreases approximately 30% using the algorithm based on link similarities. The working principle of the first model is the optimization of the clustering coefficient by selectively removing links. We observe that link removals are more likely to occur in sparse regions, while highly clustered regions are mostly unaffected (see Fig. 1(b)).

The average path length is showed in Fig. 2(b). The link removal algorithm based on link similarities increases the APL in only 20% after removing half of links in the

Table 1. Statistics from 50 spatial networks with $n = 200$ nodes, $l = 450$ u, $r = 60$

Metric	Average	Standard Deviation
$\|L\|$ initial links	979.7037	43.0071
Clustering Coefficient	0.6309	0.0152
Average Shortest Path	5.6571	0.2432

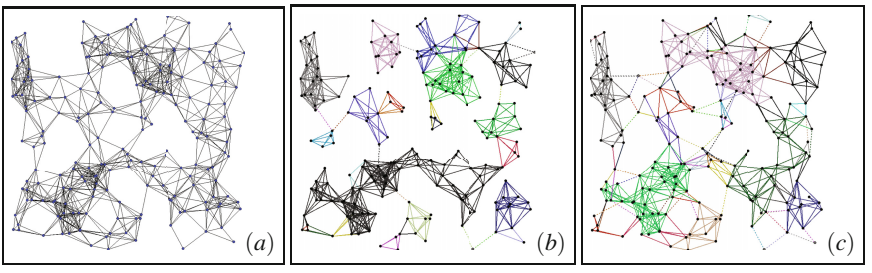


Fig. 1. Examples: (a) Spatial network with $n = 200$ nodes, $l = 450$ u, $r = 60$ u, $|L| = 1,226$ links; (b) Network with 20% nodes removed by the link removal algorithm based on clustering coefficient R_{cc}; (c) Network with 30% nodes removed by the link removal algorithm based on link similarities R_{simil}

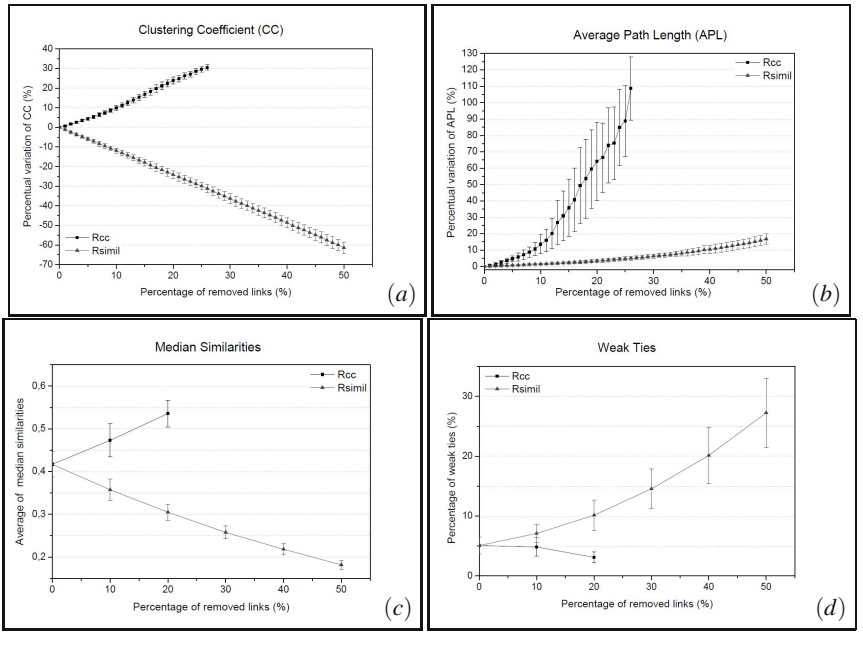


Fig. 2. Results from 50 final networks using link removal algorithms based on clustering coefficient (R_{cc}) and link similarities (R_{simil})

network. This means that R_{simil} keeps weak ties in the network. However, the link removal algorithm based on clustering coefficient increases the APL up to 125% after removing only 25% of links in the network. In this case, R_{cc} removes mainly weak ties, that significantly increase the average shortest path.

The removal of strong *vs.* weak ties is clearer in Fig. 2(c). R_{cc} significantly increases the median similarity of networks, since removing weak ties. However, R_{simil} decreases the median similarity of networks after removing strong ties.

These algorithms aim at increasing the clustering coefficient and at keeping low the average shortest path, but as a side effect, isolated links that connect clustered regions may appear. These isolated links seem to have the same functions and structures as weak ties have in social networks. If the resulting network topology is powerful in terms of information dissemination, action taking (information processing etc.) as a complex social network, but using less resources than the initial network (because links have been removed), then the presented algorithm can be used to release efficiency reserves of complex communication network design.

After calculating similarities between pairs of links and clustering links, single communities – *i.e.*, unitary communities with low similarity to other sets of links – were classified as weak ties. See Fig. 2(d) for results. The algorithm R_{simil} transforms about 30% of network connections in weak ties, after removing half of links in the network. However, the link removal algorithm based on clustering coefficient decreases the number of weak ties. As a matter of fact, the resulting topologies for the experiments based on similarities reveal visible dotted links where weak ties start to dominate (see Fig. 1(c)).

5 Conclusions

The link removal algorithm based on clustering coefficient and introduced in this paper shows that clustering does not lead to the emergence of weak ties. On the other hand, the control based on link similarities efficiently creates weak ties, but significantly decreases the clustering coefficient.

Weak ties appear to be important for the transfer of certain information that is filtered by a clustered set of nodes. And these links can be successfully classified by the link communities algorithm. Importantly, our approaches do not allow addition of new links, so our approaches rely on a procedure by removing dedicated links.

Whereby our work focuses on manipulation of an existing network to form weak ties, human social networks seem to apply different principles: they are driven by the joining and leaving of nodes and thus, use network evolution as the driving force for emerging patterns. For example, the likelihood for two persons with a common friend to become friends is higher than the possibility for two persons with no common friend to become friends [11].

In spite of these interesting considerations, it is important to keep in mind that the results were obtained for spatial networks such as unit disk graphs. We expect similar results for relational graphs, studies and analysis on network composed of dynamic nodes, and combination between link removal algorithms, but these are

subjects of further investigations. In an extended version, weights can be assigned to links to conduct the removing process.

Finally, it remains an open question if weak ties can be produced with a microscopic or localized model that does not make use of more than one-hop neighborhood information.

References

1. Ahn, Yong-Yeol Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature 466(7307), 761–764 (2010)
2. Awerbuch, B.: Complexity of Network Synchronization. Journal of the ACM (JACM) 32(4), 804–823 (1985)
3. Clark, B.N., Colbourn, C.J., Johnson, D.S.: Unit disk graphs. Discrete Mathematics 86(1-3), 165–177 (1991)
4. Granovetter, M.S.: The Strength of Weak Ties. American Journal of Sociology 78(6), 1360 (1973)
5. Helmy, A.: Small worlds in wireless networks. IEEE Communications Letters 7(10), 490–492 (2003)
6. Kahn, J.M., Katz, R.H., Pister, K.S.J.: Next century challenges: Mobile Networking for ”Small Dust”. In: Proceedings of the 5th Annual ACM/IEEE International Conference on Mobile Computing and Networking MobiCom 1999, pp. 271–278. ACM Press, New York (1999)
7. Kumpula, J., Onnela, J.-P., Saramäki, J., Kertész, J., Kaski, K.: Model of Community Emergence in weighted Social Networks. Computer Physics Communications 180(4), 517–522 (2009)
8. Santi, P.: Topology Control in Wireless Ad Hoc and Sensor Networks. Wiley (2005)
9. Vega-Redondo, F., Chesher, A., Jackson, M.: Complex Social Networks, 1st edn. Cambridge University Press (2007)
10. Wattenhofer, R.: Sensor Networks: Distributed Algorithms Reloaded - Or Revolutions? In: 13th Colloquium on Structural Information and Communication Complexity, SIROCCO (2006)
11. Watts, D.J.: Small Worlds - The Dynamics of Networks between Order and Randomness. Princeton University Press (1999)
12. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-world networks. Nature 393(6684), 440–442 (1998)
13. Weiser, M.: The computer for the 21 st century. ACM SIGMOBILE Mobile Computing and Communications Review 3(3), 3–11 (1999)

Vulnerability-Aware Architecture for a Tactical, Mobile Cloud

Anne-Laure Jousselme, Kevin Huggins, Nicolas Léchevin, Patrick Maupin, and Dominic Larkin

Abstract. Currently light infantry soldiers do not have access to many of their cyber resources the moment they depart the forward operating base (FOB). Commanders with recent combat experience have reported on the dearth of computing abilities once a mission is underway [14]. To address this, our group seeks to develop a tactical, mobile cloud implemented on a swarm of semi-autonomous robots. We provide two contributions with this work. First, provide a formal definition of the problem followed by a description of our approach to vulnerable state identification based on pattern recognition techniques. Second, we present an awareness definition as it pertains to our domain.

1 Problem Statement and Formalization

This is in essence a coverage problem. A robot is responsible for providing communication coverage to the set of clients in its area. Additionally, each robot must maintain communications with at least one other robot to ensure that the global network remains connected, see Figure 1.

First, we consider the elements of our domain. Let $R=\{r_1, \ldots, r_N\}$ be the set of robots and $C=\{c_1, \ldots, c_M\}$ the set of clients. The set C combined with their spatial location is a *configuration*. We denote ρ as a robot's unique communications range, which could be adjusted based on environmental demands. Next, $E=\{e_1, \ldots, e_N\}$ is the set of communication links between the robots and $G_c=(R, E)$ is the corresponding communication graph. Let N_c, be the set of node coordinates. Two robots r_i and r_j are separated by a distance of d_{ij} and are connected if their distance is

Anne-Laure Jousselme · Nicolas Léchevin · Patrick Maupin
Defence R&D Canada–Valcartier, 2459, Pie XI North, Quebec, G3J 1X5, Canada
e-mail: {Anne-Laure.Jousselme,Patrick.Maupin,Nicolas.Lechevin}@drdc-rddc.gc.ca

Kevin Huggins · Dominic Larkin
US Military Academy, West Point, NY 10996, USA
e-mail: {Kevin.Huggins,Dominic.Larkin}@usma.edu

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 57–65.
springerlink.com

less than ρ[1]. In later sections of this work, d_{ij} will denote the distance between two robots, two clients, or one robot and one client. We assume that (*i*) indirect links are possible through intermediate nodes acting as relays, and (*ii*) at least one of the nodes is connected to an external communication node such as a satellite or UAV. In other words, we assume that there exists a communication resource capability within the network to ensure that clients' messages are handled properly through the mobile cloud via an external wide range communication relay.

The environment is represented by a navigation graph, where each node represents a possible position for the robots or clients and each edge between nodes is a possible path [15].

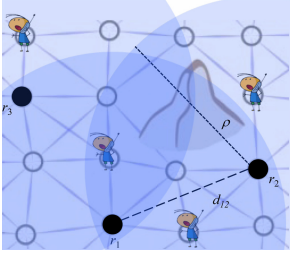


Fig. 1. A tactical mobile cloud for communication coverage. Black circles are robot nodes r_i and light blue circle represent their coverage. ρ is the communication range and d_{ij} is the distance between robot r_i and r_j.

Our formal objective is to provide continuous communications coverage for all clients while simultaneously maintaining sufficient connectivity within the network of robots. Accordingly, our research hypothesis is the following: an early identification of network vulnerabilities will prevent catastrophic events.

1.1 Coverage and Connectivity

Central to our work is a precise definition of *coverage*. Clients need to have coverage in the mobile communications architecture in order to access the tactical cloud. Similarly, robots need to provide global coverage to ensure all clients are fully connected to the cloud. For simplicity, we provide a binary definition of coverage: covered or not covered. However, this definition can be easily extended to other models, such as probabilistic ones. We now consider a robot r_i with a communications range ρ and a client c_j separated by a distance d_{ij}.

We define coverage provided by robot r_i to client c_j as

$$\begin{aligned} cov(r_i, c_j) &= 1 \quad \text{if } d_{ij} < \rho \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{1}$$

The global coverage for the network of robots at a given client c_j is

$$cov(R, c_j) = \sum_{i=1}^{N} cov(r_i, c_j) \tag{2}$$

[1] For simplicity, obstacles that modify the communication links will not be considered.

which is the number of robots that cover client c_j. In (3) we describe the inverse, *i.e.*, a definition of set of clients covered by robot r_i.

$$cov(r_i, C) = \sum_{j=1}^{M} cov(r_i, c_j) \tag{3}$$

Given the coverage definitions from the perspectives of both the client and the robot, equation (4) describes the *global coverage* of the network relative to the set of clients.

$$cov(R, C) = \sum_{i=1}^{N} cov(r_i, C) = \sum_{j=1}^{M} cov(R, c_j) = \sum_{i=1}^{N} \sum_{j=1}^{M} cov(r_i, c_j) \tag{4}$$

The other notion crucial to our work is that of network connectivity. We say that two robots are connected if they are in their respective range of communication. We define then:

$$\begin{aligned} con(r_i, r_j) &= 1 \quad \text{if } d_{ij} < \rho \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{5}$$

A value of 1 means then that a link exists between the two robots.

We say then that the global network connectivity holds if for each pair of robots (r_i,r_j) there exists a path linking r_i to r_j:

$$con(R) = 1 \text{ if } \forall (r_i, r_j) \in R^2, \exists (r_1, ..., r_m) \in R^m; (r_i, r_1), (r_1, r_2) ..., (r_m, r_j) \in E \tag{6}$$

Alternative definitions could easily replace these binary definitions and define different connectedness indices such as the algebraic connectivity [6]. Indeed, allowing real values for the connectivity would increase our flexibility in the definition of vulnerable states and lead to different cost functions (see Section 3.3). This kind of approach would lead to a multi-class problem rather a binary one and will be considered in future work.

1.2 Vulnerability

The vulnerability V of a system S can be understood as a mapping, V_S: $T \mapsto \mathcal{C}$, between an initiating threat T, whether intended or not, and a resulting consequence $\mathcal{C}$ characterized by a degree of loss [16] and related to system inoperability or state unreachability. Depending on how the threat uncertainty is characterized, the cost function may be aggregated, giving rise to an expected cost, or equivalently to a risk function [1]. Vulnerability thus corresponds to the susceptibility of a system or to the manifestation of the inherent state of a system, which can be severely affected when threatened [7]. Following the classification proposed by Klibi et al. [10], uncertain initiating events such as threats can be classified as either random, hazardous, or deeply uncertain events. Depending on the event model adopted,

various approaches can be used to deal with uncertainties arising in a decision-making problem. For instance, the approach advocated by Brown et al. [2], which is based on worst-case bi-level or simplified tri-level programming, is to be contrasted with risk analysis involving probability and event trees [5], [1]. Indeed, the fault trees used to locate the single point of failure or the minimal cut set[2] that maximizes the probability of disruption relies on the probabilistic modeling of events such as random failures [22].

In this paper, a mobile network *vulnerable* state can be defined as an instance of the network's state that may evolve in time until it affects the network's functions and the completion of its goals. Endogenous and exogenous threats to the network include the robots' inability to precede as intended, possibly due to hardware-software failures or malevolent acts, electronic warfare, obstacles, or unexpected client moves that cause some robots to move beyond their neighbors' communication range. Consider a sample set of possible clients configurations C_0 and a corresponding robot deployment represented by graph $G_0=(R, E_0)$. Include also the set $N_{c,0}$ of nodes' coordinates at time instant t_0 (encoded as attributes of the nodes). Various experiments can be conducted by triggering the loss of a robot or a subset of robots or by repositioning clients. The occurrence at t_1 of this triggering event may give rise to an adaptive robot deployment, whereby communication links can be either permanently lost or re-established, depending on the relative distance to neighboring robots. This hybrid dynamical system is characterized by switching time instants $\{t_1, t_2, \ldots, t_m\}$, where $t_{i+1}>t_i$. At t_i, the edge set jumps from E_i to E_{i+1}. An edge (i,j) is lost whenever the distance between robots r_i and r_j is greater than the communication range. It is assumed that the node set R remains invariant whether or not a robot is able to operate. The final time instant t_m is defined by the absence of any future triggering events such as a robot failure or a client move.

As further explained in Section 2.3, a component of the network (i.e., edge, node, or sub-network) is classified as *vulnerable* when a graph-connectedness-related cost associated to this component is above a prescribed threshold.

1.3 Vulnerability Awareness

The notion of awareness considered in this work is derived from the concept of limited system resources [9],[8]. Intuitively, awareness is an epistemic state, close to knowledge, referring to a limited view and a limited capacity of the agents to reach a perfect state of knowledge, the one that would be reached by perfect logically omniscient reasoners. When defining situational awareness, one must consider the concepts of attention, vigilance, intelligence and stress within the context of resource-bounded agents. Therefore, we adopt the following definition of awareness: "an agent is aware of a proposition y if it can compute the truth value of y before time t". The vulnerability awareness of one robot r_i [3] is thus directly

[2] A set of edges of a graph which, if removed (or "cut"), disconnects the graph (*Wolfram, Mathworld*, http://mathworld.wolfram.com/CutSet.html).

[3] This is a local definition but a global definition would concern a central instance having access to the global swarm's state.

linked to its ability to come with an answer the question y ="*Am I vulnerable*?" by means of an algorithm (to be detailed in the upcoming section) given its limited resources (memory, computation, move, etc). Particularly challenging is the feature selection process as (1) the more features, the higher the computation and memory costs; and (2) some feature may require a complete map of the swarm involving higher memory needs while other may be evaluated locally.

2 A Pattern Recognition Approach to Vulnerability Assessment

2.1 Principle

In [13], the authors proposed a toolbox with the goal of detecting and predicting the vulnerabilities in complex networks. These principles rely on pattern recognition techniques that leverage structural, dynamical, and functional features selected to sensitize the classifier to potential vulnerabilities in abnormal situations. Such an approach is expected to yield fast vulnerability prediction when compared with a simulation using a first-principle-based model of the network. The problem of complex systems vulnerability assessment has already been interpreted as a classificatory problem, which includes such applications as disease surveillance systems [20] and the crisis recognition [11]. Our approach integrates pattern recognition techniques applied to a time series and a network's structural and dynamical properties.

To determine vulnerability, we reason over the network using pattern recognition. With it, we design by training a mapping ψ such that:

$$\begin{aligned} \psi : \mathcal{G} &\rightarrow \{y; \bar{y}\} \\ x &\mapsto \psi(x) = \hat{y} \end{aligned} \tag{7}$$

where x is a representation of an element of $\mathcal{G}$ (e.g. a node, link or sub-graph) and $\hat{y}$ is an estimate of the detrimental effect of that element on the network, either 1 if vulnerable or 0 otherwise. Note that we use the term "vulnerable" for qualifying a node although this is extended to the network. Typically, x is a vector of k network features identified as relevant by feature selection pre-processing. As mentioned in Section 1.3, one of the crucial tasks consists in identifying the set of candidate features for the problem.

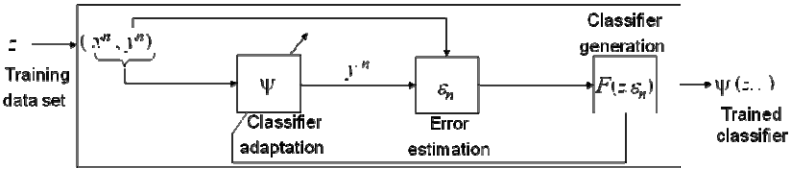


Fig. 2. Training of the network vulnerability classifier. F aims to yield a classifier ψ that minimizes the error estimation given the training data set $z=(x^n,y^n)$.

The training data set $z^n=(x^n,y^n)$ consists of n instances of R, that is a set of n samples x^n of k features together with a label y (see Table 1). z^n feeds a classifier generation mechanism F, as shown in Fig. 2, which seeks a classifier that minimizes the error estimate between the set of estimated class labels and their corresponding ground truth.

2.2 Features

In [1], the authors proposed four network feature categories. The first pertains to the structural properties of a labeled, weighted graph. These include centrality, similarity connectivity, shortest path metrics, clustering coefficient, spectral properties, vertex coreness, graph density, average nearest neighbor degree, among others [3]. The second category considers flow dynamic changes by exploiting information on signals and systems such as Fourier transform, spectral monitoring [12], bifurcation analysis [4] and efficiency measures [17]. Indicators pertaining to complex system science and statistical physics described the third category of features with examples including the exponent of the power-law distribution of failure occurrences at a crossover [18], the local shape factor of a sand pile adapted to networks [19], the Cavahlo-Rodrigues entropy, the spatial entropy, the fractal dimension, the symmetropy, the Hurst coefficient, and the self-similarity parameter [21]. The final feature group concerns functional information on key components of the network. This fourth category includes the notions of coverage introduced in Section 2.2. Our example features were drawn from this latter group as well as the structural features from the first category.

2.3 Training

The purpose of feature extraction is to build a representation that is particularly suited to vulnerability recognition problems specific to networks. Indeed, features are naturally geared to the modeling of classificatory problems. Once this model is derived, fast and efficient recognition is expected when compared to physics-based models of large interconnected networks. A parallel between pattern-recognition-based and game-theoretic approaches is proposed in [13].

Given an initial clients configuration C_0 and a corresponding robot deployment G_0, (including the robots' coordinates $N_{c,0}$), the training data set z used to derive the classifier in (7), is obtained from the disturbance sample set $\mathscr{D}=\{D_1,\ldots,D_n\}$. D_i stands for a set of sequence of disturbances $\{\delta_{ij1}, \delta_{ij2},\ldots,\delta_{ijm}\}$. This sequence is in a one-to-one correspondence with the occurrence time set $\{ t_1 ,t_2, \ldots, t_m\}$, $t_{i+1}>t_i$. δ_{ijk} represents a random realization of the j^{th} disturbance of the sequence applied to the network of robots at time t_k and for experiment i. The first disturbance δ_{ij0} corresponds to a robot failure and is followed by a series of new client configuration C_j, $\{\delta_{ij1},\ldots,\delta_{ijm}\}$. It assumed that the first triggering event (first disturbance) δ_{ij1} is a robot failure occurring at t_1 and that all δ_{ij1} span R, the set of robots. The following disturbances are a series of new client configuration C_j, $\{\delta_{ij1},\ldots,\delta_{ijm}\}$

representing new client configurations arising from clients adjusting their positions. In order to evaluate each robot's vulnerability, they will be removed successively. D_i is then defined as the following union $\cup_{j\in R}\{\delta_{ij1}, \delta_{ij2},\ldots, \delta_{ijm}\}$. D_i is instrumental in defining experiment i. Indeed, sequence D_i generates the sequence of edge sets $\{\{E_{11},\ldots, E_{1m}\},\ldots,\{E_{|R|1},\ldots, E_{|R|m}\}\}$, where $|R|$ denotes the total number of robots. The state of G_m is used to determine the cost $\mathscr{C}_{ij},(con_i,cov_i)$ associated with a disturbance sequence j of experiment i. This cost depends on the connectedness of the graph through a disconnectedness index, $con_i(G_m)$, derived from equation (6), and the coverage $cov_i(G_m,,C_m)$ of the set of clients by the robot network at time t_m. In the binary case, when G_m is disconnected, con_i is equal to 1 and when G_m is connected, con_i is set to zero. Sequence j of experiment i is thus associated with the following mapping

$$\{d_{ij1},d_{ij2},\ldots,d_{ijm}\} \rightarrow \mathscr{C}_{ij} \rightarrow y_{ij}$$

Each experiment i is characterized by the set of labels $\{y_{i1},\ldots, y_{i|R|}\}$. The classifier used to analyze the vulnerability of the robot network's dynamics in response to the occurrence of possible contingencies is obtained by exploiting the data of Table 1 established for each experiment $i\in\{1,\ldots,N\}$.

Table 1. Table of features excerpt from Experiment i. Features are divided into two classes depending on whether each robot is able or not to compute the features from local information, which is sent by adjacent nodes.

		Features						Label
		Local			Global			
	⋯	⋯			⋯			⋯
Experiment i	Robot 1	f_{11}	f_{12}	⋯	⋯	f_{1p-1}	f_{1p}	y_{i1}
	Robot 2	f_{21}	f_{22}	⋯	⋯	f_{2p-1}	f_{2p}	y_{i2}
	⋮	⋮				⋮		⋮
	Robot $\|R\|$	$f_{\|R\|1}$	$f_{\|R\|2}$	⋯	⋯	$f_{\|R\|p-1}$	$f_{\|R\|p}$	$y_{i\|R\|}$
	⋯	⋯			⋯			⋯

3 Conclusion

In this paper, we have described a formal definition of a vulnerability-aware mobile, tactical cloud architecture designed to support dismounted soldiers. We developed a problem formalization that described the network as well as a definition of coverage and vulnerability. In addition, we provided a formal definition of awareness as it pertains to our domain. We applied a novel pattern-recognition approach to vulnerability assessment, which enables robot nodes in the network to efficiently detect when they are in a vulnerable state. Future work consists of implementing and benchmarking the architecture as well as exploring multi-class problems for more robust connectivity definitions.

References

[1] Al Mannay, W.I., Lewis, T.G.: Minimizing network risk with application to critical structure protection. Journal of Information Warfare 6(2), 52–68 (2007)

[2] Brown, G.G., Carlyle, W.M., Salmeron, J., Wood, K.: Analyzing the vulnerability of critical infrastructure to attack and planning defenses. Tutorial in Operations Research, Informs, 102–123 (2005)

[3] Csárdi, G., Nepusz, T.: The igraph software package for complex network research. InterJournal Complex Systems 1695 (2006)

[4] Dobson, I.: Distance to Bifurcation in Multidimensional Parameter Space: Margin Sensitivity and Closest Bifurcations. In: Chen, D.J., Hill, X., Yu, X. (eds.) Bifurcation Control. LNCS, vol. 293, pp. 49–66. Springer, Heidelberg (2003)

[5] Garcia, M.L.: Vulnerability assessment of physical protection systems. Butterworth-Heinemann (2006)

[6] Godsile, C., Royle, G.: Algebraic Graph Theory. Springer, New York (2001)

[7] Haimes, Y.Y.: On the definition of vulnerabilities in measuring risk to infrastructure. Risk Analysis 26(2), 293–296 (2006)

[8] Halpern, J., Moses, Y., Vardi, M.Y.: Algorithmic knowledge. In: Proc. of the 5th Conference on Theoretical Aspects of Reasoning about Knowledge (TARK 1994), pp. 255–266. Morgan Kaufmann (1994)

[9] Jousselme, A.-L., Maupin, P., Garion, G., Cholvy, L., Saurel, C.: Situation awareness and ability in coalitions. In: 10th International Conference on Information Fusion, Quebec city, Canada, July 9-12 (2007)

[10] Klibi, W., Martel, A., Guitouni, A.: The design of robust value-creating supply-chain network: a review. European Journal of Operational Research 203(2), 283–293 (2010)

[11] Larsen, H.L., Yager, R.R.: A Framework for Fuzzy Recognition Technology. IEEE Trans. on Systems, Man, and Cybernetics - Part C: Applications and Reviews 30(1), 65–76 (2000)

[12] Léchevin, N., Rabbath, C.A., Maupin, P.: Toward a stability monitoring system of an asset-communications network exposed to malicious attacks. In: American Control Conf., San Francisco (2011)

[13] Léchevin, N., Jousselme, A.-L., Maupin, P.: Pattern Recognition Framework for the Prediction of Network Vulnerabilities. In: IEEE Network Science Workshop, West Point, NY (June 2011)

[14] Levine, C.: Analysers, and Users of Situational Information. In: Workshop on Information Sharing at the Front Line, Indian Wells, CA (April 2010)

[15] Maupin, P., Jousselme, A.-L., Wehn, H., Mitrovic-Minic, S., Happe, J.: A Situation Analysis Toolbox: Application to Coastal and Offshore Surveillance. In: Int. Conf. on Information Fusion, Edinburgh, UK (2010)

[16] McGill, W.L., Ayyub, B.M.: The meaning of vulnerability in the context of critical infrastructure protection. In: Critical Infrastructure Protection: Element of Risk, Critical Infrastructure Protection Program, George Mason University School of Law (2007)

[17] Nagurnay, A., Quiang, Q.: A network efficiency measure with application to critical infrastructure networks. Journal of Global Optimization 40, 261–275 (2008)

[18] Pradhan, Hansen, A., Hemmer, P.C.: Crossover behavior in burst avalanches: signature of imminent failure. Physical Review Letters 95(12), 125501-1(4) (2005)

[19] Ramos, O., Altshuler, E., Maloy, K.J.: Avalanche prediction in a self-organized pile of beads. Physical Review Letter 102(7), 078701(1-4) (2009)
[20] Shmueli, G., Fienberg, S.E.: Current and potential statistical methods for monitoring multiple data streams for bio-surveillance. In: Statistical Methods in Counter-Terrorism, pp. 109–140. Springer (2006)
[21] Sprague, K.B., Dobias, P.: Behavior in Simulated Combat – Adaptation and Response to Complex Systems Factors, DRDC CORA TM 2008-044 (2008)
[22] Stamatelatos, M., Vesely, W., Dugan, J., Fragola, J., Minarick, J., Railsback, J.: Fault tree handbook with aerospace applications. In: NASA Office of Safety and Mission Assurance, Washington, DC (2002)

Migration, Communication and Social Networks – An Agent-Based Social Simulation

Hugo S. Barbosa Filho, Fernando B. Lima Neto, and Wilson Fusco

Abstract. Due to the high dynamics present on several social phenomena, it is extremely difficult to carry out scientific investigations on social sciences. This is true especially for those phenomena most relevant for social sciences (e.g., human migration) which significantly increase the difficulty to perform an objective scientific investigation. To overcome such constraints, social scientists have been using modeling and simulation as a new approach to carry out experimental investigations on different social phenomena. In this work a multi-evolutionary agent model (MEAM) devised for social simulations was used in an experimental investigation about a plausible correlation involving migration, communication and social networks. Results suggest that the proposed model was able to outcome macroscopic behaviors adherent to actual social phenomena.

1 Introduction

Several studies pointed out that social networks play a significant role in migrants' lives. However, such aspect is treated as a complementary component for migratory phenomena, less important than other concepts (such as job market or difference in wages) [4]. Moreover, how social networks do affect migratory behaviors is a question that remains open.

Most social phenomena with scientific interest are complex in nature and have emergent behaviors of some higher orders (i.e., when individuals detect the presence of emergent features and act accordingly [5]).

Hugo S. Barbosa Filho · Fernando B. Lima Neto
Polytechnic School of Pernambuco, University of Pernambuco
e-mail: {hsbf,fbln}@ecomp.poli.br

Wilson Fusco
Joaquim Nabuco Foundation
e-mail: wilson.fusco@fundaj.gov.br

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 67–74.
springerlink.com

In a previous work, the authors investigated how the widespread of communication technologies affected migratory flows [1]. In this work, we advance that investigation, including also elements such as *short distance communication* and information exchange within social networks. Our objective is to prove that the MEAM is also able to reproduce migratory phenomena even in a more complex scenario, helping social scientist to gain a deeper understanding regarding migratory flows and social networks.

2 Background

2.1 Computer Simulation of Social Phenomena

Agent-based computational modeling has become an increasingly important tool in conducting experiments in social sciences. Phenomena such as acculturation, migration, spread of disease, groups formation, wars, among others, have been modeled successfully using this approach [6].

Only with the increased processing power of computers in recent decades and with advances in software tools and programming languages that Agent-Based Social Simulations could gain greater scalability, complexity and plausibility. Theoretical and technical advances in artificial intelligence should be listed also as an important aspect in this true paradigm shift that is observed today in social simulations.

Although social simulations are a relatively new area of research, social scientists already have important tools at their disposal to carry out their investigations. These tools allow the social scientist to represent phenomena and social behavior through computational models and investigate the dynamics of these models over time (e.g., PAX, Netlogo, Repast etc.).

2.2 Concepts on Demography and Human Migration

In demography, scientific studies are mostly related to human population and its dynamics. It encompasses features such as structures, sizes, distributions, behaviors or phenomena which can change those aspects over space or time[8].

Ravenstein, in 1885, published the first work proposing a well empirically grounded description of general aspects regarding human migration [9]. In that work, Ravenstein stated that international migration might be described according to 11 laws, which were latter named as “Ravenstein’s Laws of Migration”. Given that in this work we are not focusing only on international migration but on general aspects of the phenomena, four of those most relevant to this work are listed below:

- every migration flow generates a return or countermigration;.
- the majority of migrants move a short distance.
- migrants who move longer distances tend to choose big-city destinations.
- urban residents are often less migratory than inhabitants of rural areas.

After Revenstein, several quantitative models of migration flows and the variables that affect those flows were proposed. More on classical models of migration can be seen in [11].

In last decades other social factors are also being investigated as related to migration flows such as social networks [3, 2]. However, scientific researches carried out in order to establish the role played by social networks on migration flows are mostly based on surveys, census and official immigration data, which has problems and limitations [7, 10].

2.3 MEAM - A Multievolutionary Agent Model for Social Simulations

The agent model applied here first appeared in a previous work by the authors [1] and it was conceived as an alternative to other agent models found in literature. In the model, a multilayered architecture makes an agent able to perceive and interpret external stimuli from different perspectives. Each evolutionary layer can represent one trait of the agent's perception and of the environment comprehension. In other words, multievolutionary agents are able to observe, analyze and react to stimuli even when available actions lead the agent to conflicting situations.

Each layer has its own evolutionary process that is guided by the behavior of the agent according to the actual problem to be studied. The phenotypic and social evolutionary processes will be influenced by outcomes resulting from agent's actions throughout its existence while the genetic evolution will occur from the crossings between the players over the generations.

The cognitive module determines the actions to be performed by the agent, under the influence of phenotypic module (responsible for determining the physical and behavioral expression of the agent) and their perceptions about the environment and itself.

The phenotype evolution will occur as the outcome produced by the following factors:

- Observation and interpretation of agents' surroundings, more precisely, all entities that that may be there (e.g.,other agents and objects).
- Evaluation of the overall performance based upon an objective function (depending on the phenomena to be modeled)
- Impact of previous actions on the agents performance

The architecture of the cognitive apparatus can be seen in the form of three different modules, each one with a specific role to be played during a cognitive activity (e.g., learning or judging).

More on the inner details, processes and dynamics regarding the Multievolutionary Agent Model can be seen in the paper where it was proposed [1].

3 Experiments and Results

In this section we describe in details how the proposed agent model was applied to a social phenomenon modeling task. To test whether our model is able to produce plausible outcomes (from social sciences point of view) we carried out several experiments related to internal human migration.

The artificial society created for this experiment may be described as follows:

- two regions in a country are distant from each other enough to not allow daily flows;
- each region has its own citizens, workplaces, houses and social places;
- citizens may work, interact and establish new social ties;
- according to socio-economic attributes, citizens' happiness will may affected.
- low happiness levels may trigger different actions;

Results presented here are the average over 30 runs for all configurations. Moreover, the following parametric setup were applied:

Table 1. Simulation parameters used in all experiments

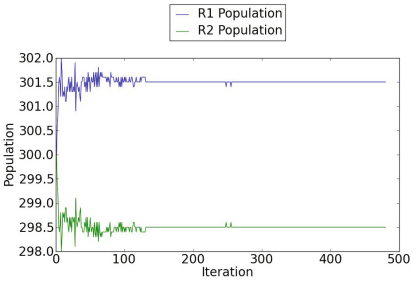
Parameter	Value
R1 average wages	15
R2 average wages	10
Initial unemployment rate	20%
Interaction probability	5%
Migration cost	5
Minimum happiness	5

Scenario 1 - No Communication

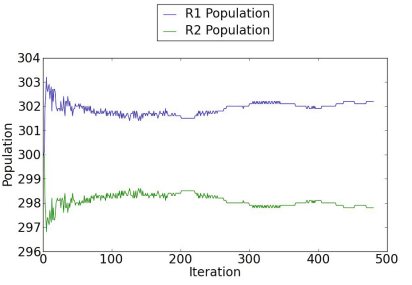
In this scenario, agents are not allowed to communicate to other agents their wages or possible job opportunities. Thus, they are not using social information in their decision-making although social interactions are still occurring. When communication is disabled, agents cannot share information with each other. Figure 1a depicts how population sizes changed over time when the experiment was executed with non-evolutionary agents.

It is noticed that right after first iterations, few agents from R2 (which have average wages lower than R1) migrated to R1 and settled down there. Comparing the behavior produced by non-evolutionary agents (Figure 1a) and the MEAM with 10% of crossover rate (Figure 1b), we can observe slight differences in both plots. In Fig. 1b we can distinguish a oscilatory behavior starting approximtely after 200th iteration.

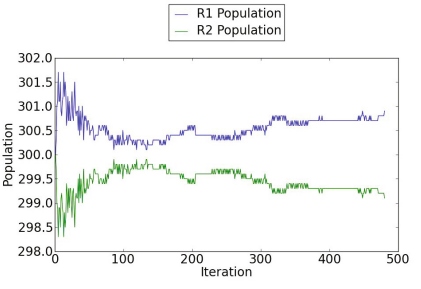
Almost the same behavior may be seen in the scenario with the MEAM with 25% of crossover rate (Fig. 1c). However, the fluctuation also observed in Fig. 1b was intensified and strengthened.



(a) Non-evolutionary agent



(b) crossover rate of 10%



(c) crossover rate of 25%

Fig. 1. Population on each region for non-communicating agents

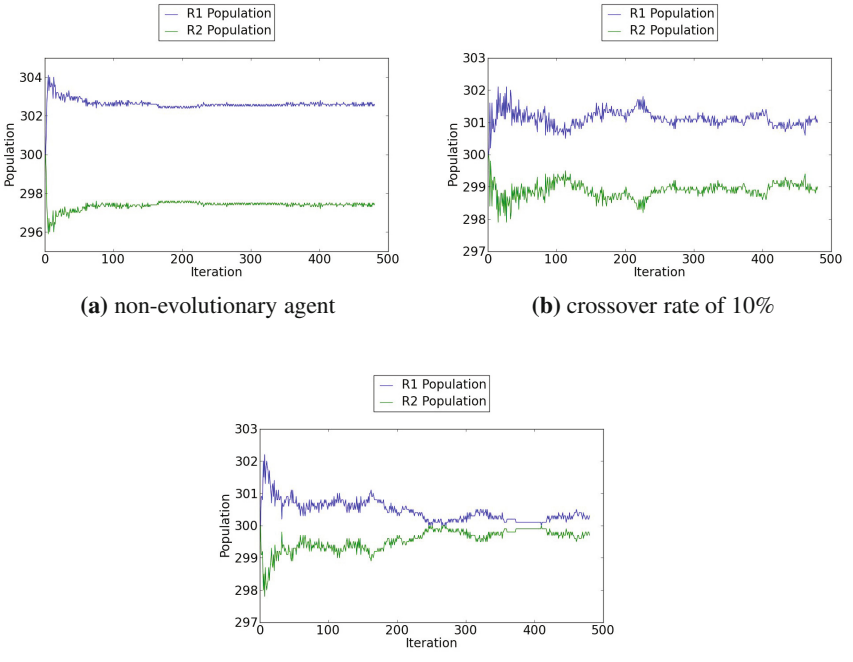
Scenario 2 - Short Distance Communication

In this scenario, agents are able to communicate with surrounding agents. Thus they are able to share social information and use them in their decision-making. In this scenario, information is exchanged within the social network but only both agents are in the same place. This scenario can be mapped to an environment with any kind of telecommunication.

Figure 2a portraits the evolution in population sizes for both regions over time when the experiment was executed with non-evolutionary agents. Similarly to the behavior observed in Fig. 1a, right after the initial iterations, few agents from R2 migrated to R1 and also settled down there.

Comparing the behavior produced by the non-evolutionary agent(Fig. 2a) and the MEAM with 10% of crossover rate (Fig. 2b), we can observe now that differences between both models are widening. MEAM now is able to quickly perceive and act accordingly to environmental changes.

In both experiments here carried out with MEAM (Fig. 2b and Fig. 2c), following the initial flow, a secondary population movement emerges. This behavior means that even those agents that are already settled down and stable in its region are responding to a subtle change in the job market and trying to find a better opportunity.



(a) non-evolutionary agent

(b) crossover rate of 10%

(c) crossover rate of 25%

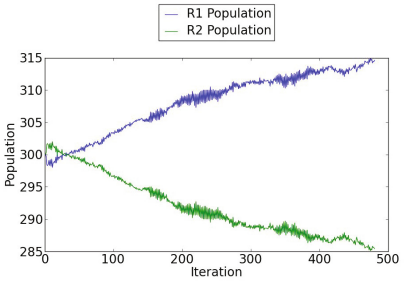
Fig. 2. Population on each region with local communication

Scenario 3 - Social Networks

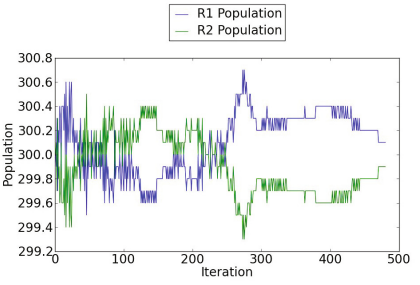
In this scenario we are investigating the influence of social networks on human migratory flows. In this experiment, agents are able to exchange information even with those agents that are in the other region. However, this information will flow only inside the social network. This scenario can be seen as an environment in which people can obtain from their acquaintances information regarding wages paid in other regions as well as possible job opportunities.

In this scenario we can observe few important changes in the overall migration dynamic. From the Fig. 3a we can see that the information flow inside the social network changed significantly the way in which the non-evolutionary agent behave. As soon as the agents from R2 obtained information regarding the better wages offered in R1, they started to migrate to there.

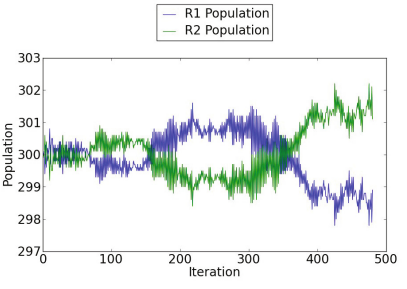
However, like in the previous scenario, those experiments were the MEAM was applied, agents were able to give quick responses to environment changes. Fig. 3b and Fig. 3c depict this behavior.



(a) non-evolutionary agent



(b) crossover rate of 10%



(c) crossover rate of 25%

Fig. 3. Population on each region with long-distance communication

4 Conclusion

The hypothesis tested in this paper was that MEAM is also able to reproduce migratory phenomena even in more complex scenarios, helping social scientist to gain a deeper understanding regarding migratory flows and social networks. In fact, from results shown in here, we can conclude that the multilayer architecture endowed the agent with behaviors not shown by the other model. That means that agent-based models are suited for social simulation of migratory phenomena.

The proposed model, despite having presented high-order behaviors, validating our main hypothesis, it may still contribute to investigations on social sciences. A detailed analysis of individual actions of the agent can provide important data related to several factors (still imponderable) of human behavior. This, we listed here a sort of possible future steps in this research:

- Compare this agent-based approach against analytical models.
- Investigate how cultural aspects may influence migratory behaviors
- Apply this model in investigations on other social contexts such as urban violence and cultural transmission

- To carry out a quantitative analysis comparing results produced by this model with data from 2010 Brazilian Demographic Census which can provide important informations regarding migratory flows.

References

1. Barbosa Filho, H.S., de Lima Neto, F.B., Fusco, W.: Migration and Social Networks - An Explanatory Multi-evolutionary Agent-Based Model. In: IEEE Symposium on Intelligent Agent (IA), pp. 1–7. IEEE (2011)
2. Fazito, D.: Análise de redes sociais e migração: dois aspectos fundamentais do retorno. Rev. bras. Ci. Soc. 25(72), 89–176 (2010), doi:10.1590/S0102-69092010000100007
3. Fusco, W.: Redes sociais na migração internacional: o caso de governador Valadares (2002)
4. Fusco, W.: Capital Social e Dinâmica Migratória: um estudo sobre brasileiros nos Estados Unidos. Textos Nepo 52 (2007)
5. Gilbert, N.: Varieties of emergence. In: Agent 2002 Conference: Social Agents: Ecology, Exchange, and Evolution, Chicago, pp. 11–12 (2002)
6. Gilbert, N., Troitzsch, K.G.: Simulation for the Social Scientist, 2nd edn. Open University Press (2005)
7. Massey, D.S., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., Taylor, J.E.: An Evaluation of International Migration Theory: The North American Case. Population and Development Review 20(4), 699 (1994), doi:10.2307/2137660
8. Petersen, W.: A general typology of migration. American Sociological Review 23(3), 256–266 (1958)
9. Ravenstein, E.G.: The Laws of Migration. Journal of the Statistical Society of London 48(2), 167–235 (1885)
10. Redfern, P.: An Alternative View of the 2001 Census and Future Census Taking. Journal of the Royal Statistical Society. Series A (Statistics in Society) 167(2), 209–248 (2004)
11. Stillwell, J., Congdon, P.: Migration Models. John Wiley & Sons Inc. (1993)

A Comparison of Methods for Community Detection in Large Scale Networks

Vinícius da Fonseca Vieira and Alexandre Gonçalves Evsukoff

Abstract. The modeling of complex systems by networks is an interesting approach for revealing the way that relationships occur and an increasing effort has been spent in the study of community structures. The main goal of this work is to show a comparative study of some of the state-of-art methods for community detection in large scale networks using modularity maximization. In this sense, we take into account not just the quality of the provided partitioning, but the computational cost associated to the method. Hence, we consider many aspects related to the algorithms efficiency, in order to provide the suitability to real scale applications. The results presented in this work are obtained from the literature, in a preliminar sense, and form a solid basis for the implementation and application of efficient algorithms for community detection in large scale networks.

1 Introduction

Many complex systems can be represented through networks, where the elements are "nodes" and the connections between them are "edges". Human societies are an example of complex systems, where persons (nodes) are related through affinities or some kind of social interaction (edges). Understanding the way the relationships among the elements occur helps to reveal the behavior of the entire system. In a network of social interactions, for example, the way the relationships among people occur can directly determine how news are spread or how opinions are formed [15].

Viícius da Fonseca Vieira
COPPE/UFRJ - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
and UFSJ - Federal University of São João del Rei, São João del Rei, Brasil
e-mail: vinicius@ufsj.edu.br

Alexandre Gonçalves Evsukoff
COPPE/UFRJ - Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
e-mail: alexandre.evsukoff@coc.ufrj.br

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 75–86.
springerlink.com

Currently, there is a great interest in researching methods that deal with real scale networks [21, 5, 10]. Particularly, the popularity of the web has led to a widespread use of social networks attracting the attention of large companies that have focused their marketing actions on them. A study released by the research-based advisory firm Altimeter Group among the 100 most valuable companies showed that companies with a deep engagement with consumers through social media channels grew 18% average in a year while the revenue of those that invested little in networks was 6% down [3]. Nowadays Facebook stands out with about 800 million active users estimated reaching over 50% of the population in developed markets such as U.S. and Canada [2, 1]. It is also worth to mention Twitter, with more than 200 million users sending over one billion tweets a week [1].

One of the most important characteristics of network interactions is the formation of communities, i.e. a division of the network in groups that show a great density of internal connections and a low density of external connections [16]. Moreover a large number of methods to detect such communities can be found in literature. Community detection in large scale complex networks is a challenge and many studies focusing this subject can be found in literature [10, 21, 12, 7, 5, 23]. We can also find in the literature a number of works that aim on the comparison of community detection methods. In [13], the authors present a comparison among methods for community detection based on the so called spectral graph theory. The work presented in [10] shows a comparison of a set of greedy algorithms for community detection. In [8], the author presents an extense comparison among methods for community detection in networks, based in different approaches.

This paper focuses on the study of state-of-the-art methods for the detection of communities concerning in the application in large scale complex networks. Based on this study, the methods are compared regarding two main aspects: the quality of the partitions they generate and the computational efficiency of each one of them. Some aspects of the methods that might allow the detection of communities in increasingly larger networks are also approached.

In Section 2 concepts related to the detection of communities in networks through modularity are presented, as well as the spectral approach which offers good results for the partitioning of networks into communities. Section 3 shows algorithms that can solve the problem of detection of communities in large scale networks. A comparison of these algorithms is shown in Section 4, where relevant aspects of each method and their suitability to large networks are pointed out. Finally, Section 5 presents final considerations and suggestions for further work.

2 Community Detection in Networks

The sense of community becomes more evident when the difference between internal and external connection density increases, being necessary to define the criteria for measuring the quality of the partitioning of a network into communities. From the analysis of such criteria, it is possible to define methods that, without any previous information on the structure of the network, are able to differentiate networks

with a well defined community structure from those that present a structure basically formed by random connections [16].

To this end, we consider a graph $G(V,E)$ where V represents the set of n nodes and E the set of m edges and k_i is de degree of node i. We also consider C as a community structure, i.e. a partition of V in c communities, such that $C = \{C_1,...,C_c\}$, where $V = \bigcup_{i=1}^{c} C_i$ and $\bigcap_{i=1}^{c} C_i = \emptyset$. Taking n_{C_i} the number of nodes in C_i, m_{C_i} the number of intra-community connections in C_i, such that $m_{C_i} = |\{(u,v) : u \in C_i, v \in C_i\}|$ and b_{C_i} the number of inter-community connections, such that $b_{C_i} = |\{(u,v) : u \in C_i, v \notin C_i\}|$.

2.1 Modularity

The most frequently used metrics for community structure assessment is modularity [16], which is based on the idea that a vertex subset may be considered a community if the number of internal connections is greater than expected in a random formation. Considering that the chance of a connection of i be one of the k_j connections of j is $k_j/2m$ the modularity Q of a partitioning can be calculated by:

$$Q = \frac{1}{2m}\sum_{ij} A_{ij}\delta(c_i,c_j) - \frac{1}{2m}\sum_{ij}\frac{k_i k_j}{2m}\delta(c_i,c_j), \tag{1}$$

where the first term is the number of connections between vertices of the same community and the second one is the expected for a random version. $\delta(a,b)$ returns 1, if $a = b$ and 0 otherwise. A modularity matrix B can then be defined as:

$$B = A - \frac{kk^T}{2m}. \tag{2}$$

Based on this definition, it is expected that a community structure that corresponds to the maximum modularity value is the best division scheme, or, at least a very good one. This leads to the need of modularity metrics maximization and several studies are described in literature[10, 6, 14, 21, 7, 5, 11].

2.2 Spectral Optimization of Modularity

An efficient way of modularity optimization in networks derives from the spectral partitioning of graphs and the relationship between its structural properties and the eigenvectors of its adjacency matrix. Given a matrix A of dimension $n \times n$, the eigen problem is defined as the search for a scalar λ and a non-zero vector v:

$$Av = \lambda v, \tag{3}$$

where the scalar λ is the eigenvalue and v is the correspondent eigenvector.

Considering the division of a network in only two communities, such division may be represented by a vector $s \in \{-1,1\}^n$, where, for each vertex v_i of the network, we associate a value $+1, if\ v_i \in C$ and $-1, if\ v_i \notin C$. Thus:

$$Q = \frac{1}{4m} s^T B s. \tag{4}$$

The objective then is to find the value of s that maximizes the Equation 4 for a certain matrix B. The elements of s are restricted to the values ± 1, but relaxing this restriction, it points to any direction and an approximate solution of the optimization problem Q is given by:

$$Bs = \beta s, \tag{5}$$

where s is one of the eigenvectors of the modularity matrix B. Considering that $s^T s = \sum_i s_i^2 = n$, we obtain

$$Q = \frac{n}{4m} \beta. \tag{6}$$

For modularity maximization, we choose the value s as the eigenvector u_1 corresponding to the highest eigenvalue of the modularity matrix. However, we cannot choose $s = u_1$, because the elements of s are subjected to the restriction $s_i = \pm 1$. But s may be approximated adopting the values $+1,\ if\ [u_1]_i \geq 0$ and $-1,\ if\ [u_1]_i < 0$. In summary, the eigenvector of the modularity matrix corresponding to the largest eigenvalue is calculated and then vertices are assigned to the communities according to the signs of the elements: elements with positive signs are assigned to a community and elements with negative signs are assigned to another.

It is assumed that the nodes can be divided into c groups and there is no assumption about the value of c. For this reason, the original formulation of the algorithm, proposed by Newman, must be adapted, allowing detecting several communities[13], and this can be done through a process of successive bipartitions . In this way, a division of a community is only performed if it results in some gain in modularity and the possibility of division into smaller groups is analyzed in each one of both parts until it is accepted that a further division will not result in any gain.

To this end, Newman defined the measure ΔQ that assess the modularity variation of the network generated by the bipartition of a community C [15]:

$$\Delta Q = \frac{1}{2m} \left[\frac{1}{2} \sum_{i,j \in C} B_{ij}(s_i s_j + 1) - \sum_{i,j \in C} B_{ij} \right] = \frac{1}{4m} s^T B^{(c)} s, \tag{7}$$

3 Algorithms for Community Detection in Large Scale Networks

A negative aspect of the Newman's spectral optimization algorithm is that the calculation of modularity ΔQ (Equation 7) involves the resolution of an eigenvalue problem, computationally very expensive. This section presents some methods for

community detection aiming at overcoming this inconvenience, enabling the solution of the problem in large scale networks.

3.1 Clauset, Newman and Moore (CNM)

The algorithm proposed by Clauset, Newman and Moore (CNM) uses a greedy strategy for community detection. As described in [6] and [12], it initially associates each vertex of the network to a community and then, repeatedly merges the communities the union of which produces the highest elevation of Q, given by Equation 1. Considering a network with n vertices, after $n-1$ combinations the result is just a big community containing all the vertices and the algorithm stops. Given a vertex v belonging to a community c_v, e_{ij} can be defined as:

$$e_{ij} = \frac{1}{2m} \sum_{vw} A_{vw} \delta(c_v, i) \delta(c_w, j), \tag{8}$$

the fraction of the edges that connect the vertices in community i to the vertices in community j. We can also define a_i as

$$a_i = \frac{1}{2m} \sum_{v} k_v \delta(c_v, i), \tag{9}$$

the fraction of the extremities of the edges associated to the vertices in the community i.

Modularity Q is then written as

$$Q = \frac{1}{2m} \sum_{vw} \left[A_{vw} - \frac{k_v k_w}{2m} \sum_{i} \delta(c_v, i) \delta(c_w, i) \right] = \sum_{i} (e_{ii} - a_i^2). \tag{10}$$

The algorithm aims at finding the combination of communities i and j that results in a greater increase in Q and, then, perform that operation, thereby joining communities with higher affinity [7].

The CNM method performs well in time, reaching up to $O(n\ log^2\ n)$ and, therefore, can be used in the detection of communities in large scale networks [15]. On the other hand, being a greedy algorithm, the CNM method often results in modularity values worse than other methods, such as the spectral. The following sections present some variations of CNM found in literature aiming at detecting and improving its deficiencies, to ensure better modularity values. Some of these alterations are also intended to improve the execution time of the CNM.

3.2 Danon, Diaz and Arenas (DDA)

Danon, Diaz and Arenas (DDA) identified that when there are inhomogeinities in sizes of the communities being processed, the CNM algorithm presents a big trend for the detection of the largest communities, sacrificing the smaller communities and considerably impairing the quality of the community structure found [7]. Starting

from the original CNM formulation, in [7] they propose a modification of CNM which treats communities of different sizes on an equal footing.

To this end, they normalized the modularity variation ΔQ, dividing it by k_i, which represents the fraction of all the edges with vertices in community i. Thus, the calculation of modularity variation ΔQ is given by

$$\Delta Q^*_{ij} = \frac{\Delta Q_{ij}}{k_i}. \tag{11}$$

3.3 Wakita and Tsurumi (WT)

Another modification of the CNM algorithm proposed by Wakita and Tsurumi (WT) is based on the observation that merging communities of very different sizes greatly impacts the computational efficiency of the CNM algorithm [21]. From this idea, merging communities in a balanced way should improve the efficiency of the algorithm.

In the modification proposed by WT, a heuristics is defined that attempts to merge community structures in a balanced manner consequently improving the computational efficiency of the CNM method. The heuristics, called consolidation ratio $rcons(c_i, c_j)$ can be defined as:

$$rcons(c_i, c_j) = min\left(\frac{|c_i|}{|c_j|}, \frac{|c_j|}{|c_i|}\right). \tag{12}$$

where c_i may be interpreted in terms of the degree of the community in process or in terms of the number of vertices.

Thus, the modification of the CNM algorithm is performed choosing the community pair (i, j). Instead of choosing the pair with maximum value of ΔQ_{ij}, as in CNM, the pair with maximum value $\Delta Q_{ij} \times rcons(c_i, c_j)$ is chosen.

3.4 Leon-Suematsu and Yuta Method (LY)

Leon-Suematsu and Yuta (LY) analyzed the CNM algorithm and identified possible forms of vertex incorporation when merging communities in process i and j for updating ΔQ [10].

Through computational experiments, in [10] it was observed that when there is a great number of combinations of communities with one vertex linked to community i and not-linked to community j (approximated by the number of interconnected communities *nci*), the behavior of the execution is seriously harmed. The following factor was defined to reduce this effect:

$$factor(i, j) = \frac{1}{\max(nci_i, nci_j)} \tag{13}$$

Initially, $max_k \Delta Q_{ik}$ is identified for a community i and then the factor is applied just to $max_k \Delta Q_{ik}$, different from WT modification that applies the consolidation ratio to all the elements. The selection of the community pairs to be operated is performed through

$$\max_i [(\max_j \Delta Q_{ij}) \times factor(c_i, c_{k_i})]; k_i = \arg\max_j \Delta Q_{ij}. \quad (14)$$

3.5 Blondel, Guillaume, Lambiotte and Lefebvre (BGLL)

Blondel, Guillaume, Lambiotte and Lefebvre (BGLL) proposed a hierarchical method for community detection based on modularity maximization performed in two phases [5]. In the first phase, the BGLL algorithm assigns a community to each vertex of the network, thus generating n unary communities. Then, for each vertex i of the network, the algorithm takes into consideration its neighbor j and places i in the community where the result is the maximum modularity value. This process is iteratively applied to all the vertices until the modularity cannot be further improved. The second phase of the algorithm starts creating a new network where each new vertex represents a community found in the previous phase. The iterative process defined in the first phase is applied on this new network. The simulations performed by the authors suggest that the execution time presents an almost linear complexity, but it is not possible to exactly assess this statement.

3.6 Parallel Computing Applied to Community Detection

The algorithms for community detection in complex networks found in literature, although being potentially applied to large scale complex networks, present bottlenecks that prevent the scalability of their use, due to high execution time or the impossibility of storing data structures in memory.

According to [21], the WT method was not applied to larger networks due to processing issues. Besides the limitation of the execution time, it was pointed out that a possible limitation on the size of the network was the capacity of storing data structures in memory. The BGLL method was executed in a network with 118 million nodes in only 9120 seconds. However, the scalability of the BGLL method was limited by the storage in memory. In these cases, an approach widely used to reduce execution time is parallelism, i.e. the simultaneous use of the computing power of several processing cores for executing an algorithm [19]. Currently, there is a strong tendency to use personal computers with multi-core processors spreading the use of parallelism to solve computational problems. It is, thus, a powerful tool for reducing the execution time of large scale applications such as community detection in real complex networks [19]. There are several studies in literature which successfully apply parallel computing to problems involving large scale networks [18, 4].

In the methods based on greedy heuristics, the analysis of the nodes to assess the best way of community merging may be divided among several processors, reducing the global processing time. Besides, the graph storage may be divided among

several memory units, enabling processing increasingly larger networks. The main bottleneck of the spectral optimization methods is the solution of the eigenvalue problem. Using the power method [9], the process is reduced to iterative multiplications matrix-vector, which is a very important operation in numerical methods and has motivated the application of parallelism to this kind of operation [22, 20]. This strategy applied to the solution of the eigenvalue problem turns feasible the usage of spectral approaches in large scale networks.

4 Algorithm Comparison

This section presents a comparison of the methods previously presented in two ways: in terms of the modularity of their results and in terms of their computational efficiency. Through these comparisons, some aspects which allow an implementation of these methods more suitable for large scale networks are enhanced.

4.1 Quality of Community Division

The quality of community partition by the methods addressed here was compared through results found in literature [13, 10]. Table 1 summarizes the comparison between the CNM and Newman's spectral method (refered below as Newman) in networks of up to 27 000 nodes, found in [13] and a comparison of the methods based on greedy heuristics in larger scale networks, found in [10].

Table 1. Algorithm Comparison

		Modularity Q				
Network	n	CNM	Newman	DDA2	WT	LY
Karate	34	0.381	0.419	-	-	-
Jazz musicians	198	0.439	0.442	-	-	-
Metabolic	453	0.402	0.435	-	-	-
E-mail	1133	0.494	0.572	-	-	-
Key signing	10680	0.733	0.855	-	-	-
Physicists	27519	0.668	0.723	-	-	-
Mixi	360,802	0.601	-	0.666	0.466	0.615
YouTube	1,138,499	0.705	-	0.703	0.552	0.646
LiveJournal	5,204,176	0.686	-	0.737	0.433	0.648
Orkut	3,072,441	-	-	0.663	0.380	0.540
Facebook	63,731	0.606	-	0.592	0.385	0.500

The Newman's spectral method offers results with better modularity values in all the networks where it was applied. When compared to other methods based on greedy heuristics, the DDA method presents the best modularity values, indicating that the concern in avoiding the trend of attraction of larger communities

really results in a better partitioning quality. Although of fundamental importance for the understanding of the methods, the analysis of the results found in literature shows the need to conduct a more comprehensive study on the same execution conditions. Thus, we could compare the methods in a more complete and therefore more conclusive way.

4.2 Computational Efficiency

The calculation of the dominant eigenvector, performed in the methods based on spectral approaches is conducted in $O(mn)$, i.e. $O(n^2)$ on a sparse matrix and $O(n^3)$ on a dense matrix. Thus, a big problem emerges from the fact that the modularity matrix B is not spase. Nevertheless, [13] presents a way to explore special properties of B, enabling to take advantage of the network sparsity. The power method applied to the matrix B corresponds to iterative multiplications Bx, where x is the vector that will converge to the eigenvector. Applying the multiplication Bx to Equation 2, we obtain:

$$Bx = Ax - \frac{k(k^T x)}{2m}. \tag{15}$$

The term Ax is a sparse matrix-vector multiplication, executed on time $O(m+n)$. The second term corresponds to an inner product $(O(n))$. Thus, the multiplication Bx is $O(m+n)$ and, considering $O(n)$ multiplications, the total time to find the dominant eigenvector is $O[(m+n)n]$, although in practice, this value is much lower. Then, in a sparse matrix, the execution time tends to $O(n^2)$. Considering the division of the network into several communities, that, in a real case is of order $log\ n$, we obtain a total time for the execution of the algorithm to unfold communities in networks is of order $O(n^2 log\ n)$ for a sparse network.

An application of the spectral optimization method in a network of 27 000 nodes, executed in about 20 minutes can be found in [13]. However, the author states that the method is reasonable only for networks of up to 100 000 nodes on personal computers. It is known that real situations present networks on larger scales than those approached by [13] and thus heuristics need to be frequently used for modularity optimization.

The CNM algorithm uses a greedy heuristics to merge communities and thus, it does not ensure optimality of the partitioning, which is one of the main disadvantages of this method when compared to the spectral approach. On the other hand, its ability to work with large scale networs allows the application to real scale networks. The CNM method was formulated with $O(n^2)$, and reduced to $O(n\ log^2\ n)$ in following studies[15]. Therefore, it is capable of dealing with networks on a scale impossible for other methods, such as Newman's spectral method.

Some methods based on CNM aim on reducing the computation time while being suitable for large scale networks. By the application of the WT method, it has been noticed, empirically, the execution time reduction of the CNM method [21], and so, it could be applied to a network of about 5.5 million nodes. However, the authors

could not determine the causes of such improvement and besides, and a weak point of the WT method is the low quality of the communities found, as observed in Table 1. LY method was proposed in order to control the combination of communities, which can directly affect the performance of the CNM [10]. The BGLL method is enhanced because of its application to larger scale networks than those found in other studies, and was applied to a network of about 118 million nodes [5]. In [5], the authors indicate that the order in which the vertices of a network are analyzed can affect the time of execution of the method, although this was concluded based on empirical results.

With the purpose of comparing the efficiency of methods based on greedy heuristics in relation to time, Table 2 presents a comparison of CNM based methods, as seen in [10]. In that work, the methods, implemented in C++, were executed in a CPU Xeon 2.8 GHz with 64 GB Ram PC running a Red at Linux. It can be observed that in Table 2 only some of the networks presented in Table 1 were used. The reason is the absence of such results in the original study [13].

Table 2. Comparison between the execution time of the greedy algorithms in seconds

		Time (s)			
Network	n	CNM	DDA2	WT	LY
Mixi	360,802	4,747	417	639	288
YouTube	1,138,499	11,892	2,091	3,852	2,631
LiveJournal	5,204,176	810,302	114,898	56,767	43,059
Orkut	3,072,441	-	275,154	41,691	26,960
Facebook	63,731	72	15	11	11

In Table 2 it can be observed that the LY method presented good results in relation to execution time when compared to the other methods.

From the analysis of the results we can notice that the LY modification allowed a drastic reduction in the number of chain-like operations, which reached 81 times in some cases. Also, there is a great reduction in the execution for the LY modification when compared to the original CNM method. In the experiments, the execution time was reduced in a factor between 3 and 45 when compared to the original CNM. In respect to the modularity, comparing the LY and the CNM methods, LY presented better results than CNM in about 40% of the tests. When the results are compared with the DDA modification, better results are obtained in about 60% of the cases; CNM presented better results in about 40% tests. When Tables 1 and 2 are analyzed together, we can observe that the LY method presents a good execution time while preserving the quality of the partitioning which makes its use very attractive. Again, it is important to enhance that a more detailed study is necessary to be conducted, which allows a more solid base of comparison of the methods presented. It is particularly important to compare directly CNM based methods with spectral approaches

and BGLL, because results of a wide application of these methods to benchmark large scale networks could not be found.

5 Conclusions and Future Directions

The formation of communities is one of the properties most frequently observed in complex networks and the study of methods for the partitioning of networks into communities has received great attention. The present paper presents a study of some of the main methods for community detection in large scale networks and a comparison of these methods using results found in literature.

The study was conducted focusing on two aspects: partitioning quality and computational cost. The partitioning quality was assessed through the modularity value obtained and it was observed that the Newman's spectral optimization approach presented better results than the heuristics methods. However, the high computational cost associated to the eigenvalue problem makes its application difficult in real scale networks. In these cases, the use of algorithms based on heuristics proves to be more efficient. We pointed out several features which, when properly worked, can reduce the execution time of the algorithms.

The comparison of the methods enabled the formation of a solid base for the understanding of the characteristics of each one of them. However, we intend to include in future studies, the implementation of each one of the methods considering characteristics that lead to a reduction of the computational cost. Then, the methods should be compared in equal experimental conditions, applied on networks with different characteristics, so as to obtain an exact knowledge of the suitability of each method to each context. Additionally, we intend to extend the range of methods compared, including different approaches, such as the particle competition method [17], which proved to be an efficient approach for dealing with the community detection problem. In future works, we also intend to use parallel computing, enabling the best use of multiprocessor computers, very popular nowadays.

Acknowledgements. The authors are grateful to the agencies CNPq and CAPES for their financial support and to the agreement CAPES/Cofecub which supports the work.

References

1. Facebook reaches 800 million users, http://migre.me/7vEqr (last access in January 12, 2012)
2. Facebook statistics by country, http://www.socialbakers.com/facebook-statistics/ (last access in January 12, 2012)
3. New study: Deep brand engagement correlates with financial performance, http://www.altimetergroup.com/2009/07/engagementdb.html (last access in January 12, 2012)

4. Arjomandi, E., Corneil, D.G.: Parallel computations in graph theory. In: 16th Annual Symposium on Foundations of Comp. Science, pp. 13–18 (October 1975), doi:10.1109/SFCS.1975.24
5. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment (October 2008)
6. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Phys. Rev. E 70 (2004)
7. Danon, L., Diaz-Guilera, A., Arenas, A.: Effect of size heterogeneity on community identification in complex networks. Journal of Stat. Mech.: Theory and Experiment 6 (November 2006)
8. Fortunato, S.: Community detection in graphs. Physics Reports 486, 75–174 (2010)
9. Heath, M.T.: Scientific Computing: An Introductory Survey, 2nd edn. McGraw-Hill (2002)
10. Leon-Suematsu, Y.I., Yuta, K.: Framework for fast identification of community structures in large-scale social networks. In: Data Mining for Social Network Data, Annals of Information Systems, vol. 12, pp. 149–175. Springer, US (2010)
11. Newman, M.E.J.: Coauthorship networks and patterns of scientific collaboration. Proc. of the National Academy of Sciences of the USA 101, (suppl. 1), 5200–5205 (2004)
12. Newman, M.E.J.: Fast algorithm for detecting community structure in networks. Physical Review E 69(2), 1–5 (2004)
13. Newman, M.E.J.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences of the United States of America 103(23), 8577–8582 (2006)
14. Newman, M.E.J.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103(23), 8577–8582 (2006)
15. Newman, M.E.J.: Networks: An Introduction, 1st edn. Oxford University Press, USA (2010)
16. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Review. E, Stat., Nonlinear and Soft Matter Physics 69(2) (2004)
17. Quiles, M.G., Zhao, L., Alonso, R.L., Romero, R.A.F.: Particle competition for complex network community detection. Chaos Woodbury NY 18(3), 107 (2008)
18. Quinn, M.J., Deo, N.: Parallel graph algorithms. ACM Comput. Surv. 16, 319–348 (1984), doi: http://doi.acm.org/10.1145/2514.2515
19. Sloan, J.D.: High Performance Linux Clusters with OSCAR, Rocks, OpenMosix, and MPI. O'Reilly (2004)
20. Tavakoli, F.: Parallel sparse matrix-vector multiplication (1997)
21. Wakita, K., Tsurumi, T.: Finding community structure in mega-scale social networks. Analysis 105(2), 9 (2007)
22. Williams, S., Oliker, L., Vuduc, R., Shalf, J., Yelick, K., Demmel, J.: Optimization of sparse matrix-vector multiplication on emerging multicore platforms. Parallel Computing 35(3), 178–194 (2009)
23. Zhang, Y., Wang, J., Wang, Y., Zhou, L.: Parallel community detection on large networks with propinquity dynamics. In: Proceedings of the 15th ACM International Conference on Knowledge Discovery and Data Mining, p. 997 (2009)

Stable Community Cores in Complex Networks

Massoud Seifi, Ivan Junier, Jean-Baptiste Rouquier, Svilen Iskrov, and Jean-Loup Guillaume

Abstract. Complex networks are generally composed of dense sub-networks called communities. Many algorithms have been proposed to automatically detect such communities. However, they are often unstable and behave non-deterministically. We propose here to use this non-determinism in order to compute groups of nodes on which community detection algorithms agree most of the time. We show that these groups of nodes, called community cores, are more similar to Ground Truth than communities in real and artificial networks. Furthermore, we show that in contrary to the classical approaches, we can reveal the absence of community structure in random graphs.

1 Introduction

Complex networks appear in various contexts such as computer science (networks of Web pages, peer-to-peer exchanges), sociology (collaborative networks), biology (protein-protein interaction networks, gene regulatory networks). These networks can generally be represented by graphs, where vertices represent entities and edges indicate interactions between them.For example, a social network can be represented by a graph whose nodes are individuals and edges represent a form of social relationship. Likewise, a protein-protein interaction network can be modeled by a graph whose nodes are proteins and edges indicate known physical interactions between proteins.

Massoud Seifi · Jean-Loup Guillaume
Pierre and Marie Curie University, Laboratory of Computer Sciences, Paris VI (LIP6), 4, place Jussieu 75005 Paris, France
e-mail: {firstname.lastname}@lip6.fr

Ivan Junier · Jean-Baptiste Rouquier · Svilen Iskrov
Centre for Genomic Regulation (CRG), Barcelona, Spain
Institute of Complex Systems, Paris le-de-France, 57/59, rue Lhomond 75005 Paris, France
e-mail: {firstname.lastname}@iscpif.fr

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 87–98.
springerlink.com

An important feature of such networks is that they are generally composed of highly interconnected sub-networks called communities [6]. Communities can be considered as groups of nodes which share common properties and/or play similar roles within the graph. The automatic detection of such communities has attracted much attention in recent years and many community detection algorithms have been proposed, see [11] for a survey. Most of these algorithms are based on the maximization of a quality function known as *modularity* [14], which measures the internal density of communities. Modularity maximization is an NP-complete problem [3], and most algorithms use heuristics. For several reasons related to the modularity, as well as the non-determinism of the algorithms or randomness in initial configuration, such algorithms may produce different partitions of similar quality and there is no reason to prefer one above another. Besides, such algorithms may find communities with a high modularity in networks which have no community structure, e.g. random networks [8]. This is related to the instability of algorithms as shown in [1]: small perturbations of the input graph can greatly influence the output.

Here, we assume that, if several community detection algorithms, or multiple executions of a non-deterministic algorithm agree on certain sets of nodes, then these sets of nodes are certainly more significant. On this basis, we study the tendency of pairs of nodes to belong to the same community during multiple executions of a non-deterministic community detection algorithm. Experimental results on both artificial and real networks show the performance of this concept and we show in particular that it allows to distinguish random from non-random networks.

We provide a general description of algorithms used for detecting consensus communities in Section 2. We then present our previous contributions in Section 3. Finally, we describe the experimental results on artificial and real networks in Section 4 and on random networks in Section 5 before concluding in Section 6.

2 Algorithms for the Identification of Consensus Communities

Two main methods have been used to combine different partitions into a set of consensus communities. One is based on network perturbations. The other one takes advantage of changing the initial configuration of the algorithms.

Network perturbations: Since most community detection algorithms are deterministic, small perturbations can be made on the network to obtain different results. Then, communities are found in each modified network and compared to the partition of the original network to obtain consensus communities. Several methods of network perturbations are proposed in the literature. For example in [9] the method involves removing a fraction of links and putting them back between randomly selected pairs of vertices. Another

technique consists in adding noise to the weight of links, i.e. slightly change them in order to influence the algorithm. For example, in [17] it is proposed to change the weight of links using a Poisson distribution whose parameter is the average weight of links in the original graph. In [5], the noise added to the weight of a link between nodes i and j, initially equal to w_{ij}, is given by a distribution between $-\sigma w_{ij}$ and σw_{ij}, where σ is a constant parameter. A weakness of this method is that it needs an additional parameter σ, whose value is in principle arbitrary. In addition, these studies consider only pairs of adjacent nodes. We will see later that we may identify nodes with a strong tendency to be in the same community even if there is no direct link between them. Also, in these studies, the comparison was made with the partition of the original network, whose significance is not obvious.

Changing the initial configuration of an algorithm: Most algorithms start with an initial partition which is modified many times until a high quality partition is obtained. In general, the algorithms are very sensitive to the initial partition and modifying it may lead to different outcomes. This method is used in [16], to identify overlapping communities by identifying stable and unstable nodes. In [10] this method is used in order to detect communities in multi-scale networks.

There are also similar methods in ensemble clustering like [19] but here we study networks, i.e. structured data, not an unordered set of vectors. In this article, we use the second approach by randomizing the order in which nodes are considered. In addition, we consider all pairs of nodes and not only connected pairs of nodes.

3 Community Cores

Given a graph $G = (V, E)$ with $n = |V|$ vertices, we apply $\mathcal{N}$ times a non-deterministic community detection algorithm $\mathcal{A}$ to G. In the following we use the non-deterministic algorithm known as Louvain method [2]. At the end of an execution, each pair of nodes $(i, j) \subseteq V \times V$ can be classified either in the same community or in different communities. We keep track of this in a matrix of size $n \times n$, which we denote by $P_{ij}^{\mathcal{N}} = [p_{ij}]_{n\times n}^{\mathcal{N}}$, where p_{ij} represent the fraction of the $\mathcal{N}$ executions in which i and j were classified in the same community. Note that $p_{ij} = p_{ji}$, and we set $p_{ii} = 0$. From $P_{ij}^{\mathcal{N}}$, we create a complete weighted graph $G' = (V, E', W)$, where the weight of the link (i, j) is p_{ij}. Finally, given a threshold $\alpha \in [0, 1]$, we remove all links having $p_{ij} < \alpha$ from G' to obtain the *thresholded virtual graph*, ${G''}_{\alpha}$. The connected components in ${G''}_{\alpha}$ obtained with a given α are called α-cores, which are non-overlapping sets of nodes.. A pseudo-code version of this algorithm is given in Algorithm 1. We now analyze the impact of the parameters on the results.

Algorithm 1. Core detection

1. Input: a graph $G = (V, E)$, a threshold α, a number of executions $\mathcal{N}$, a non-deterministic community detection algorithm $\mathcal{A}$
2. Apply $\mathcal{N}$ times the algorithm $\mathcal{A}$ to G
3. Create a matrix $P_{ij}^{\mathcal{N}} = [p_{ij}]_{n\times n}^{\mathcal{N}}$ where p_{ij} is the proportion of times that i and j belonged to the same community
4. Create a complete weighted graph $G' = (V, E', W)$ with $|V|$ nodes and p_{ij} as weights
5. Remove all edges ij where $p_{ij} < \alpha$ from G' to obtain ${G''}_{\alpha}$
6. The connected components of ${G''}_{\alpha}$ are α-*cores*

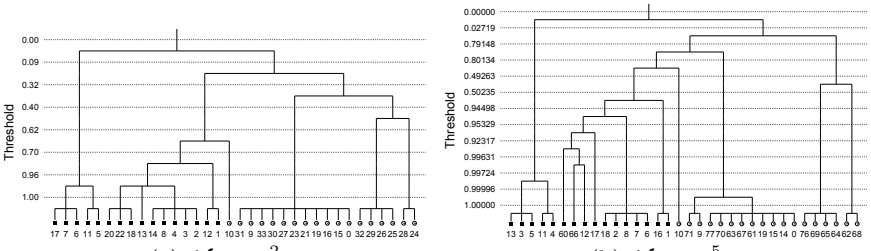
Number of executions: We can estimate the variation of p_{ij} after each execution of algorithm $\mathcal{A}$ by calculating the Euclidean distance between p_{ij} values as a function of the number of executions $\mathcal{N}$. As shown in [16], the variation of p_{ij} converges when the number of executions $\mathcal{N}$ increases. It is therefore possible to terminate the iteration when the variation of p_{ij} is small enough. We derive no theoretical bound on the minimum number of executions to ensure good statistical significance on the estimators p_{ij}. However, we observe that even with an order of magnitude larger of the number of executions, the results do not change much.

Threshold: The threshold α has a strong influence on the results of the algorithm. The proposed algorithm does not aim at finding the largest sets of nodes that are *all* connected to each other with a $p_{ij} \geq \alpha$. There are two reasons for this: (i) the calculation of cores in this case would consists in finding the largest cliques in G', which is an NP-complete problem and (ii) cores could then overlap, which is not allowed in our case. More precisely, given a threshold α, a core may contain pairs of nodes connected with a probability smaller than α.

3.1 Hierarchical Structure of Cores

The parameter α has a strong influence on the size of the cores, it furthermore allows to obtain a hierarchical structure of cores. Indeed α_1-cores are included in α_2-cores if $\alpha_1 > \alpha_2$, i.e. α_1-cores are sub-cores of α_2-cores. Let us discuss this on an example.

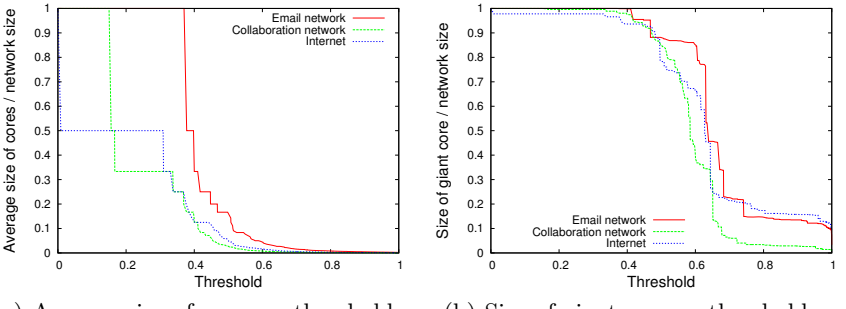
The Algorithm 1 is applied to the famous friendship network of Zachary's karate club [21]. Figure 1 shows the dendrograms of this network for $\mathcal{N} = 10^2$ and $\mathcal{N} = 10^5$, while Figure 3 shows the cores identified by our algorithm. We can see that the division found by the algorithm with $\mathcal{N} = 10^2$ and $\alpha = 0.32$ corresponds almost perfectly to the Ground Truth and only node 10 is misplaced. Note that the number and size of cores is greatly influenced by the choice of α.



(a) $\mathcal{N} = 10^2$ (b) $\mathcal{N} = 10^5$

Fig. 1. Hierarchical structure of cores of Zachary's network for $\mathcal{N} = 10^2$ and $\mathcal{N} = 10^5$

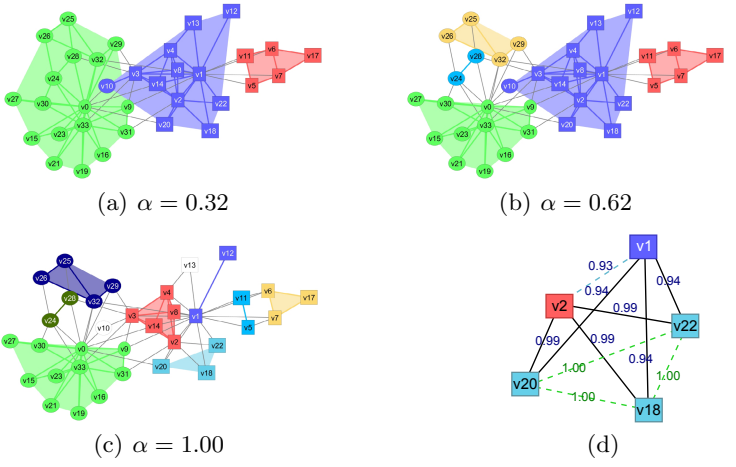
We also applied our algorithm to graphs of different sizes from different domains, including a collaboration network [13], an email network [7] and a snapshot of the Internet (created by M. Newman, unpublished). As Figure 2(a) shows, with a threshold close to zero we obtain very large cores (even larger than the communities) and a strict threshold e.g. $\alpha = 1$ will lead to tiny cores, most of which consisting in only one single node (called trivial cores). We also observe in Figure 2(b) that with an $\alpha < 0.5$, we have a giant core containing the majority of nodes. When the threshold increases, the cores will split quickly into small cores. But in the Internet or email network we still have a giant core containing 10% of the nodes even with an α equal to 1. Note that community partitions also contain a giant community.



(a) Average size of cores vs threshold α (b) Size of giant core vs threshold α

Fig. 2. Impact of threshold to the size of the cores

It must be noted that, as explained above, the nodes inside a core are not necessarily connected in the original network. For example, in Figure 3(c), a core containing the nodes 18, 20 and 22 is identified with a threshold $\alpha = 1$, however, there is no direct link between these three nodes in the original graph. As Figure 3(d) illustrates, these nodes were always together either in the community $c_x = \{1, 18, 20, 22, \ldots\}$ or in the community

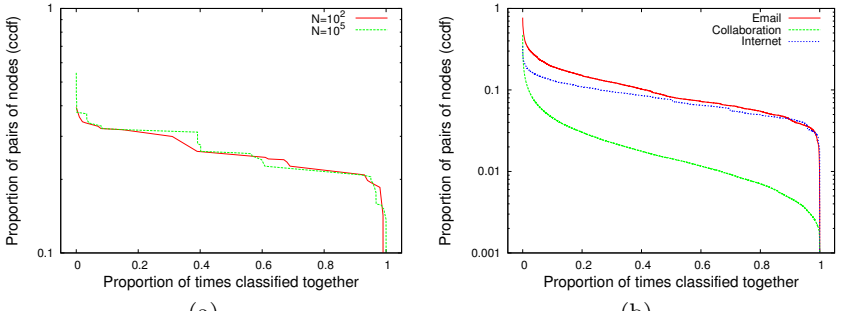


(a) $\alpha = 0.32$ (b) $\alpha = 0.62$ (c) $\alpha = 1.00$ (d)

Fig. 3. (a), (b) and (c) Cores for Zachary's network using three different thresholds. The shape of the nodes (circle/square) is the manual classification made by Zachary. (d) A subgraph of the virtual graph of Zachary's network.

$c_y = \{2, 18, 20, 22, \ldots\}$. This property is interesting and shows that we can identify groups of nodes with a strong tendency to be together even if they have no direct link.

We studied the distribution of p_{ij} in the matrices. The Figure 4(a) shows the p_{ij} distributions of the Zachary's network for $\mathcal{N} = 10^2$ and $\mathcal{N} = 10^5$. As we can see, most pairs are nearly always grouped or separated, but there are some pairs of nodes in the middle which are sometimes together and sometimes separated. The nodes constituting those pairs are less stable. We observe on Figure 4(b) that even on large graphs the majority of pairs are never classified together, and that a significant number of pairs of nodes are always in the same cluster. A large fraction of these pairs are linked in the original network.



(a) (b)

Fig. 4. p_{ij} distribution for (a) Zachary's network and (b) three real-world networks

4 Significance of Cores

We now apply our method to some artificial and real networks having a known community structure, to evaluate the significance of the identified cores.

A classical approach to evaluate the quality of a cluster partition consists in comparing the similarity of the clusters with known communities (or Ground Truth). Various measures of similarity between two clusterings have been proposed [15], and the most widely used is the mutual information, from information theory. It counts the number of bits shared by two random variables. Despite the popularity of mutual information, there are many ways to normalize it, which lead to different values, without definitive solution. Also, it is shown that the mutual information depends on the size of the partitions [20], therefore we used an adjusted version of this metric, called AMI [20]. We also used the edit distance presented in [1] which gives similar results and in some cases is simpler to interpret (data not shown).

Girvan and Newman artificial network [6]: Each graph is constructed with 128 vertices in 4 groups of 32 vertices each. Vertices of the same group are linked with a probability p_{in}, whereas vertices of different groups are linked with a probability p_{out}. Each subgraph corresponding to a group is therefore an Erdös-Rényi random graph [4] with connection probability p_{in}. The probabilities are chosen so as to obtain an average degree $z = 16$. With $p_{in} > p_{out}$ the intra-cluster edge density exceeds the inter-cluster edge density and the graph has a community structure. Figure 5(a) shows a comparison of the similarity of cores and communities to Ground Truth for $z_{out} = 8$ which is a value of z_{out} for which most community detection algorithms fail to identify communities. As we can see, for some α, cores are more similar to Ground Truth than communities.

American College football: This network is also a popular test network with a known community structure [6]. We compare our results with the known partitioning and we find that our algorithm reliably detects the known structure: cores are more similar to known community structure than communities for a wide range of α (see Figure 5(b)).

Another metric that we have used to evaluate the significance of cores is the p-value. The p-value is the probability of obtaining a test statistic at least as extreme as the observed one, assuming that the null hypothesis is true, i.e. assuming that the observed structure is only due to chance. The p-value varies between 0 and 1. The lower the p-value, the stronger the test rejects the null hypothesis, i.e. confirms the significance of the results.

Proteome network: We used this metric to evaluate the significance of identified cores on the Baker's yeast proteome network [18]: nodes are proteins and there is a link when two proteins have been shown to interact. Proteins can work together to achieve a particular function and we used these functions (for instance metabolism or replication) as Ground Truth: a correlation

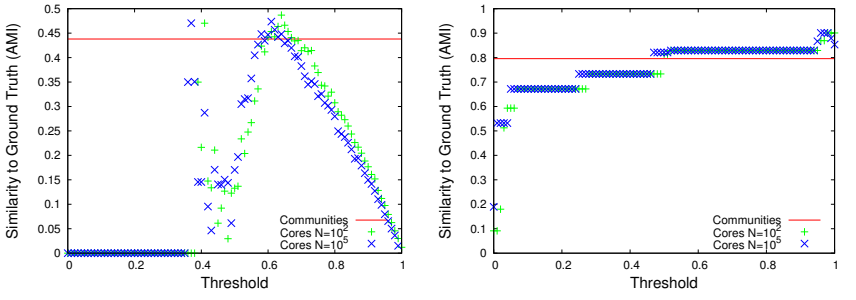


Fig. 5. (a) Girvan-Newman artificial network. (b) American College football

between the clustering and the functions would validate the clustering. We define the null hypothesis as stating that a core is a random subset of the nodes, of a given size. Thus, for a given function, the number of proteins (or nodes) in the core having this function should follow a hypergeometric law. A small p-value thus denotes the fact that many more proteins than expected have the mentioned function: the nodes have not been chosen at random, but with a bias towards this function.

In Table 1, by comparing the lines having the same label, e.g. "GO:0070478", between the cores and communities, one can see that cores have smaller p-values, except for a few big groups with extremely small p-values, where our method removes some nodes from the group yielding a slightly worse p-value. Also, in cores table, the p-values are smaller when $\alpha < 1$, which means that there is a higher correlation between cores and functions. These findings show that our methodology helps to find relevant sets of cofunctional nodes.

5 Random Graphs

We have shown that cores are efficient at finding a Ground Truth on real and artificial networks. In random graphs, the nodes are linked independently to each other so a strong inhomogeneity in the density of links on these graphs is not expected. Therefore random graphs should not have communities. But, as shown in [8], due to fluctuations it is possible to find a partition which has a high modularity for random networks. A good algorithm should indicate both the presence and the absence of community structure. In the following we show that cores cannot be found in random graphs, using two different random graphs model: the classical Erdös-Rényi model [4] and the configuration model [12], which is a construction model that has the degree distribution as an input but is random in all other respects.

First of all, Figure 6(a) shows the impact of the number of execution $\mathcal{N}$ on the distribution of the p_{ij} for a random graph $G(n, M)$. While there are some high values of p_{ij}, there is a high concentration of p_{ij} at an average value

Table 1. Table of p-values for Baker's yeast proteome network . The parameters to compute the p-value are: g_s: total number of nodes in the network, g_f: number of proteins having this function among all the nodes of the network. c_s: size of the core c_f: number of protein in the present core having this function

(a) Cores

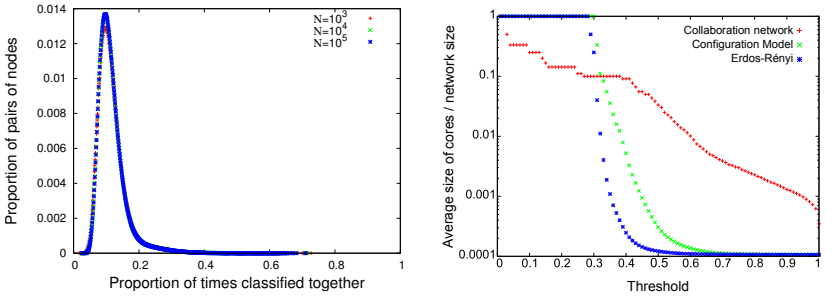
	function	p-value	g_s	g_f	c_s	c_f
$\alpha = 1.00$	GO:0016021	3e-134	5033	927	332	258
	GO:0055085	3e-58	5033	244	332	100
	GO:0005763	1e-45	5033	28	30	21
	GO:0005847	7e-36	5033	15	20	14
	GO:0005789	8e-36	5033	187	332	69
	GO:0016455	7e-35	5033	23	19	15
	GO:0016592	1e-34	5033	24	19	15
	GO:0051123	1e-33	5033	17	15	13
	GO:0032040	1e-33	5033	41	22	17
	GO:0000176	2e-33	5033	13	14	12
$\alpha = 0.99$	GO:0016021	4e-161	5033	927	398	307
	GO:0055085	2e-67	5033	244	398	116
	GO:0005730	1e-51	5033	180	180	65
	GO:0000398	6e-50	5033	58	164	41
	GO:0005763	1e-47	5033	28	32	22
	GO:0005762	2e-41	5033	36	31	21
	GO:0005789	3e-41	5033	187	398	80
	GO:0046540	2e-40	5033	28	164	27
	GO:0016455	3e-37	5033	23	142	23
	GO:0071004	9e-37	5033	29	164	26
$\alpha = 0.98$	GO:0016021	2e-161	5033	927	407	311
	GO:0055085	3e-66	5033	244	407	116
	GO:0005730	5e-51	5033	180	223	70
	GO:0000398	9e-51	5033	58	173	42
	GO:0005763	1e-47	5033	28	32	22
	GO:0005789	1e-42	5033	187	407	82
	GO:0005762	2e-41	5033	36	31	21
	GO:0046540	1e-39	5033	28	173	27
	GO:0071004	4e-36	5033	29	173	26
	GO:0070478	2e-35	5033	17	18	14

(b) Communities

function	p-value	g_s	g_f	c_s	c_f
GO:0016021	8e-170	5033	927	456	338
GO:0055085	9e-075	5033	244	456	129
GO:0005730	2e-050	5033	180	343	82
GO:0005789	1e-044	5033	187	456	88
GO:0000398	6e-044	5033	58	345	46
GO:0005680	3e-031	5033	16	14	12
GO:0046540	3e-031	5033	28	345	27
GO:0008054	3e-030	5033	12	14	11
GO:0031145	2e-029	5033	13	14	11
GO:0007091	2e-029	5033	13	14	11
GO:0030687	5e-029	5033	35	343	29
GO:0071004	6e-028	5033	29	345	26
GO:0045449	6e-028	5033	167	352	59
GO:0016455	1e-027	5033	23	352	23
GO:0006350	1e-026	5033	308	352	79
GO:0005847	2e-026	5033	15	107	15
GO:0016592	3e-026	5033	24	352	23
GO:0006406	5e-026	5033	53	632	40
GO:0006378	8e-026	5033	18	107	16
GO:0004298	1e-025	5033	14	92	14
GO:0000022	3e-025	5033	23	14	11
GO:0005762	7e-025	5033	36	345	27
GO:0005484	1e-024	5033	24	212	20
GO:0005763	1e-024	5033	28	356	24
GO:0000070	2e-023	5033	31	14	11
GO:0005666	7e-023	5033	18	160	16
GO:0006611	1e-022	5033	31	632	28
GO:0032040	1e-022	5033	41	343	27
GO:0005886	5e-021	5033	222	456	68
GO:0005685	1e-020	5033	17	345	17
GO:0070478	1e-020	5033	17	128	14

(0.1 with the selected parameters: 1000 nodes, 20000 links). We obtain similar results with a wide range of parameters, see Figure 7. This means that even if we can find partitions with a good modularity, algorithms cannot choose between these partitions.

These results can be explained by the fact that using low values of threshold, our algorithm finds a single core comprising all nodes and since there is



(a) (b)

Fig. 6. p_{ij} distribution for different $\mathcal{N}$ (a). Absence of cores in random graphs (b)

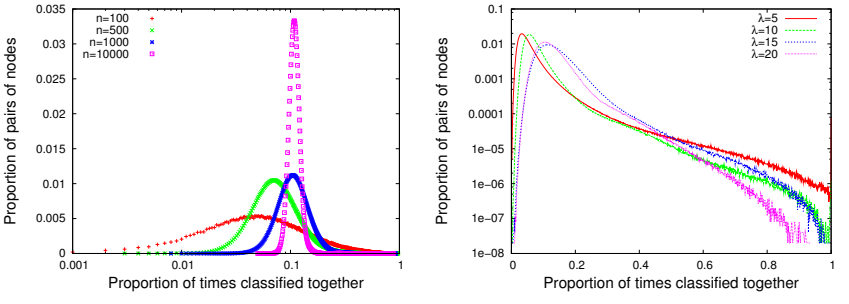


Fig. 7. Distribution of the p_{ij} averaged over 100 realizations, with $\mathcal{N} = 10^3$. Networks with different number of nodes n and an average degree of $\lambda = 20$ (left). Networks with $n = 1000$ nodes and different values of λ (right).

nearly no high values in random networks, there is no core with high values of the threshold, while real-world networks have high threshold cores (see Figure 6(b)). Interestingly, in random networks there is a sharp transition (as shown by the cusp at a threshold value around 0.3) between the situation where one single core is present and the intermediate threshold values where several cores are present, which is not present in real-world networks.

To further validate these results, we compared the cores of two real-world networks with random graphs that have the same size (and same degree distribution for the configuration model), see Figure 8. In the case of the Erdös-Rényi model, there is no pair of nodes with $p_{ij} = 0$, which means that all pairs of nodes have been grouped together at least once during 1000 execution of the Louvain algorithm. Conversely, there is nearly no pair of nodes which are always grouped together, but for the leaves (nodes of degree 1) of the network which are always grouped with their only neighbor.

All these results show that random and real-world network behave very differently from a core perspective, while both can exhibit a "classical" community structure, as measured by the modularity. This result gives a strong advantage to cores versus communities.

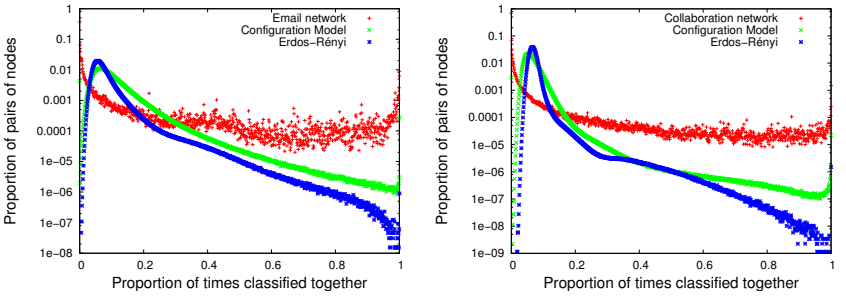


Fig. 8. p_{ij} distribution for two real-world networks together with Erdös-Rényi and configuration model random graphs with the same size

6 Conclusion

In this paper, we have investigated community structure of complex networks, using community cores which may improve the significance and the stability of groups of nodes detected by current community detection algorithms. We showed that community detection algorithms use heuristics methods which lead to different partitions of similar quality and there is no reason to prefer one above another. Furthermore, community detection algorithms are highly unstable and can find communities in graphs that have none.

If multiple executions of a non-deterministic community detection algorithm agree on certain sets of nodes, then these sets of nodes can be considered as more significant. We showed that cores have a hierarchical structure which can be obtained using different thresholds in our proposed algorithm. We applied our method to both artificial and real networks and showed the performance of our approach when comparing cores to Ground Truth. More particularly, in random networks we find an absence of cores for high enough values of the parameter α. This might provide a robust way to distinguish random networks from real-world networks.

The perspectives of our work are to find a meaningful way to select the threshold, even if the whole hierarchy can be useful as it gives a multi-scale view of the network, and to study the dynamical networks and the evolution of cores in such networks[1].

References

1. Aynaud, T., Guillaume, J.: Static community detection algorithms for evolving networks. In: Proceedings of the 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt), pp. 513–519. IEEE (2010)

[1] This work is supported in part by the French National Research Agency contract DynGraph ANR-10-JCJC-0202.

2. Blondel, V., Guillaume, J., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008, P10,008 (2008)
3. Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., Wagner, D.: On Finding Graph Clusterings with Maximum Modularity. In: Brandstädt, A., Kratsch, D., Müller, H. (eds.) WG 2007. LNCS, vol. 4769, pp. 121–132. Springer, Heidelberg (2007)
4. Erdös, P., Rényi, A.: On random graphs. i. Publicationes Mathematicae (Debrecen) 6, 290–297 (1959)
5. Gfeller, D., Chappelier, J., De Los Rios, P.: Finding instabilities in the community structure of complex networks. Physical Review E 72(5), 056, 135 (2005)
6. Girvan, M., Newman, M.: Community structure in social and biological networks. In: Proceedings of the National Academy of Sciences, vol. 99(12), p. 7821 (2002)
7. Guimera, R., Danon, L., Diaz-Guilera, A., Giralt, F., Arenas, A.: Self-similar community structure in a network of human interactions. Physical Review E 68(6), 065, 103 (2003)
8. Guimera, R., Sales-Pardo, M., Amaral, L.: Modularity from fluctuations in random graphs and complex networks. Physical Review E 70(2), 025, 101 (2004)
9. Karrer, B., Levina, E., Newman, M.: Robustness of community structure in networks. Physical Review E 77(4), 046, 119 (2008)
10. Lambiotte, R.: Multi-scale modularity in complex networks. In: 2010 Proceedings of the 8th International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt). IEEE (2010)
11. Lancichinetti, A.: Community detection algorithms: a comparative analysis. Physical Review E 80(5), 056, 117 (2009)
12. Molloy, M., Reed, B.: A critical point for random graphs with a given degree sequence. Random Structures & Algorithms 6(2-3), 161–180 (1995)
13. Newman, M.: The structure of scientific collaboration networks. In: Proceedings of the National Academy of Sciences, vol. 98(2), p. 404 (2001)
14. Newman, M., Girvan, M.: Finding and evaluating community structure in networks. Physical review E 69(2), 026, 113 (2004)
15. Pfitzner, D., Leibbrandt, R., Powers, D.: Characterization and evaluation of similarity measures for pairs of clusterings. Knowledge and Information Systems 19(3), 361–394 (2009)
16. Qinna, W., Fleury, E.: Detecting overlapping communities in graphs. In: European Conference on Complex Systems (ECCS 2009), Warwick Royaume-Uni. (2009), http://hal.inria.fr/inria-00398817/en/
17. Rosvall, M., Bergstrom, C.: Mapping change in large networks. PloS one 5(1), e8694 (2010)
18. Salwinski, L., Miller, C., Smith, A., Pettit, F., Bowie, J., Eisenberg, D.: The database of interacting proteins: 2004 update. Nucleic Acids Research 32(suppl. 1), D449–D451 (2004)
19. Strehl, A., Ghosh, J.: Cluster ensembles—a knowledge reuse framework for combining multiple partitions. The Journal of Machine Learning Research 3, 583–617 (2003)
20. Vinh, N., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: is a correction for chance necessary? In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 1073–1080. ACM (2009)
21. Zachary, W.: An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 452–473 (1977)

An Empirical Study of the Relation between Community Structure and Transitivity

Keziban Orman, Vincent Labatut, and Hocine Cherifi

Abstract. One of the most prominent properties in real-world networks is the presence of a community structure, i.e. dense and loosely interconnected groups of nodes called communities. In an attempt to better understand this concept, we study the relationship between the strength of the community structure and the network transitivity (or clustering coefficient). Although intuitively appealing, this analysis was not performed before. We adopt an approach based on random models to empirically study how one property varies depending on the other. It turns out the transitivity increases with the community structure strength, and is also affected by the distribution of the community sizes. Furthermore, increasing the transitivity also results in a stronger community structure. More surprisingly, if a very weak community structure causes almost zero transitivity, the opposite is not true and a network with a close to zero transitivity can still have a clearly defined community structure. Further analytical work is necessary to characterize the exact nature of the identified relationship.

1 Introduction

In a complex network, a *community* is a cohesive subset of nodes with denser inner links, relatively to the rest of the network [1]. The presence of such groups is a

Keziban Orman
Galatasaray Uni versity & University of Burgundy
e-mail: korman@gsu.edu.tr

Vincent Labatut
Galatasaray University
e-mail: vlabatut@gsu.edu.tr

Hocine Cherifi
University of Burgundy
e-mail: hocine.cherifi@u-bourgogne.fr

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 99–110.
springerlink.com

common feature in networks modeling different types of real-world systems, including biological, social, information or technological ones [2]. When a network takes the form of a set of interconnected communities, it is said to possess a *community structure*.

The presence of a community structure is presumably related to other topological properties of the network. Uncovering what causes a community structure to appear, and what its effects are, would be valuable for a better understanding of the complex networks structure and dynamics. In particular, it would allow improving or explaining the existing community detection methods, and provide tools to interpret the communities identified in real-world networks. This angle was adopted in a few studies, with different objectives and/or in different contexts.

Pastor-Satorras *et al*. [3] showed how the presence of a hierarchical community structure and a power law degree distribution are sufficient conditions to cause a high transitivity (also called clustering coefficient). For this matter, they defined a generative model implementing these properties and studied the obtained networks. Moreover, they derived a new use for the transitivity measure, by utilizing its distribution to characterize the network hierarchical structure. Clauset *et al*. [4] proposed a different hierarchical approach: they defined a parameterized hierarchical model which they fit to various real-world data. The obtained hierarchical structures possess various properties present in real-world networks, including being scale-free (power law distributed degree) and having a high transitivity. This seems to indicate the hierarchical structure alone is enough to get both a scale-free and highly transitive network. Lie & Hu [5] proposed a model able to generate networks with community structures of various strengths, and showed the transitivity of the resulting networks depend on this strength. They used their model to study the effect of community structure on the network epidemic threshold. Interestingly, the generated networks are neither scale-free nor have a hierarchical structure, which seems to indicate these are sufficient, but not necessary conditions. Wang and Qin [6] had the same objective, but used a different model. It is a mixture of Watts-Strogatz's small-world model [7] and Newman's community structure model [1]. It is therefore not hierarchical either, nor is it scale-free.

The previous studies intended at studying the effects of the community structure on some topological properties of interest. In the works by Jin *et. al* [8] and Boguñá *et. al* [9], the community structure is, on the contrary, a byproduct. The authors focused on social networks and designed their models as multi-agent systems mimicking social interaction. The generated networks turned out to possess some properties observed in real-world social networks, including hierarchical community structure and high transitivity. Interestingly, the degree is not power law-distributed in social networks, which seems to confirm the scale-free property is not a prerequisite to get highly transitive and/or community structured networks.

In this article, our goal is to study how transitivity and community structure can mutually affect each other in realistic networks. Contrarily to the first cited studies [3, 4], we consider non-hierarchical networks, since this property does not seem to be a necessary condition to the presence of a community structure. The obvious difference with studies [5, 6] is our focus on transitivity, which intuitively seems to be a good candidate to explain the presence of a community structure (cf.

section 3). Another important difference is our aim of evaluating not only the effect of the community structure on this property, like in [3-6], but also the effect of the property on the community structure. Finally, we are not interested in the specific process resulting in the network structure, like in [8, 9], but rather in the general relationships between community structure and transitivity.

To study this relationship, we adopt an empirical approach based on several generative models. First, we use an existing model to generate realistic networks possessing a community structure with a controlled strength, [10] and study its transitivity. Second, an existing model [11] and a new model of our own are used to generate networks with a high or controlled transitivity, and we study the strength of their community structure. The rest of the document is structured as follows. In the next two sections, we review the notion of community structure and justify our choice of the transitivity as a property of interest relatively to its study. Section 4 is dedicated to the description of our methods, and more particularly the models we are using. We then present the results of our simulation and discuss the nature of the uncovered relationships in sections 5.1 and 5.2 . Finally, we conclude by highlighting our contributions and the possible extensions of our work.

2 Community Structure

The concept of community can be formally defined in several ways: mutually exclusive vs. overlapping, hierarchical vs. flat, local vs. global, etc. [12]. The nature of the community structure directly depends on the considered definition of a community. Independently from this choice, stating the presence or absence of a community structure is itself an ambiguous task. For this matter, one can clearly distinguish two extreme cases: on the one hand, the complete absence of any community structure (e.g. a complete network, in which all nodes are connected to each others), and on the other hand a perfect community structure (a network made up of several disconnected components). Between these two extremes lies a continuum of networks exhibiting community structures of various strengths. It makes therefore more sense to measure this strength rather than the presence or absence of a community structure.

In this article we selected the *modularity* [1] for this matter. It is certainly the most widely spread measure to assess the strength of a community structure. It is based on the numbers of intra- and inter-community links, and consists in comparing the proportion of intra-community links present in the network of interest, to the expectation of the same quantity for a randomly generated network of similar size and degree distribution. It is worth noticing some limits have been identified since the creation of this measure [12]. The most important seems to be its resolution limit, causing it to fail identifying communities considered as small relatively to the network size and community interconnection pattern [13]. However, we considered it to be sufficient for this exploratory work.

Let us note e_{ij} the proportion of links connecting nodes in community i to nodes in community j. Then the proportion of intra-community links for the whole

network is $\sum_i e_{ii}$. Let us note $e_{i+} = e_{+i}$ the proportion of links connecting at least one node from community i. For the same community, Newman defines the expected number of inter-community links as e_{i+}^2, in a network whose links are distributed randomly. The modularity is therefore: $Q = \sum_i (e_{ii} - e_{i+}^2)$.

3 Transitivity

The *transitivity* (also called *clustering coefficient*) of a network is the relative proportion of triangles among all connected triads it contains [14]: $C = n_\Delta / n_\Lambda$ where n_Δ and n_Λ are the numbers of triangles and connected triads, respectively. A triangle is a set of three completely connected nodes, whereas a triad can be either a triangle, or a set of three nodes connected by only two links (instead of three). The transitivity can be interpreted as the probability of finding a direct connection between two nodes having a common neighbor. The measure therefore ranges from 0 to 1. Besides this global version, a local one exist, defined at the level of some node i [7]: $C_i = \frac{\delta_i}{k_i(k_i-1)/2}$, where k_i is the degree of i, and δ_i the number of triangles containing this node. The denominator corresponds to the number of combinations of two neighbors of i, in other words: the number of connected triads centered on i. The ratio can therefore be interpreted as the probability of finding a direct connection between two neighbors of i. The local transitivity can be averaged over the whole network to obtain a global measure. Real-world networks are characterized by a high transitivity, whatever the considered version [2].

Transitivity and community structure are frequently jointly observed in real-world networks. Let us consider for instance the comparative study conducted in [15]. The authors classify networks depending on the systems they model, and analyze their community structures. According to our processing, the transitivity values associated to these community-structured networks are significantly higher than for same-sized random networks, by several orders of magnitude and for all considered classes.

The relationship between transitivity and community structure may seem trivial at first. Intuitively, a high transitivity appears to be the natural consequence of a community structure: links are concentrated in communities and should therefore form many triangles. Reciprocally, it seems a high transitivity indicate the links are form clusters, and therefore communities. However, it is relatively easy to find counter-examples to refute these propositions. First, consider a network whose communities are fully connected multipartite networks: the community structure can be very strong, with dense communities, but the transitivity is nevertheless zero. One could alternatively consider communities taking the form of connected stars, for the same result. Second, consider a fully connected network: the transitivity is maximal, but there is no community structure (just a single community).

To avoid this kind of situation, we based our analysis on randomly generated networks with realistic properties. When possible, we selected generative models able to mimic the topological properties consensually considered to be present in real-world networks.

4 Methods

The empirical approach we adopted to study the relationship between community structure and transitivity is two-stepped. First, we generate artificial realistic networks with controllable community structure and analyze how changes in the community structure affect the transitivity. Second, we use two different models able to generate transitive networks, and analyze how changes in the transitivity affect the community structure. The identification of the community structures is performed by applying two different and complementary algorithms. In this section, we describe all three generative models, and summarize the principle of both community detection algorithms.

4.1 Community Structure Model

To generate networks possessing a community structure, we used a modified version of the LFR model [10]. This model applies a three-stepped generative process based on the use of a more basic model, i.e. one not supposed to produce a community structure. First, the basic model is used to generate an initial network. Second, virtual communities are randomly drawn so that their sizes follow a power law distribution. Third, an iterative process takes place to rewire certain links, in order to make the community structure appear while preserving the degree distribution of the initial network.

The strength of the community structure is controlled by a specific parameter called the *mixing coefficient* μ. This parameter allows us to produce networks with various community structure strengths and analyze how this affects the transitivity. The mixing coefficient represents the desired average proportion of links between a node and nodes located outside its community, called inter-community links. Consequently, the proportion of intra-community links is $1-\mu$.

By construction, the LFR model guaranties to obtain power law-distributed community sizes, which is a property present in community-structured real-world networks [10]. Since the degree distribution is preserved during the rewiring step of the generative process, the rest of the topological properties depend mainly on the basic model used at the first step. The original LFR process relies on the Configuration Model (CM) [16], which is able to produce networks with a specified degree distribution. In LFR, it was used to obtain a power law-distributed degree, also a well identified feature of many real-world networks [2]. To detect any potential effect the basic model could have on the transitivity measured in the final networks, we selected two alternatives to the CM, both able to produce scale-free networks too. Barabási–Albert's model (BA) [17] implements a completely different, more realistic, generative process based on preferential attachment. The Evolutionary Preferential Attachment model (EV) [18] is a variant of BA able to produce networks with a higher transitivity.

4.2 Transitive Models

We used two different models to study the effect of the transitivity on the community structure. We first selected a model by Newman [11] (NM), which could be considered as an adaptation of the CM able to produce networks with a controlled transitivity. Instead of specifying the degree k_i of each node i like in the CM, one has to define both the number of single links s_i and the number of distinct triangles t_i attached to the node. In other words, a distinction is made between the links depending on their belonging to a triangle. Both are mutually exclusive, meaning one link is either a single link or appears in only one triangle. In the end, the total degree is $k_i = s_i + 2t_i$. For our study, we wanted to obtain scale-free networks for matters of realism, and we therefore needed to control k_i. We consequently introduced in our implementation of NM a parameter called transitivity coefficient $\tau \in [0,1]$, in order to control the proportion of the degree dedicated to triangles (vs. single links). Let $[\ \ ]$ denote the round function, then we have $t_i = [\tau\, k_i/2]$ and $s_i = [(1-\tau)k_i]$.

The main advantage of NM is it allows artificially changing the transitivity of the generated networks. However, for our study, it also has an important limitation: the obtained transitive structure is not very realistic. Indeed, the created triangles are all distinct, i.e. they cannot share more than one node. Put differently, it is not possible for them to have a common side. This also limits the transitivity (both the global and local versions). The maximal local transitivity some node i can reach is $1/(k_i - 1)$ when $s_i = 0$.

In order to overcome this disadvantage, we developed our *Highly Transitive* model (HT). It is able to randomly generate networks with both a specified degree distribution and a high transitivity. The process starts with a ring network, in order to avoid isolated nodes or components in the final network. Links are then randomly added while respecting the desired degree distribution and favoring the connection of nodes with common neighbors (in order to increase the transitivity).

Our model allows obtaining networks whose transitivity is much higher than in NM networks. However, we are not able to control it with a parameter like we did for NM. Both models are therefore complementary: NM allows us to test for the effect of various level of transitivity, even if the maximal transitivity obtained is not very high (greater than in random networks though, so still realistically high). HT allows us to test for the effect of a very high transitivity on the community structure.

4.3 Community Detection

In the first part of our experiment, the community structure of the generated networks is known, because it is defined by construction. However, this is not the case in the second part, and we therefore need to identify it. For this purpose, we used two recent algorithms: Louvain [19] and Infomap [20].

Louvain (LV) is an optimization algorithm proposed by Blondel et *al.* [19]. It uses a two-stepped hierarchical agglomerative approach. During the first step, the algorithm performs a greedy optimization of the modularity (cf. section 2) to

identify small communities. During the second step, it builds a new network whose nodes are the communities found during the first step. In this new network, the intra-community links are represented by self-loops, whereas the inter-community links are aggregated and represented as links between the new nodes. The process is repeated on this new network, and stops when only one community remains.

Infomap (INP) is an algorithm developed by Rosvall and Bergstorm [20]. The task of finding the best community structure is expressed as a compression problem. The authors want to minimize the quantity of information needed to represent the path of some random walker traveling through the network. The community structure is represented through a two-part nomenclature based on Huffman coding: the first part is used to distinguish communities in the network and the second to distinguish nodes in a community. With a partition containing few inter-community links, the walker will probably stay longer inside communities, therefore only the second level will be needed to describe its path, leading to a compact representation. The authors optimize their criterion using simulated annealing.

As mentioned in section 2 , many different definitions of the concept of community exist. Louvain optimizes directly the modularity, whereas Infomap relies on a completely different definition of what a community is. The first is from far the most widely spread, and the second proved to be very efficient [21]. From this point of view, these two algorithms are complementary, which is why we selected them. This allows us to detect if the community structure induced by a high transitivity favors one definition or the other.

5 Results

5.1 Effects of Community Structure on Transitivity

By applying the LFR rewiring process to the three basic models (CM, BA and EV), we generated three different sets of community structured networks. We selected our parameters values based on previous experiments regarding artificial networks generation [10], and descriptions of real-world networks measurements from the literature [2, 22], so that the produced networks were the most realistic possible. Some parameters are common to all three processes: we fixed the size $n = 5000$ and the power law exponent for the community sizes distribution $\beta = 2$, and made the mixing coefficient μ range from 0.05 to 0.95 with a 0.05 step. Other parameters are model-dependent. CM allows a precise control of the degree, since it is possible to specify the desired power law exponent γ for the degree distribution, and the average $\langle k \rangle$ and maximal degrees k_{max}. We used the values $\gamma = 3$, $\langle k \rangle \in \{15{,}30\}$ and $k_{max} \in \{45{,}90\}$. Both other models do not let as much control, and we had to adjust their parameters so that the resulting networks had approximately the same degree-related properties. Preferential attachment does not give any control on γ, which tends towards 3 by construction. We produced 25 networks for each combination of parameters, and averaged the transitivity, as shown in Fig. 1 (left). Results were very similar for $\langle k \rangle = 15$ and 30, so we only present the latter here, but comments apply to both.

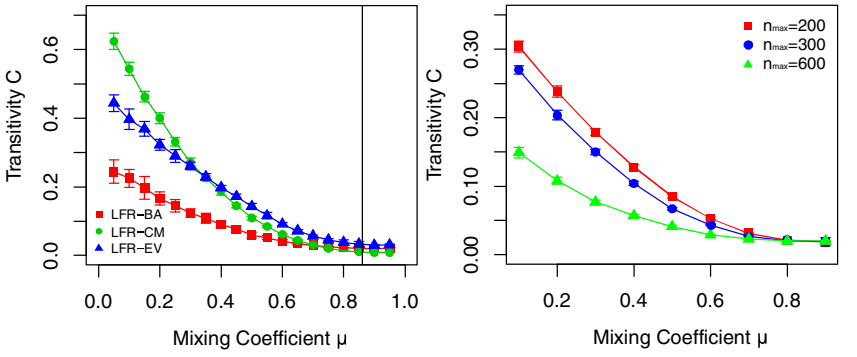


Fig. 1. Effect of the mixing coefficient μ on the transitivity. Each point corresponds to an average over 25 networks generated with $\gamma = 3$ and $\beta = 2$. Left: for each LFR variant, with $n = 5000$, $n_{max} = 700$ and $\langle k \rangle \approx 30$. Right: for several values of n_{max} on LFR-CM, with $n = 1000$ and $\langle k \rangle \approx 15$.

The transitivity of the networks generated by the basic models before rewiring are 0.020, 0.008 and 0.030 for CM, BA and EV, respectively. After rewiring, CM leads to the highest transitivity, with values around 0.6 for $\mu \approx 0$, but it reaches almost zero for $\mu \approx 1$, exhibiting a serious sensitiveness to changes in the community structure strength. Both other models also show a decreasing transitivity when μ increases, but the range is much smaller, partly because their values for $\mu \approx 0$ are significantly smaller: around 0.25 and 0.45 for BA and EV, respectively. Like CM, their transitivity is close to zero when $\mu \approx 1$. In the literature, real-world networks with a 0.3 transitivity are considered highly transitive [22], so we can state all three models exhibit a realistic transitivity for a small μ (clearly separated communities). The fact the transitivity decreases when the communities become more and more difficult to discern, for all three models, supports the assumption that a realistic community structure causes a high transitivity.

Besides its strength, a community structure can be characterized by its community size distribution. For realism matters, we chose a power law with fixed exponent $\beta = 2$, but the practical draw of the community sizes requires specifying the size of the largest community n_{max}. In order to study the effect of this limit on the transitivity, we generated another batch of networks with $n = 1000$, $\langle k \rangle = 15$ and $n_{max} \in \{200,300,600\}$, the other parameters being the same than before. Transitivity values for different largest community sizes are shown on Fig. 1 (right). When using a smaller n_{max}, the size difference between the smallest and the largest communities decreases, making the community size distribution more homogeneous (or rather: less contrasted, since it still follows a power law). It also affects the number of communities, which decreases when n_{max} increases: the numbers of communities are 40, 30 and 15 for $n_{max} = 200$, 300 and 600, respectively.

It turns out the transitivity measured on the obtained networks decreases when n_{max} increases. In other words, the number of triangles increases when there are less communities, with more similar sizes. This makes sense considering the links

constituting triangles have more chance to fall between communities when there are more of them, especially if they are smaller. This is confirmed by the fact the observed effect is stronger for clearly separated communities ($\mu \approx 1$).

5.2 Effects of Transitivity on Community Structure

We specified the parameter values for our HT model so that they were the most similar possible to what was used with LFR. We consequently generated 25 networks with $n = 5000$ and $\langle k \rangle \in \{5,15,30\}$, and a power law-distributed degree ($\gamma = 3$). We obtained an average transitivity of 0.5, 0.45 and 0.3 for $\langle k \rangle = 5$, 15 and 30, respectively. This is consistent with the values observed in real-world networks. Both community detection algorithms return modularities close to 0.90, 0.72 and 0.74, respectively, indicating a strong community structure. This observation support the assumption a high transitivity allows obtaining a community structure.

As mentioned before, on the one hand NM does not reach a very high density, but on the other hand it can control it through the transitivity coefficient τ, allowing to analyze how changes in this parameter affects the community structure. It is therefore complementary to our model. Because of the local transitivity limit mentioned in section 4.2 , we had to use different parameters (compared to HT) to obtain a relatively high transitivity. We generated 6 networks with $n = 1000$, $\langle k \rangle = 5$, 10, and made τ range from 0 to 1 with 0.1 steps. Although sparse, the generated networks are connected.

We first focus on the networks obtained for $\langle k \rangle = 5$. Fig. 2 (left) shows how the modularity obtained by both community detection algorithms varies in function of τ. They differ in the amplitude of the measured modularity, which is higher for Louvain than for Infomap. This might be due to the fact Louvain directly optimizes this criterion. However, and more importantly, the trend is the same for both algorithms: the detected community structures get stronger when the transitivity increases. This is particularly true when $\tau > 0.4$. It turns out below this value, the actual transitivity does not change very much ($C < 0.05$), as shown in Fig. 2 (right), certainly due to the rounding performed during the generative process (cf. section 4.2).

The highest modularity is obtained for $\tau = 1$, i.e. when most links are used to create triangles. However, because of the model characteristics, this does not necessarily translates into a very high transitivity value, as shown in Fig. 2 (right). More surprisingly, even the smallest modularity values (close to 0.5), obtained for $\tau = 0$, are still considered as large in the literature, and reveal a clear community structure. The networks generated for $\tau = 0$ are not supposed to contain many triangles (only those obtained by chance, i.e. a negligible number [11]), as confirmed by the measured transitivity ($C < 0.05$). This indicates a high transitivity is not a prerequisite to the existence of a strong community structure, at least when considering the definition implemented by the modularity measure.

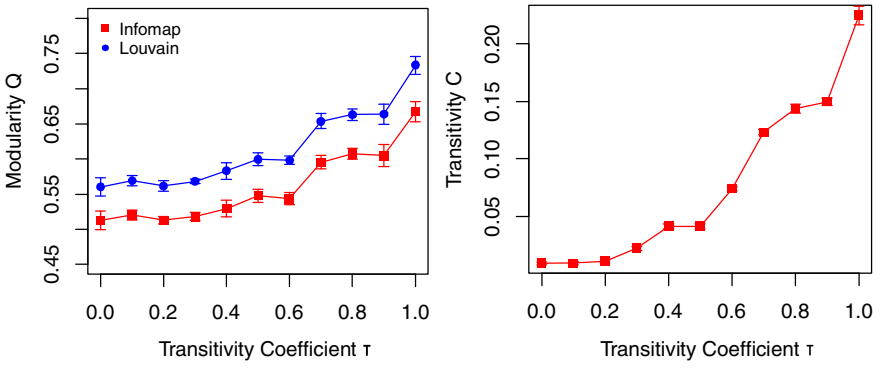


Fig. 2. Effect of the transitivity coefficient on the modularity. Each point corresponds to an average over 6 networks generated by NM with $n = 1000$ and $\langle k \rangle = 5$.

For the denser networks ($\langle k \rangle = 10$), the evolution of both the transitivity and modularity are similar to what was observed with $\langle k \rangle = 5$. However, as mentioned before, due to the local transitivity limit present in NM, the transitivity reaches a much lower value of only 0.12: this cannot be considered as high. The modularity is also lower, ranging from 0.33 to 0.41, however these values are still considered as relatively high, even those obtained when τ is close to zero. This confirms our previous remark regarding the coexistence of both a low transitivity and a significant community structure.

6 Conclusion

In this study, we took advantage of several generative models to investigate the relation between the community structure and the transitivity of complex networks. We first applied three variants of the LFR model [10] to generate artificial networks with known community structures. We observed similar results for all three variants: the rewiring process allowing the community structure to appear also causes a large increase in the transitivity. Moreover, the obtained transitivity is directly affected by the strength of the community structure and the distribution of the community sizes. So for this model, transitivity seems to be an offspring of community structure. Secondly, we used two models HT and NM [11] to generate transitive networks. The first, designed by us, produces a very high transitivity, but cannot control it. The transitivity is clearly lower with the second, but a specific parameter allows controlling it. Besides this point, the models are also complementary in the sense they produce networks with very different topologies. We used two state-of-the-art algorithms, Louvain [19] and Infomap [20], to identify the community structures in the generated networks. It turns out the strength of the modularity structure, expressed in terms of modularity, increases with the transitivity, for both generative models and according to both community detection algorithms. This also supports our point concerning the relationship between community structure and transitivity. More surprisingly, according to the obtained

modularity, the networks with almost zero transitivity also have a clear (although not as strong) community structure. For NM, it therefore seems the transitivity affects the community structure strength, but is not a prerequisite.

Our main contribution was to study the relationship between community structure and transitivity, which, although intuitively trivial, was not objectively analyzed before. For this purpose, we developed a new random generative model able to produce highly transitive networks with a desired degree distribution. We also modified the other models used in this article, in order to adapt them to our objectives. Our work can be extended in various ways. It would be possible to develop our model, in order to generate more realistic networks, and allow controlling the transitivity. We could also use alternative models, for the production of both community structure and controlled transitivity, in order to ensure our results are not model-dependent. The quality and nature of the community structures could be assessed in a deeper way, through various additional tools like community profile [23] or some alternative to the modularity [12]. There also are generalized versions of the transitivity, dealing with cycles of higher order. But a more important point would be to characterize the *nature* of the relationship between transitivity and community structure. Complementarily to our empirical study, an analytical work would allow identifying the necessary and/or sufficient conditions for the existence of a community structure. This might require to consider other topological properties, especially the network density and degree distribution.

References

1. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E 69 (2004)
2. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45, 167–256 (2003)
3. Pastor-Satorras, R., Rubi, M., Diaz-Guilera, A., Barabási, A.-L., Ravasz, E., Oltvai, Z.: Hierarchical Organization of Modularity in Complex Networks. In: Statistical Mechanics of Complex Networks, vol. 625, pp. 46–65. Springer, Heidelberg (2003)
4. Clauset, A., Moore, C., Newman, M.E.J.: Hierarchical structure and the prediction of missing links in networks. Nature 453, 98–101 (2008)
5. Liu, Z.H.: Bambi: Epidemic spreading in community networks. Europhysics Letters 72, 315–321 (2005)
6. Wang, G.-X., Qin, T.-G.: Impact of Community Structure on Network Efficiency and Communicability. In: 2010 International Conference on Intelligent Computation Technology and Automation (ICICTA), vol. 2, pp. 485–488 (2010)
7. Watts, D., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393, 409–410 (1998)
8. Jin, E.M., Girvan, M., Newman, M.E.J.: Structure of growing social networks. Phys. Rev. E 64, 046132 (2001)
9. Boguña, M., Pastor-Satorras, R., Diaz-Guilera, A., Arenas, A.: Models of social networks based on social distance attachment. Phys. Rev. E 70, 056122 (2004)
10. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Phys Rev E 78, 046110 (2008)
11. Newman, M.E.J.: Random Graphs with Clustering. Phys. Rev. Lett. 103 (2009)

12. Fortunato, S.: Community detection in graphs. Physics Reports 486, 75–174 (2010)
13. Fortunato, S., Barthelemy, M.: Resolution limit in community detection. PNAS USA 104, 36–41 (2007)
14. Luce, R.D., Perry, A.D.: A method of matrix analysis of group structure. Psychometrika 14, 95–116 (1949)
15. Lancichinetti, A., Kivelä, M., Saramäki, J., Fortunato, S.: Characterizing the Community Structure of Complex Networks. PLoS ONE 5, e11976 (2010)
16. Molloy, M., Reed, B.: A critical point for random graphs with a given degree sequence. Random Structures and Algorithms 6, 161–179 (1995)
17. Barabasi, A., Albert, R.: Emergence of scaling in random networks. Science 286, 509 (1999)
18. Poncela, J., Gomez-Gardeñes, J., Floría, L.M., Sanchez, A., Moreno, Y.: Complex Cooperative Networks from Evolutionary Preferential Attachment. PLoS ONE 3, e2449 (2008)
19. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. Stat. Mech. P10008 (2008)
20. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. PNAS 105, 1118 (2008)
21. Orman, G.K., Labatut, V., Cherifi, H.: Qualitative Comparison of Community Detection Algorithms. Communications in Computer and Information Science 167, 265–279 (2011)
22. da Fontoura Costa, L., Oliveira Jr., O.N., Travieso, G., Rodrigues, R.A., Villas Boas, P.R., Antiqueira, L., Viana, M.P., da Rocha, L.E.C.: Analyzing and Modeling Real-World Phenomena with Complex Networks: A Survey of Applications (2008), arXiv 0711.3199
23. Leskovec, J., Lang, K.J., Dasgupta, A., Mahoney, M.W.: Statistical Properties of Community Structure in Large Social and Information Networks. In: WWW. ACM, Beijing (2008)

Detecting Overlapping Communities in Complex Networks Using Swarm Intelligence for Multi-threaded Label Propagation

Bradley S. Rees and Keith B. Gallagher

Abstract. We propose a unique approach to finding overlapping communities within complex networks that leverages swarm intelligence, for decentralized multi-threading processing, with label propagation, for its fast identification of communities. The combination of the two technologies offers a high performance approach to overlapped community detection that allow for the processing of very large networks in tractable time.

Keywords: Community detection, complex networks, multi-agent system.

1 Introduction

Complex Networks are a popular way to capture the relationship between objects. Social networks are a key example, but other real-world networks come from biology, physics, and computer science, to name a few [12]. One property exhibited by complex networks is a tendency for nodes to collect in groups, or communities. The discovery of these communities helps provide insight into group formation and social structures.

Algorithms for community detection have garnered significant interest in the past decade [2], however problems still remain. Namely, that the development of a detection algorithm requires a rigorous mathematical definition of community and no such definition exists [4, 15]. Furthermore, many of the algorithms assume that communities are disjointed and limit nodes to a single community. Additionally, as the size of the networks increase, the runtime complexity of many algorithms becomes prohibitive. Performance and scalability is further limited when

Bradley S. Rees · Keith B. Gallagher
Department of Computer Science, Florida Institute of Technology, Melbourne, Florida, USA
e-mail: brees2011@my.fit.edu, kgallagher@fit.edu

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 111–119.
springerlink.com

algorithms' require a central control step that looks across the whole network. Divisive hierarchical algorithm, based on edge-between centrality [3] for example, requires that a metric score be computed for each edge and then the edge with the highest score selected for removal.

In this paper we present a different approach to community detection that allows individual nodes' to negotiate with its neighbors to determine community membership, including multiple overlapping memberships. Additionally, we use *Swarm Intelligence* as a means to remove the need for a central control processes while providing the ability to multi-thread the algorithm.

Due to page limitations, we are not presenting a review of related works in this paper. Instead we point interested readers to: Fortunato's [2] extensive survey of community detection algorithms; Xie, Kelley, and Szymanski's [19] survey of overlapping community detection algorithms; for label propagation, work by Raghavan, Albert, and Kumara [16] and also by Gregory's [4]; lastly, Ant Colony Optimized swarm work by Liu et al. [10] and by Leung, Kothari, and Minai [9].

2 Our Algorithm

The term *Swarm* defines a multi-agent system [1] where the behavior is common across all agents and the end goal of the algorithm is an emergent behavior. In our case, the detection of overlapping communities is an emergent behavior and not the goal of the agents. Additionally, each agent is reactive, able to respond to its environment, and social, able to communicate with other agents [18]. For the rest of this paper we will use the term agent, node, and vertex interchangeably.

2.1 Processing Sequence

The algorithm executes in four distinct sequential stages, with each agent keeping track of which stage it is in, and determines when it should move to the next stage. The stages are:

1. Initialize each agent
2. Building egonets and identify friendship-groups
3. Find non-propagating nodes within each friendship-group
4. Propagate

***Initialization*:** At start-up, each agent is assigned a unique integer identifier (ID). The ordering of the agents is not important, only that each is uniquely identified.

***Egonets and Detecting Friendship-Groups*:** an agent can only see and communicate with its neighbors. Therefore, its view of a community, or multiple communities, is limited to just those parts visible within an egonet. From the viewpoint of an agent, communities appear as disjoint components, which we term friendship-groups. We define a friendship-group [17] to be the local view of communities within an egonet from the perspective of the ego node, adhering to the principle

that communities have multiple paths [11]. A distinction is made between communities and friendship-groups since a friendship-group can be only a small portion of a community, and one or more friendships-groups can be combined to form a community.

For example, given the simple network in Figure 1(a), the egonet for vertex D is just those vertices connected to D and any edges between the neighbors of D, as highlighted in Figure 1(b). From the viewpoint of vertex D, B and C are friends and E and F are friends. The removal of D, grayed out in Figure 1(c), creates the disjointed view of the two components. The friendship-groups are then the sets {B, C, D} and {D, E, F}. This was a simple example; in real networks a friendship-group can range in size from a minimum of 3 nodes to a maximum of *(n -1)* nodes.

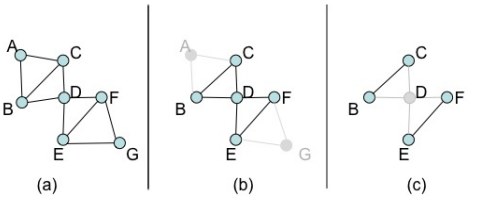


Fig. 1. Friendship-Groups

Each friendship-group detected is assigned a unique ID (label) based off the ID of the agent. For this work we append a decimal value to the agents integer ID value.

***Finding Non-Propagating Nodes*:** Not all members of a friendship-group are treated equally. Neighbors can have a different view of the friendship-group. This difference in view is what allows us to reconstruct the full community, additionally it allows for determining where potential overlap occurs. The process consists of each agent asking each of its neighbors for their view of the friendship-group and then comparing the two views. If the sets are outside an acceptable threshold, then information being propagated from that neighbor is not further propagated, and the agent flagged as "non-propogating". A discussion of similarity thresholds is presented in section 2.2.

Label Propagating: labels are not propagated to all neighbors as typically done. Instead, labels are only propagated within friendship-groups. The propagation of labels (friendship-group IDs) needs to be presented from the perspective of the agent publishing the ID and from the view of an agent being informed of the ID value.

***Propagating Agent*:** if any of the friendship-groups belonging to this agent have their IDs changed - initially setting the value counts as a change - then notify each agent within that friendship-group of the new ID.

***Neighboring Agent*:** if the passed in friendship-group ID value is lower than the currently value, then the friendship-group value is updated. The indication of whether or not that update is considered a change is dependent on the status of the calling agent. If the calling agent is considered non-propagating, then the update in value is not considered as a change; hence the value is not further propagated. Otherwise the friendship-group value is marked as being changed and the value further propagated.

The process is repeated until all agents reach a steady state and no longer need to propagate information: e.g. the assigned values on the friendship-groups do not change.

Obtaining the Communities: Once propagation has stabilizes, communities can be determined simply by asking each agent for a list of its assigned communities. Communities will be all agents that share a common ID value.

2.2 Options for Determining Non-propagation Agents

Without the inclusion of additional information, the selection of non-propagating nodes is the only way to influence community boundaries. As stated, selection is based on the similarity between friendship-groups. That comparison is performed using the Overlap Coefficient, a variant of the Jaccard Index that is better suited for comparing sets of different sizes. Overlap Coefficient returns a value between 0.0 and 1.0, with a score of 1 indicating that the smaller set is a proper subset of the larger. The problem with using the Overlap Coefficient is that the threshold value needs to be selected *a priory*. Setting the threshold to 1.0 causes a large number of very small overlapping communities to be detected. While, setting the value of 0.0 causes the detection of a few large communities.

$$\text{Overlap Coefficient } O_c = \frac{|X \cap Y|}{\min(|X|,|Y|)}$$

The appropriate threshold between the two extremes is unknown. For this work we have selected to use two similarity settings: the first is a very restrictive measure where the sets can only differ by a maximum of one node (off-by-one); the second allow the sets to differ by up to 25% (75%-overlap).

2.3 Complexity

Runtime complexity of our algorithm is tied to the density of the network. The reason being that an egonet needs to evaluate for each node and the size of the egonet is tied to the average degree of the network. For this work, we use the estimate that the average degree in a spare network is *log n* [5]. Complexity for each stage is:

1. Initialization involves setting a unique value on each agent and specifying neighbors: *O(n log n)*.

2. Finding friendship-groups requires: (a) building an egonet, *O(log n)*; and (b) finding friendship-groups, which is done using union-find O(*log n)*. Since the process is done for each node, complexity is *O(n log n)*.
3. Finding non-propagating nodes involves each agent asking its neighbors, *log n*, for a set and doing set intersection, *log n*. Complexity: $O(n \log^2 n)$.
4. Label propagation involves pushing a label to neighbors: *O(n log n)*.

Complexity reduces to just $O(n \log^2 n)$ on sparse networks. Since the algorithm can be multi-threaded, we define time complexity as: $O((n \log^2 n) / T)$.

3 Application

3.1 Zachary

The Zachary karate club dataset [20] is a widely studied real-world network representing club members that interacted. By chance, a dispute broke out between two members that caused the club to splinter into two smaller groups.

Figure 2 shows the network with the splintering indicated by the dashed line. Our algorithm detected four communities within the dataset using the *off-by-one* threshold, as illustrated by the diamonds, circles, squares, and hexagons.

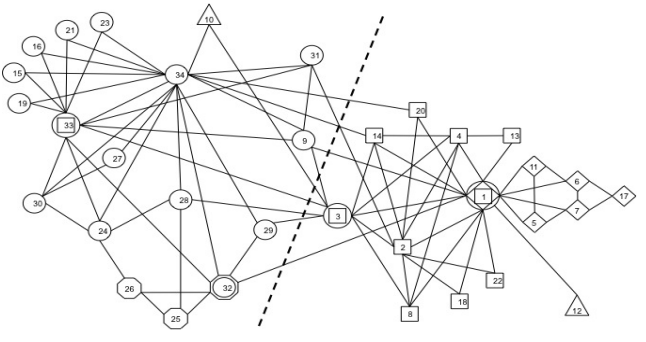


Fig. 2. Zachary Karate Club Dataset

Community A: {1, 5, 6, 7, 11, 17} - diamonds
Community B: {1, 2, 3, 4, 8, 9, 13, 14, 18, 22, 20, 33} - squares
Community C: {1, 3, 9, 15, 16, 19, 21, 23, 24, 27, 28, 29, 30, 31, 32, 33, 34} - circles
Community D: {25, 26, 32} - hexagons
Not a member of a community: 10, 12 - triangles

The detection of four communities, versus two, might seem to contrast the findings of Zachary. However, the Zachary paper addressed group fission and not community detection. Detection of more than two communities has been addressed by Zhang et al. [21], who used a variant of k-means clustering to uncover three communities. Since Zhang's algorithm does not including overlapping

communities, a true comparison cannot be performed. Nevertheless, we can do a comparison with special treatment given to overlapping nodes. Ignore overlap, we match on Community A. Community B is close, except we identified node 12 as not belonging to any community rather than community B. Lastly, the Zhang algorithm merged Communities C and D into a single set.

Looking at overlapping community results from the CFinder algorithm, Palla et al. [14], that algorithm identifies three communities. We achieved 100% match using the 25% similarity setting (Overlap Coefficient set to 0.75).

3.2 Other Datasets

A number of other well-studied datasets were run through our algorithm and results are shown in Table 1. Results are for using the off-by-one threshold setting and single threaded execution.

3.3 Generated Datasets

We used the LFR network generator[1] [6] to create a number of artificial networks for both performance and scale testing, and for evaluation against the Girvan and Newman (GN) benchmark [3]. For performance testing, we set the average degree to 7, the max degree to 100, and the mixing ratio to 10%, allowing the LFR algorithm determined the number of edges. The following two tables show runtime performance on networks from 1,000 nodes to 50,000 nodes, Table 2, and on 100,000 to 400,000[2] node, Table 3.

Against the GN benchmark, Figure 3, we were not expecting our algorithm to have reasonable performance since the benchmark does not consider overlapping communities and is biased towards producing centrality based graphs. The results, however, are in-line with what was achieved by the Cfinder algorithm [8].

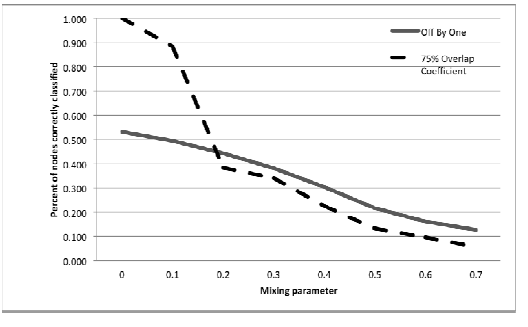


Fig. 3. Results of our algorithm against the GN Benchmark

[1] http://sites.google.com/site/santofortunato/inthepress2
[2] Due to hardware limitations, 400,000 nodes was the max we could test.

Table 1. Additional Datasets

Dataset	Nodes	Edges	Avg. Degree	Density	Communities Detected	Runtime (seconds)
Dolphins (a)	62	159	5.1	0.08	5	0.04
Zachary [20]	34	78	4.6	0.14	4	0.02
Football [3]	115	613	10.7	.094	18	0.11
Jazz (b)	198	2,742	27.69	0.14	2	0.62
Email (c)	1,133	5,452	9.6	0.01	40	0.83
PGP (d)	10,680	24,316	4.55	0.00043	791	2.37
Cond-mat-2003 [13]	31,163	120,029	7.70	0.00025	4065	8.24
Cond-mat-2005 (e)	40,421	175,693	8.69	0.00022	4794	14.23

(a) D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations, *Behavioral Ecology and Sociobiology* 54, 396-405 (2003).

(b) Gleiser, P. M. and Danon, L. (2003). Community structure in jazz. *Advances in Complex Systems* (ACS), 6(04):565–573

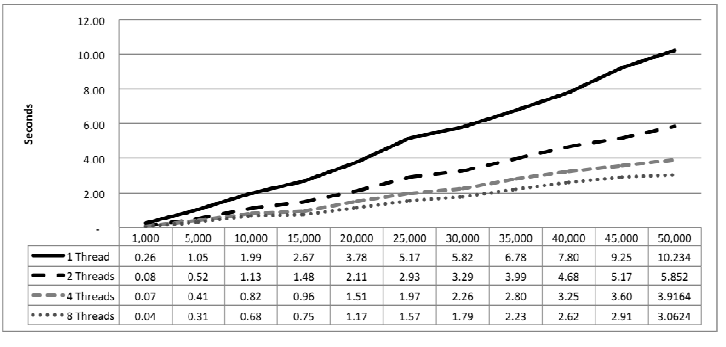
(c) Guimera, L. Danon, A. Diaz-Guilera, F. Giralt and A. Arenas, *Physical Review E*, vol. 68, 065103(R), (2003).

(d) M. Boguña, R. Pastor-Satorras, A. Diaz-Guilera and A. Arenas, Models of social networks based on social distance attachment. *Physical Review E*, vol. 70, 056122 (2004)

(e) Newman, M. E. J. The structure of scientific collaboration networks, *Proc. Natl. Acad. Sci.* USA 98, 404-409 (2001).

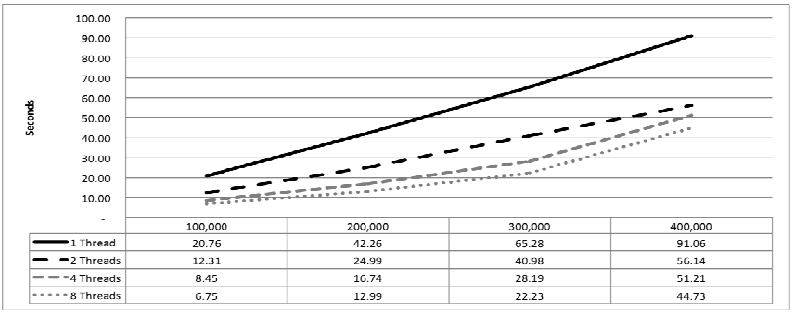
Datasets from `http://deim.urv.cat/~Eaarenas/data/welcome.htm`

Table 2. Runtimes on Artificial Networks from 1,000 to 50,000 Nodes



	1,000	5,000	10,000	15,000	20,000	25,000	30,000	35,000	40,000	45,000	50,000
1 Thread	0.26	1.05	1.99	2.67	3.78	5.17	5.82	6.78	7.80	9.25	10.234
2 Threads	0.08	0.52	1.13	1.48	2.11	2.93	3.29	3.99	4.68	5.17	5.852
4 Threads	0.07	0.41	0.82	0.96	1.51	1.97	2.26	2.80	3.25	3.60	3.9164
8 Threads	0.04	0.31	0.68	0.75	1.17	1.57	1.79	2.23	2.62	2.91	3.0624

Table 3. Runtimes on Artificial Networks from 100,000 to 400,000 Nodes



	100,000	200,000	300,000	400,000
1 Thread	20.76	42.26	65.28	91.06
2 Threads	12.31	24.99	40.98	56.14
4 Threads	8.45	16.74	28.19	51.21
8 Threads	6.75	12.99	22.23	44.73

4 Conclusion

The detection of communities within complex networks is a challenging problem [7, 8], made even more difficult when nodes are allowed to exist in multiple overlapping communities. In this work we presented a unique approach to the problem that uses each individual's view of the communities, allowing the network to be analyzed in pieces. The aggregation of individual perspective engenders an algorithm that shifts away from the traditional requirement of viewing the graph as a whole. That paradigm change allowed us to leverage swarm intelligence to further remove the need for a central control mechanism. The combination of those two items allows for overlapping communities to be detected with good performance, $O(n \log^2 n)$, while being scalable to large data set sizes.

References

1. Ferber, J.: Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence, 1st edn. Addison- Wesley Longman Publishing Co., Inc., Boston (1999)
2. Fortunato, S.: Community detection in graphs. Physics Reports 486, 3-5, 75–174 (2010)
3. Girvan, M., Newman, M.E.: Community structure in social and biological networks. In: Proceedings of the National Academy of Science, vol. 99, pp. 7821–7826 (June 12, 2002)
4. Gregory, S.: Finding overlapping communities in networks by label propagation. New Journal of Physics 12, 10,103018 (2010)
5. Hwang, W., Kim, T., Ramanathan, M., Zhang, A.: Bridging centrality: graph min-ing from element level to group level. In: KDD 2008: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 336–344. ACM (2008)
6. Lancichinetti, A., Fortunato, S., Radicchi, F.: Benchmark graphs for testing community detection algorithms. Physical Review E 78,4, 046110 (2008)
7. Lancichinetti, A., Fortunato, S.: Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. Physical Review E 80(1), 016118 (2009)
8. Lancichinetti, A., Fortunato, S.: Community detection algorithms: A comparative analysis. Physical Review E 80(5), 56117 (2009)
9. Leung, H., Kothari, R., Minai, A.A.: Phase transition in a swarm algorithm for self-organized construction. Physical Review E 68(4), 046111 (2003)
10. Liu, Y., Wang, Q., Wang, Q., Yao, Q., Liu, Y.: Email Community Detection Using Artificial Ant Colony Clustering. In: Chang, K.C.-C., Wang, W., Chen, L., Ellis, C.A., Hsu, C.-H., Tsoi, A.C., Wang, H. (eds.) APWeb/WAIM 2007. LNCS, vol. 4537, pp. 287–298. Springer, Heidelberg (2007)
11. Moody, J., White, D.R.: Structural cohesion and embeddedness: A hierarchical concept of social groups. American Sociological Review 68(1), 103–127 (2003)
12. Newman, M.E., Girvan, M.: Finding and evaluating community structure in net-works. Physical Review E 69(12), 026113 (2003)
13. Newman, M.E.: Fast algorithm for detecting community structure in networks. Physical Review E 69(6), 066133 (2004)

14. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the overlapping commu-nity structure of complex networks in nature and society. Nature 435, 814 (2005)
15. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. In: Proceedings of the National Academy of Sciences of the United States of America, vol. 101(9), pp. 2658–2663 (2004)
16. Raghavan, U.N., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Physical Review E 76, 036106 (2007)
17. Rees, B.S., Gallagher, K.B.: Overlapping community detection by collective friendship group inference. In: International Conference on Advances in Social Network Analysis and Mining, pp. 375–379 (2010)
18. Weiss, G.: Multiagent systems: a modern approach to distributed artificial intelligence. MIT Press, Cambridge (1999)
19. Xie, J., Kelley, S., Szymanski, B.K.: Overlapping community detection in networks: the state of the art and comparative study. CoRR, abs/1110.5813 (2011)
20. Zachary, W.: An information flow model for conflict and fission in small groups. Journal of Anthropological Research 33, 452–473 (1977)
21. Zhang, S., Wang, R., Zhang, X.: Identification of overlapping community structure in complex networks using fuzzy cc-means clustering. Physica A: Statistical Mechanics and its Applications 374, 483–490 (2007)

A Genetic Algorithm to Partition Weighted Planar Graphs in Which the Weight of Nodes Follows a Power Law

Rodrigo Palheta and Vasco Furtado

Abstract. This research makes use of evidence that the distribution of crime by census tracts in large cities follows a Power Law. This means that there are few places that concentrate many crimes and many places that concentrate few crimes. In this article we investigate how modeling complex networks and genetic algorithms can help to understand the behavior of samples representing views of part of the map of crimes of a large metropolis. The representation of the network is a planar graph where the nodes are the centroids of census tracts, the edges represent the adjacency between the tracts, and each node has a weight representing the number of crimes recorded in the census tract. The problem of this research lies in the context of the study of sampling distributions that have long tails (e.g. the weight of the nodes of the graph follows a Power Law). In particular, we describe a genetic algorithm to explore the space of possible samples of the initial distribution (plotted crimes throughout the city) so that the maximum number of samples holds features to follow a Power Law with an exponent close to the original distribution.

1 Introduction

This research is based around a system for collaborative crime mapping, called WikiCrimes (Furtado et al. 2010) (www.wikicrimes.org). WikiCrimes subverts the traditional logic of the handling of information about crimes that have occurred, because it allows citizens to build their own map of crime. Wikis, in general, and WikiCrimes in particular, are based on the concept of radical trust, i.e., it is believed that individual participation mostly includes correct information. Nevertheless, the identification of fraud or attempted vandalism is necessary. The challenge

Rodrigo Palheta · Vasco Furtado

University of Fortaleza, Av. Washington Soares 1321, Edson Queiroz, Fortaleza, Ceará, Brazil,
e-mail: rodrigo.palheta@gmail.com, vasco@unifor.br

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 121–130.
springerlink.com

imposed on WikiCrimes and collaborative maps in general is to ensure the credibility of the information recorded on the map.

This challenge is the main motivation that brings us to consider the use of information that can be modeled as a complex network. The distribution of crime by census tract is one type of such information. Previous studies (Melo 2008) (Cançado 2005) show that this distribution follows a Power Law. In this context, there are few places that concentrate many crimes and many places that concentrate few crimes.

This finding is very useful because it can support the identification of a malicious activity by identifying abnormalities in the original Power Law distribution. However, the problem in doing this is that the most prejudicial and difficult malicious activity to identify in this kind of maps is specific to a local area (in short, a local geographical trend). These localized activities typically do not affect the original distribution (e.g. the crime distribution for an entire city).

The perception of the trends on a digital map is typically made via kernel density algorithms, which identify areas that have a high concentration of crimes. Kernel Density Estimation or kernel smoothing is a statistical method for determining the density of crimes at different locations. The method is used to produce a continuous event density surface from crime point data. Kernel smoothing results in a continuous 'heat map' that shows geographic variation in the density or intensity of the crimes, also called hot spots (McLafferty 2000).

Note that a hot spot computed via kernel density methods depends on the number of crimes being analyzed, which in turn depends on the geographical area and the period to be considered. Whenever a user is viewing a digital map, the screen is the limitation for perceiving the events on the map. The manner the user has to vary the view of a geographical area is by regulating the zoom level. Therefore, the identification of hot spots must be done for different zoom levels, and for each one of these levels, the configuration of hot spots can be different.

More broadly, the problem of this research requires an understanding of sample analysis of distributions that are stable and have a long tail (Pickering et al. 1995), (Gatterbauer 2011). As in the present context, in order to identify malicious activity, we need to explore a set of data at different zoom levels; this equates to exploring a network and then, successive divisions thereof. These subdivisions or subnets are actually samples of the larger network.

In particular, in this article we describe and evaluate a genetic algorithm to explore the space of possible samples of the initial distribution (plotted crimes throughout the city) to optimize the samples, in the sense that they follow a Power Law with a similar exponent to the original distribution.

2 Related Work

2.1 Sampling of Power Law Distributions

Works on complex networks have identified properties of samples of the network, a very current theme and explored in various fields. Understanding the impact of

measuring a sample rather than a complete distribution attracts the attention of researchers who intend to characterize the network of links that form the Web, for example. It is known that, due to the size and dynamism of the Web, any analysis of the properties of the network of links and pages can only be done by sampling. The impact of how samples are collected is investigated in (Clauset and Moore 2005). In this work it is shown that the strategy of sampling used by crawlers, called traceroute, tends to underestimate the exponent of the Power Law of the degree distribution of the graph representing the links on the web. This is particularly true when the sample has more edges than vertices. Another relevant work in this context is (Pickering et al. 1995), which examines the impact of deficiencies in sampling and collecting data on earthquakes in the magnitude distribution of earthquakes in regions where they occur. A categorization of the types of sampling problems is performed, as well as an analysis of the impact of the difference in scale between the original distribution and the samples. Although it is known that the samples of random networks (Erdos and Renyi 1959) retain the properties of the initial distribution, the same does not seem to occur in networks where the distribution of degree of connectivity of the vertices is free of scale. Similar results were identified (Stumpf 2005), which showed that sub-networks of scale-free networks (scale-free nets) do not maintain the same properties as the original distribution.

2.2 Genetic Algorithm

The main steps of a basic genetic algorithm are shown in Fig. 1. The basic elements are the encoding of the solution (chromosome), the operators (mutation, crossover and selection), initial population, and fitness function.

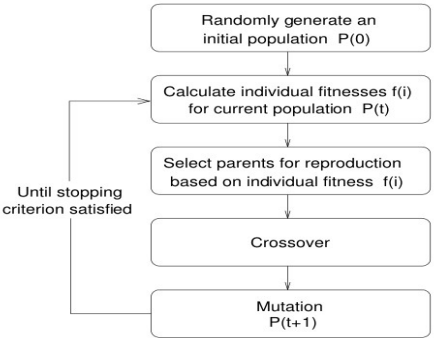


Fig. 1. Basics steps of a genetic algorithm according to (Cole 1998)

Several works in evolutionary computing have sought to define methods for partitioning graphs and discovering communities using genetic algorithms as in (Tavares-Pereira et al. 2007), (Datta et al. 2008) and (Semaan et al. 2009). The evolutionary approach allows one to create specific algorithms for any problem by modifying the representation of the solution or operators. In (Datta et al. 2008) several uses are cited, such as circuit design, gene ontology, simulation and other

products. In particular, it is worth mentioning (Tavares-Pereira et al. 2007) and (Semaan et al. 2009), where genetic algorithms are applied to *district problems*, in which a planar graph representing a geographic area is divided into districts that maintain similarity between them. Although several constraints have been considered in this process of planar graph partitioning in the literature, none of them refers to keeping the same distribution of the weight of nodes as the original distribution.

Typically the encoding for representing partitions is based on group-number (Cole 1998) in which each node receives a label indicating the partition to which it belongs. However, this representation for graph partitioning problems suffers to keep the constraint of connectivity. (Semaan et al. 2009) proposed an alternative solution for this, by using a Spanning Minimum Tree (SMT) generated from the original graph. From the SMT, an individual is created from the adjacent nodes of two nodes selected at random. The idea is to walk around the path from these nodes by adding the nodes to the partitions (until the maximum number is reached). The size of the partition is driven by a restriction defined *a priori*. Fig. 2(a) shows an SMT in which nodes A and B were randomly selected, and the maximum number of nodes per partition is six. Starting from A, an initial partition P1 is created. The nodes connected to A and that are not in the path to B are added to P1 (just A is added). Then C is visited and Y and W are added to P1. Z and Y are also added and the maximum number is reached. The process goes on, with Z and V being added to another partition (P2) and so on. The final partitions are shown in Fig. 2(b).

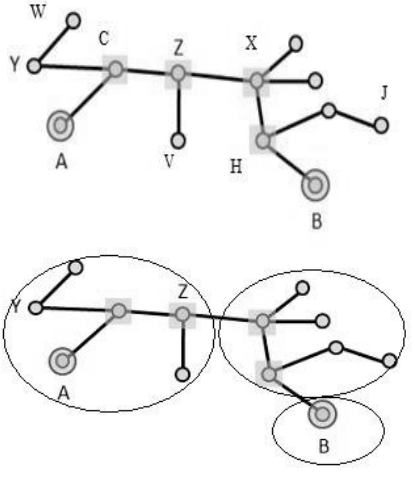


Fig. 2. An SMT and partitions generated from the methods of (Seeman et al. 2009)

The purpose of a crossover operator is to generate offspring by exploiting a search space, where some beneficial portion between two chromosomes is exchanged. Typically this is done randomly, but (Datta et al. 2008) has proposed a crossover operator that generates a new chromosome by inserting a random zone from one chromosome into another chromosome. It also takes care of any overlapping during this insertion by redefining the partially overlapped zones, as well

as other zones, if required. The mutation is developed by altering the sizes of various zones. The operator shifts a random boundary node of a zone to one of its adjacent zones, thereby reducing the size of the first zone and increasing that of the second zone.

3 A Genetic Algorithm for Partitioning a Weighted Planar Graph

Here we are going to describe the genetic algorithm (GA) capable of partitioning a weighted planar graph. As we work with graphs in which the distribution of the node weights follow a Power Law, the algorithm tries to partition the initial graph in a way such that the weight distribution of the nodes for each partition also follows a Power Law with a similar exponent. In addition, other restrictions also must be obeyed during the creation of individuals of the initial population and at the end of a mutation or crossover. There are three restrictions for checking integrity (each partition must have nodes that belong to only one partition), connectivity (the nodes of a partition must be connected), and minimum and maximum number of nodes per partition.

The proposed evolutionary algorithm is based on the canonical structure described in Section 2.2, but has modifications in the solution representation, crossover operator, and fitness function. Below we describe these modified elements.

3.1 Solution Representation

The representation of the solution is the following: an individual $I = \{ y_1, y_2, \ldots, y_k \}$ represents a set of partitions, y_i, where each partition is an undirected graph G=(V,E). In Fig. 3 it is possible to see a schema of a graph G and two possible partitions y_1 and y_2. Nodes represent a geographic area (e.g. census tract) with an associated weight (reported crimes), edges connect the regions, and the partitions are subgraphs of G. A chromosome of an individual is $I=\{y_1,y_2,y_3,..,y_k\}$, where *y* is a subgraph (or partition) of G. Each *y* is represented by a pair of sets. The first set represents the nodes of a partition. The second set represents the set of edges, which in turn is also represented by a pair describing the nodes to which the edge connects. The chromosome representation of y_1, for example, is as follows:

y1={{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16},{(1,3),(3,2),(3,4),(3,11),(11,16),(11,4),(11,5),(11,10),(5,7),(10,8),(10,9),(10,12),(9,14),(14,15),(14,13)} }

One can also see, in Fig. 3, that the nodes have a weight and that the distribution of the frequency thereof follows a Power Law.

The procedure to generate an initial population is inspired by (Semaan et al. 2009), but we had to adapt the original idea to cope with the fact that we would also like to guarantee the restriction of minimum and maximum number of nodes per partition. Therefore, during the process of generating the initial population, we did not generate individuals that do not satisfy these constraints.

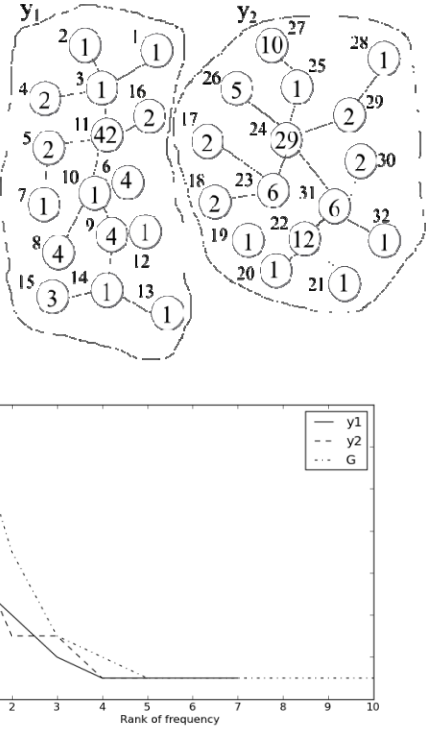


Fig. 3. A possible solution is represented by a partition of the original graph in two subgraphs that follows a PL

3.2 Fitness Function: Taking into Account the Power Law Distribution of the Weight of the Nodes

The main feature of our proposal is that it maximizes the number of partitions in which the distribution of the weight of the nodes follows a Power Law with an exponent close to the exponent of the Power Law of the initial graph (*slopeGlobal*). A distribution follows a Power Law when:

$$P(k) \approx k^{-slope} \tag{1}$$

Where ***k*** represents the weight of the node. Two functions have been defined to guide the algorithm in the exploitation of the space of possible partitions. The first one is F_s, as follows:

$$F_s = \frac{\sum_{i=1}^{partitions} (slopePartition_i - slopeGlobal)^2}{partitions} \tag{2}$$

Where partitions represents the number of partitions of each individual, *slopePartition* is the slope of the distribution of weights of each partition *i*, and *slopeGlobal* is the slope of the distribution of the weights of the whole graph. F_s tends to zero

when all of the slopes of the distribution of weights of the partitions are close to the slope of the Power Law of the distribution of the initial graph (*slopeGlobal*).

The fitness function, F_f, is the inverse of F_s multiplied by the ratio between the number of partitions (goodpartitions) in which *slopePartition* is greater or less than a predefined threshold (typically 10%) and the number of partitions. The greater F_f, the better the individual is.

$$F_f = (1/Fs) \times (goodpartitions/partitions) \quad (3)$$

3.3 Mutation and Crossover Operators

Mutation consists basically of distributing a node into a partition. The crossover operator we have defined is inspired by (Datta et al. 2008), but we take into account knowledge about how strong the variability of the partitions is after the application of the operators. Given two individuals, we search for partitions in one individual that are already present in the other individual. We then keep this partition, and all the other nodes that exceed it are spread out to other partitions. The algorithm used for the crossover is shown below:

Crossover

Input: *I1*, *I2*

Output: New Individual

1. Select randomly a partition of *I1*
2. Say *Ix* is I2's partition that *I1* is inserted into.
3. Insert *Ix*'s nodes that are not in *I1* into *Ix*'s adjacent partitions.
4. Return New Individual if the other constraints are satisfied.

The proposed algorithm will get the initial population and iterate until is reaches the stop condition (maximum number of generations defined by the user). For each iteration, the GA uses the selection wheel to choose the two individuals that will be used to generate a new individual by crossover, and eventually this new individual will be modified in mutation.

4 Results

We evaluated our approach with one dataset synthetically generated using (Devroye 1986). The dataset has an exponent equal to 1.48. The size of the distribution is 2194. This is the same as the number of census tracts of the city of Fortaleza, Brazil. To each node we randomly attributed a value of the distribution. The number of edges is 6801, representing the adjacency of the census tracts. The initial population was 80 individuals. The mutation probability is 0.012 and crossover 1. The number of generations was 100. Empirical tests indicate that the minimum and maximum number of nodes should be 34 and 200, respectively.

The first evaluation was aimed at measuring the number of partitions the GA discovered with weights following the Power Law. Therefore, we ran the GA with an initial population having individuals with at least 20, 30 and 40 partitions. Table 1 indicates the results and allows a comparison with a GA that uses the same operators as those proposed by (Seeman et al. 2009). For the best case, 27 out of 35 partitions follow a Power Law.

Table 1. Comparison of the GA using our operators against Datta's operators with the synthetic dataset

Initial population with % Partitions follows a PL	at least 20 partitions	at least 30 partitions	at least 40 partitions
Our approach	70%	77%	61%
Using Seeman's operators	62%	57%	59%

The other way we evaluated our approach was to plot a graphic showing whether the GA is evolving. Fig. 4 shows how the similarity function (Fs) behaves in the 90 generations**.** It is possible to see that the GA evolves until generation 40. The graphic also shows the peaks representing the moment in which the fitness function prefers the number of partitions following a PL rather than continuing to approach the distribution of the original graph.

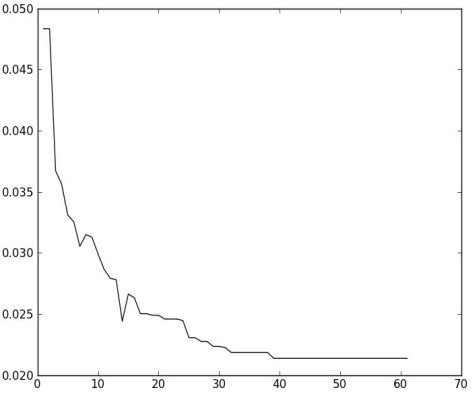


Fig. 4. Evolution of the GA

We realized that the number of partitions of the individuals that form the initial population of the GA determines the final number that the GA can reach. This is due to the fact that the fitness function considers two aspects: the number of partitions following a PL and how close the exponent of the distributions of the partitions is to the exponent of the original graph. We plan in the future to evaluate this using a multi-objective GA in order to cope with this particularity.

5 Conclusion

In this paper we characterize the problem of crime reporting by means of concepts from complex networks. We investigated how genetic algorithms can assist in finding samples of an original graph in which the distribution of the weights of the nodes follows a Power Law with an exponent close to the distribution of the original graph. The best result we found thus far was to partition the initial graph in which 77% of the partitions having the distribution of the node's weights follow a Power Law with an exponent close to the exponent of the distribution of the entire graph. We have also shown that the variation in the crossover operator that we implemented leads to better results compared to another approach that uses a typical operator used for graph partitioning in general.

We intend to continue this research by trying to improve the method, in particular, the genetic operators. We are going to insert knowledge about how a mutation and crossover disturbs a partition in which the weights follow a Power Law. We also intend to test the approach using a multi-objective GA (Deb et al. 2002) for considering the fact that we have two goals to consider, namely, to have the maximum number of partitions following the PL and to have these partitions with an exponent close to the original graph exponent. Tests with other datasets in order to evaluate the generality of the results are also necessary.

Acknowledgments. The first author thanks CAPES for the PROSUP scholarship. Part of the second author work is financed by CNPq grants 306266/2008 and 559977/2010.

References

Cançado, T.: Alocação e despacho de recursos para o combate à criminalidade. Dissertação de mestrado, Departamento de Ciência da Computação, UFMG (2005)

Clauset, A., Moore, C.: Accuracy and Scaling Phenomena in Internet Mapping. Phys. Rev. Lett. 94(1), 018701 (2005)

Cole, R.M.: Clustring with Genetic Algorithms. Master's thesis, Department of Computer Science, University of Western Australia (1998)

Datta, D., Figueira, J.R., Fonseca, C.M., Tavares-Pereira, F.: Graph Partitioning Through a Multi-Objective Evolutionary Algorithm: A Preliminary Study. In: Keijzer, M. (ed.) Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation (GECCO 2008), pp. 625–632. ACM, New York (2008)

Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation 6(2) (2002)

Devroye, L.: Non-Uniform Random Variate Generation. Springer, New York (1986)

Erdős, P., Rényi, A.: On Random Graphs I in Publ. Math. Debrecen 6, 290–297 (1959)

Furtado, V., Ayres, L., de Oliveira, M., Vasconcelos, E., Caminha, C., D'Orleans, J.: Collective In telligence in Law Enforcement: The WikiCrimes System. Information Science 180 (2010)

Gatterbauer, W.: Rules of Thumb for Information Acquisition from Large and Redundant Data. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 479–490. Springer, Heidelberg (2011)

McLafferty, S.: Identification, development and implementation of innovative crime mapping techniques and spatial analysis, p. 27. U.S. Department of Justice, Washington, D.C (2000)

Melo, A.: Um modelo multiagente de simulação criminal bio-inspirado. Dissertação de mestra do, Mestrado em Informática Aplicada. In: UNIFOR (2008)

Pickering, G., Bull, J.M., Sanderson, D.J.: Sampling Power-law Distributions. Tectonophysics 248, 1–20 (1995)

Semaan, G.S., Brito, J.A.M., Ochi, L.S.: Um algoritmo evolutivo híbrido aplicado ao problema de clusterização em grafos com restrições de capacidade e contiguidade. In: Anais do IX Congresso Brasileiro de Redes Neurais e Inteligência Computacional (IX CBRN), Ouro Preto/MG (2009)

Stumpf, M., Wiuf, C., May, R.: Subnets of scale-free networks are not scale-free: Sampling properties of networks. PNAS,102 (March 22, 2005)

Tavares-Pereira, F., Figueira, J.R., Mousseau, V., Roy, B.: Multiple criteria districting problems: The public transportation network pricing system of the Paris region. Annals of Operations Research 154(1), 69–92 (2007)

Measuring a Category-Based Blogosphere

Priya Saha and Ronaldo Menezes

Abstract. Blogs form an essential part of the Web and they are one of the main sources of information to millions of people around the world. Blogs such as Gizmodo, Slashdot, and many others, receive a very large number of daily visitors and consequently are a main force on driving what information becomes known to the public. Furthermore, information in blogs have become crucial to established news agencies such as CNN and NBC, which have dedicated programs and reporters to discuss information in the *Blogosphere*. This paper looks at the structure the blogosphere using Blogspot—Google's blog hosting service—as a case study. We created networks for 12 different blog categories and a combined network. We show that these networks are very similar to the structure of the whole WWW and that the blogosphere is highly connected regardless of category divisions.

1 Introduction

The argument about the ubiquity of the Web has become redundant nowadays because the Web has become so important to our lives that it is inconceivable to think of our world without it. According to the latest numbers from the Internet World Statistics, 2.1 billion people (or about 30% of the world's population) are connected to the Internet and have access to Web content[1]. Web Logs, popularly known as *Blogs*, are personal journals published on the Web. However, the interest in blogs has gone mainstream given its wide adoption. Today most news agencies have

Priya Saha
BioComplex Laboratory, Department of Computer Sciences,
Florida Institute of Technology, Melbourne, Florida, USA
e-mail: psaha2010@my.fit.edu

Ronaldo Menezes
Bio-Inspired Computing Lab, Department of Computer Sciences,
Florida Institute of Technology, Melbourne, Florida, USA
e-mail: rmenezes@cs.fit.edu

[1] March 2011 numbers from http://www.internetworldstats.com/

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 131–139.
springerlink.com

reporters dedicated to the scouting of information in blogs to bring interesting posts to people watching TV. But the effect is not limited to news, blogs are being used for group mobilization such as the Egyptian uprising (in January 2011), London Riots (in August 2011) and even the Occupy Wall Street (in October 2011). Blogs and micro-blogs enable these movements to become "viral" influencing millions of people worldwide.

The number of blogs worldwide has grown rapidly from a total of 50 blogs in 1999 to about 4.1 million just 5 years later. At the time of writing of this paper, BlogPulse indicated that there were 170 million active blogs on the Web with about 70 thousand being created everyday.[2] Today, blogging has become part of our online activities [10]; we can connect to our friends, families, co-workers, but more importantly we can discuss subjects of different categories using an open-forum format. Blogs are partially responsible for the globalization of news because they even help us to overcome geographical barriers [4].

Let us use one example to make the motivation of this work clearer. We built a network of blogs based on physical links that exist in blog sites. That is, a blog is connected to another if one has a hyperlink to the other. However, the flow of information may not be the same for every hyperlink because the blogs have different categories. We investigate the structure of blogosphere but categorized into 12 subjects. We show that all the sub-categories have small-world properties, which lead us to believe that information can flow in the blogosphere independently of the blog category (main subject of the blog). Furthermore, we show that the categories overlap which means that a post in a blog of one category can easily flow to another.

The paper concentrates on blogs hosted by Google's Blogspot hosting service. We have crawled Blogspot sites (sites ending at *blogspot.com* and attributed to each blog a category based on its reachability from 10 starting points of 12 blog categories.

2 Related Works

There has been many works studying blogs and their influence, and how this influence can be computed for blogs, even how blog visits and the number of posts correlates to sales in websites such as amazon.com [9, 12].

The importance of blogs is well motivated but researchers have also looked at other aspects of the Blogosphere such as the style of writing. Huffaker and Calvert [11] studied the characteristics of bloggers (writers of blogs) and showed that the style of writing depends strongly on their gender, age and language. Although this is only slightly related to our structural study here, Huffaker and Calvert's study demonstrates that there is diversity in blogs which leads to a confirmation that the approach we take in this paper considering differences in categories (themes of blogs) is reasonable.

Politics is another important issue that has been studied in the context of blogs and information spread. Farrell and Drezner [8] have performed a study on political

[2] blogpulse.com is no longer available as of January 2012. `http://web.archive.org` maintains the latest version of BlogPulse.

posts in the Blogosphere and found that blogs can shape the political discorse by creating specific focal points. They found that 7% of the general population on the Web use blogs but a staggering 83% of journalists used blogs with 43% of them using/reading blogs on a weekly basis.

Inspired from his work on language acquisition, Roy [14] has recently shown that blogs are excellent predictors of people's likes and dislikes; he has launched a startup company called Bluefin Labs[3] to commercially explore this market niche. The company focuses on monitoring activity in micro-blogs such as Twitter. Roy's work has generated a lot of attention because his approach to measure flow of information in blogs and social media can be used by advertisers to made decisions on where to place their content. Statistical evidence in favor of blogs as generators of content was found in Dautrich and Barnes in a survey of 300 TV programs [5].

Another kind of work related to network of blogs deals with the identification of influential bloggers in a network [1]. Although the original study was done to identify bloggers (people posting on blog sites), it easy to see how this work could be generalized to the entire site. Aggarwal [1] proposes to use four factors to decide on the importance of blogs (bloggers in his study): *(1)* Recognition of the blog site by many people which can be approximated by the number of in-links the blog has; *(2)* Activity Generation of the blog is represented by the number of posts it receives; *(3)* Novelty of the post correlates to it being more important—blogs with more out-links generally represent the fact that the posts are not novel since they refer to content already published elsewhere; and finally *(4)* Eloquence of posts in a blog can be represented by the length of the posts—longer posts generally represent eloquence.

Our approach is different from the ones above because we examine the blogosphere network divisions from the point of view of area of interest or "categories". The paper delves into an analysis of each individual network as well as a full network combining all categories. We show that the small-world properties of each category network may explain why information spreads fast in the Blogosphere.

3 Network Properties

The literature in Network Sciences include a number of metrics that can be calculated to characterize networks which, in turn, may reveal interesting patterns in the relationships of nodes.

Degree is a measure of a node. For the network, we generally look at the *degree distribution* which represents the frequency in which a node with degree k appears in the network, given by $p(k)$. It has been observed that in many real networks [13] the degree distribution roughly follows a power law, that is:

$$p(k) = ck^{-\lambda}, \tag{1}$$

where c and λ are constants. In directed networks, such as the ones we study in this paper, we look at the degree distribution divided into indegree and outdegree.

[3] `http://bluefinlabs.com`

An interesting characteristic of some networks is the transitivity of relations between nodes, generally referred to as the *clustering coefficient* of the network or the is just the average of all C_i.

$$C_i = \frac{2m_i}{k_i(k_i - 1)}, \tag{2}$$

where m_i is the number of links between the k_i neighbors of i.; the clustering coefficient of the entire network (C) is given by $C = \frac{1}{n}\sum_{i=1}^{n} C_i$. As a transitivity measure, C may used to identify small-world networks [15] which are expected to have high clustering (when compared to random networks) and short average path lengths.

For many real networks, their growth is based on the idea that modules combine to form larger modules in a sort of hierarchical manner. The scaling law is given by:

$$C(k) = k^{-\gamma}, \tag{3}$$

where γ is expected to be approximately 1. The hierarchy of a network can be characterized quantitatively by the work of Dorogovtsev et al. [6] who demonstrated that in hierarchical networks the distribution of clustering coefficients of nodes with degree k, given by $C(k)$.

4 Blogosphere Networks: The blogspot.com Case Study

There are many providers that specialize in hosting blogs. Among the most common ones are: Blogspot and Wordpress. Blogspot is Google's blog hosting service and contains thousands of blogs.

The first step in our study is to build a network representing blogspot.com. In our network nodes represent blog sites and the directed edges represent a connection between two of these sites. The general assumption is that if a blog has a link to another, there exists a potential exchange of information from the later to the former. The first step was to collect the a dataset. We used Google to get the names of the top-10 blogs of blogspot.com in 12 common categories: Arts, Business, Cooking, Education, Finance, Health, Music, Politics, Religion, Science, Sports and Travel. These categories reflect general topics found in many newspapers around the world. For each of these topics a Google search was performed following the format: *site:blogspot.com [CATEGORY]*, where *[CATEGORY]* is one of the 12 mentioned.

Once we had the starting points, we wrote a crawler to explore the starting points and the other "blogspot.com" sites they link to until a given *depth*. For instance, if we crawled the starting points using depth 1, it would include the 10 starting points and all sites hosted at blogspot.com that any of the original 10 starting points link to. In this paper we use a dataset that we gathered by crawling the sites until depth 3. During the crawling process we added the initial category as an attribute of every reached site; if a site S is reached while we were crawling the starting nodes for the "cooking" category, site S would also be considered as a "cooking" site. The premise is that if one is reading a blog on cooking and it contains a link to another

site, there is a probability that he will follow that link and maybe post information on the connected site. To summarize, a *[CATEGORY]* tag indicates that information from a *[CATEGORY]* blog can flow to the blog in question.

The network we have created is directed given that it represents the link between blogs (a blog link to another exists independent of the inverse link). The network is also weighted since we want to represent close connections. We assume that if a blog has many links to another, then it considers that blog important. The larger the number of links, the stronger the connection (perceived importance).

It is easy to understand that if we have 12 individual networks and we try to combine them into one, we may have nodes that appear in many of the networks. In other words, the crawling from different locations may lead us to the same blog. This overlap is important because it tells us how information may flow from one category to another.

Figure 1 depicts the full network where nodes are colored according to their category. This figure shows that many nodes are reached from more than one category (showed as black nodes). The visualization alone demonstrate that at depth 3, information may flow from any topic to any topic although the number of black nodes is not as high as one might think. We see the network dominated by music blogs which accounts for nearly 40% of the blogs in the dataset.

5 Network Measurements

We analysed the blog networks for each category as well as the full network. We are mainly concerned with the network measures that we have described previously because they can help us characterize the kind of network formed by the Blogspot blogosphere. The results in Table 1 indicate that each of the category-based blog networks display small-world properties.

In order to better understand the values we show in Table 2, the corresponding values of our Blogspot network is placed against values for other real networks, namely the Internet and the Film actors networks studied by Newman [13].

One issue that has to be observed from Figure 2 is that the outdegree distribution does not follow Barabási and Albert [3] description of a scale-free network. The problem is quite simple to understand and pertains to the fact that the preferential attachment—where hub nodes tend to become even more connected as time goes by—does not account for the fact that the entity being represented by the node ages and hence become less attractive. In the context of blogs, it means that there is a cap on the number of its out-links, which is very natural to understand. Dorogovtsev and Mendes [7] have described a model in which aging is considered and this model is a better fit to the behavior of out-degree distribution in Figure 2. In network terms, the out-degree distribution appears to be a case where the data is better fitted with a generalized power law supplemented with an exponential cut-off (GPL-EC) [2].

Next we try to look at the growth of the network from blogspot.com. Figure 3 shows the clustering coefficient distribution for nodes with degree k. As explained earlier, hierarchical networks tend to closely follow the scaling law.

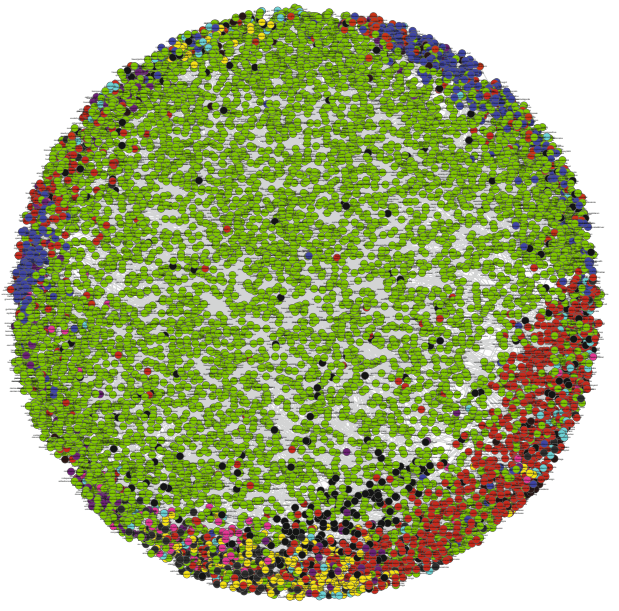


Fig. 1. The the full Blogspot network considering all categories (crawled to depth 3). The categories are denoted by the colors. Due to the density of the network this picture shows edges if they have weight 3 or more. Each color represent a different category but the nodes which can be reached from more than one category are represented in black.

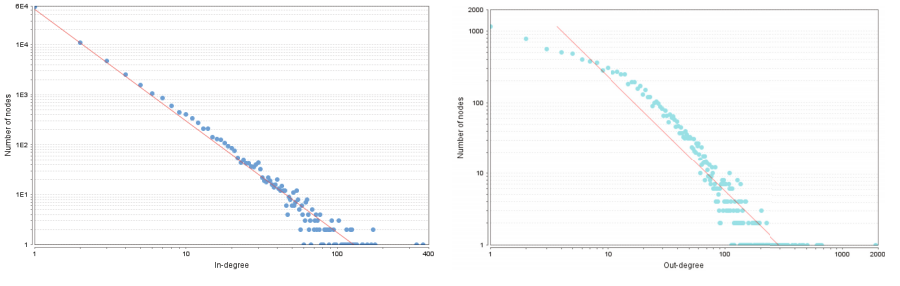


Fig. 2. Figure shows the indegree (left) and outdegree (right) distributions for blogs in the blogosphere. Both distributions follow a power law except that the outdegree has an exponential cutoff.

The full blogspot network follows the scaling law perfectly with $\gamma = 1$. This means that blogs link to each other and form a hierarchy where small tightly-connected groups join into larger groups of blogs which in turn become more tightly connected just to later join again into larger networks.

Table 1. Category-based blog networks. n represents te number of nodes, m is the number of edges, λ_{in} and λ_{out} represent the exponent of the power-law degree distribution of the indegree and outdegree respectively, ℓ is the average path length between pairs of nodes, and C refers to the clustering coefficient of the network

Network	n	m	λ_{in}	λ_{out}	ℓ	γ	C	Color in Figure 1
Finance	252	311	2.5	0.8	2.40	1.4	0.04	Dark Green
Business	1,125	1,379	2.9	0.6	3.05	1.5	0.07	Light Gray
Religion	3,203	5,404	1.8	0.9	3.50	1.1	0.05	Orange
Travel	3,669	5,073	2.7	0.8	3.56	1.2	0.05	Dark Gray
Education	4,160	6,314	2.0	0.8	3.61	1.1	0.06	Pink
Sports	4,949	7,783	2.3	1.1	3.69	1.3	0.07	Brown
Health	6,347	9,455	2.0	0.9	3.80	1.5	0.05	Navy Blue
Politics	8,252	13,368	2.2	0.9	3.91	1.1	0.05	Purple
Arts	10,615	15,572	2.3	1.1	4.02	1.4	0.06	Cyan
Cooking	12,697	26,416	2.1	1.1	4.10	1.1	0.06	Yellow
Science	21,328	37,512	2.1	1.2	4.32	1.2	0.05	Red
Music	32,631	93,647	2.0	1.4	4.51	0.8	0.08	Green
Full Network	82,921	197, 206	2.2	1.6	4.91	1.0	0.07	—

Table 2. The comparison of the full unfiltered blog network with the Internet and the Film Actors statistics

Variable Name	Blog Network	Film Actors	Internet
Number of Nodes (n)	82,921	449,913	10,697
Number of Edges (m)	197,206	25,516,482	31,992
Exponent of power law distr. (λ)	2.2 / 1.6	2.3	2.5
Average Clustering Coefficient(C)	0.07	0.20	0.035
Average path length (ℓ)	4.9	3.48	3.31

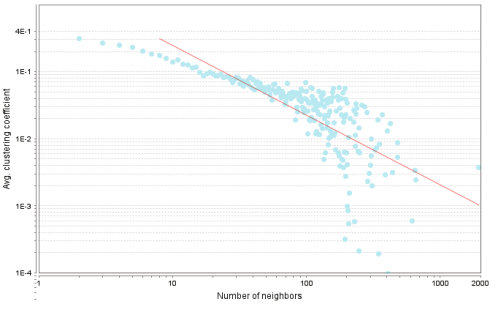


Fig. 3. The clustering coefficient distribution of the blogspot network. The distribution demonstrates that the network is hierarchical given it perfectly follows the scaling law.

6 Conclusion

In this paper, we used a dataset of Google blog hosting service called Blogspot and showed that even categorized blog networks have small-world characteristics. This is an indication that regardless of the subject discussed in blogs, a fad, gossip, or other important information can spread rapidly. This is also reinforced by the fact that if we look at the amount of overlap in these networks, that is, the number of sites that are reached from many different categories (shown as black nodes in Figure 1). From a total 82,921 nodes, 10,099 are reached from different categories (approx. 12% of the nodes). We believe this tells us that information in the blogosphere can easily cross these category-based boundaries. We intend to study the behavior of these overlap as a function of the distance from the starting nodes (depth).

References

1. Agarwal, N.: A study of communities and influence in blogosphere. In: Proceedings of the 2nd SIGMOD PhD Workshop on Innovative Database Research, IDAR 2008, pp. 19–24 (2008)
2. Amaral, L.A., Scala, A., Barthelemy, M., Stanley, H.E.: Classes of small-world networks. Proc. Natl. Acad. Sci. 97(21), 11149–11152 (2000)
3. Barabási, A., Albert, R.: Emergence of scaling in random networks. Science 286(5439), 509 (1999)
4. Cointet, J.-P., Roth, C.: Socio-semantic dynamics in a blog network. In: IEEE International Conference on Computational Science and Engineering, vol. 4, pp. 114–121. IEEE Computer Society (2009)
5. Dautrich, K., Barnes, C.: Freedom of the press survey: General population 2005. University of Connecticut, Department of Public Policy (May 2005)
6. Dorogovtsev, S.N., Goltsev, A.V., Mendes, J.F.F.: Pseudofractal scale-free web. Phys. Rev. E 65(6), 066122 (2002)
7. Dorogovtsev, S.N., Mendes, J.F.F.: Evolution of networks with aging of sites. Phys. Rev. E 62(2), 1842–1845 (2000)
8. Farrell, H., Drezner, D.: The power and politics of blogs. Public Choice 134(1), 15–30 (2008)
9. Gruhl, D., Guha, R., Kumar, R., Novak, J., Tomkins, A.: The predictive power of online chatter. In: KDD 2005: Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 78–87. ACM Press, New York (2005)
10. Guo, J., Fan, C., Guo, Z.: Weblog patterns and human dynamics with decreasing interest. The European Physical Journal B - Condensed Matter and Complex Systems 81(3), 341–344 (2011)
11. Huffaker, D.A., Calvert, S.L.: Gender, identity, and language use in teenage blogs. Journal of Computer-Mediated Communication 10(2) (2005)
12. Java, A., Kolari, P., Finin, T., Oates, T.: Modeling the spread of influence on the blogosphere. In: Proceedings of the 15th International World Wide Web Conference (March 2006)

13. Newman, M.: The structure and function of complex networks. SIAM Review 45(2), 167–256 (2003)
14. Roy, D.: New horizons in the study of child language acquisitio. In: Proceedings of Interspeech, Brighton, England (2009)
15. Watts, D., Strogatz, S.: Collective dynamics of small-world networks. Nature 393(6684), 440–442 (1998)

Ripple Effects: Small-Scale Investigations into the Sustainability of Ocean Science Education Networks

Robert Chen, Catherine Cramer, Pam DiBona, Russel Faux, and Stephen Uzzo

Abstract. Education Networks are an important way for educational institutions to develop and share knowledge and resources. Yet, methods of evaluating what makes them successful have been elusive. Here, we present a network analysis of the New England Ocean Science Education Collaborative (NEOSEC), a successful ocean science literacy collaborative and an effort to reveal characteristics inherent to successful education networks. NEOSEC is a network comprised of more than 40 institutions, with a stated goal of advancing ocean literacy in the region. Analysis of the evolution of this network suggests that network analysis adds an important dimension to evaluating education networks, and that successful educational networks may exhibit network characteristics that could aid in understanding their functionality and sustainability. Preliminary results also indicate that as

Robert Chen
University of Massachusetts Boston
e-mail: Bob.Chen@umb.edu

Catherine Cramer
Centers for Ocean Science Education Excellence
e-mail:catherinecramer@comcast.net

Pam DiBona
New England Aquarium
e-mail: pdibona@neaq.org

Russel Faux
Davis Square Associates
e-mail: russell@davissquare.net

Stephen Uzzo
New York Hall of Science
e-mail: suzzo@nysci.org

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 141–147.
springerlink.com

these networks increase in complexity they may exhibit characteristics of other kinds of complex networks.

1 Introduction

Developing effective networks is critical to the success of educational programs, the spread of excellence across scales of educational practice, and the sustainability of communities devoted to a shared mission (Austin 2000). Yet there have been relatively few attempts to look at the structure and dynamics, and resulting effects and sustainability of these communities through a network lens (Durland & Fredericks 2005). The evaluation of consortia in general has tended to look at the effects of the efforts of individuals or organizations rather than investigating the structures and dynamics of networks (Cross et al. 2002).

Over the past year the Center for Ocean Science Education Excellence-Ocean Communities in Education And social Networks (COSEE OCEAN), an NSF-funded ocean science literacy center, has assembled an interdisciplinary team to look at the effectiveness and sustainability of networks of education communities developed to increase ocean literacy among multiple audiences. We performed an analysis of NEOSEC in an effort to reveal characteristics inherent to successful education networks. Analysis of this network suggests that successful educational networks exhibit characteristics that could aid in understanding their functionality and sustainability. As these networks aggregate and increase in complexity, they may also reveal complex network characteristics like clustering, and preferential attachment, albeit further study is needed to substantiate this.

2 Analysis

Founded in 2006, NEOSEC is a diverse, networked collaboration of 43 institutions from across New England, including aquaria, museums, universities, government entities, and science and research centers. COSEE OCEAN is a member of this network.

Network analysis for this study was conducted based on the Himmelman model (Himmelman 2002) to assess the increase in collaboration among members. The intent was to investigate the following questions: “What changes can be seen in the inter-organizational collaborations within NEOSEC?”; and “Are there organizational characteristics that affect participation in the network?”

The sample is comprised of organizational members of NEOSEC (N=43) (in NEOSEC, individuals act as representatives of their institutions), with 38 of these submitting analyzable responses after the data were cleaned (for a final response rate of 88%). The survey asked about depth of interactions with fellow collaborative members at two time points: 2005, prior to NEOSEC forming; and then again at August 2011. It utilized a scale ranging from “We did/do not know of this group” to “We had/have sustained collaborations with this group.” All analyses were conducted in UCINet and SPSS, with network visualizations done in NetDraw, Gephi and Pajek.

2.1 Node-Level Metrics

An ego network is composed of a single network node, referred to as the *ego* (in this case, a single NEOSEC member institution) and the other nodes with whom the ego claims to have a relation, known as the *alters*.

Table 1. Summary of ego network density gain (*significant at $p<0.05$ [paired samples t test])

Time	Mean	Std. Deviation	Effect Size
T1_Ego_Density	72.85	6.66	
T2_Ego_Density	88.51*	1.38	0.88

In Table 1, *Ego density* refers to the extent to which the alter organizations are linked. Note that gains seen here are significant, with a sharp downward turn to the standard deviation. Effect size here is also large. NEOSEC member organizations are interacting more with other organizations that are in turn interacting with one another. Changes in standard deviation convey sharp increases in ego network densities.

We theorized that meeting attendance and funding levels might have significant impact on network effects and calculated the correlations between these variables and ego network density gains (Table 2). Limiting analysis to only organizations that participated in joint projects with federal funding to NEOSEC, we found a non-significant and negative correlation between funding score and meetings attended. We concluded that these 3 variables are only weakly correlated for funded projects. Multiple regression (not shown) using state, meetings attended, and composite funding as independent variables and ego network density gains as a dependent variable also yielded inconclusive results.

Table 2. Correlations between meeting attendance, funding received, and centrality gains for organizations in jointly funded projects

		Meetings Attended	Composite Funding	EgoNet Density Gain
Meetings Attended	Pearson Correlation	1	-0.127	0.176
	Sig. (2-tailed)		0.652	0.529
Composite Funding	Pearson Correlation	-0.127	1	-0.336
	Sig. (2-tailed)	0.652		0.220
Network Density Gain	Pearson Correlation	0.176	-0.336	1
	Sig. (2-tailed)	0.529	0.220	

2.2 Dyadic-Level Metrics

In this section two metrics, namely degree centrality (out) and Eigenvector centrality (out) were analyzed. *Out* ties are those a given organization claims to have with others in NEOSEC, as opposed to ties others claim to have with a given organization. *Out degree* refers to the number and strength of ties a given organization claims to have with the others. *Out Eigenvector* refers to the number and strength of ties that one's ties claim to have. An organization with a high level of Eigenvector centrality will be connected to other organizations that have many connections of their own, connections that may or may not be connected to the original organization.

Table 3 shows that NEOSEC members reported significant increases in both the number and strength of their collaborations with other NEOSEC members. In addition, these connections themselves increased in their collaborative ties with other NEOSEC members. The approximate η^2 effect sizes (eta-squared) are very solid, leading one to conclude that the network is has grown significantly in complexity.

Table 3. Pre-Post Dyadic-Level Metrics (*Significant at $p<0.05$ [paired samples t test])

	Mean	Std. Deviation	Effect Size
T1_Out Degree	23.82	8.36	
T2_Out Degree	32.45*	4.69	0.57
T1_Out Eigenvector	0.68	0.23	
T2_Out Eigenvector	0.89*	0.12	0.52

NEOSEC organization ties have increased in number and strength accompanied by smaller standard deviations. Geographically, organizations with greater centrality are clustered in Boston and southern New England coastal areas, and there is a much more even distribution of centrality extending over the entire region.

The question then arises as to whether the centrality increases were caused by the members associating with one another within the context of NEOSEC. We looked at whether there might be an association between the number of NEOSEC meetings attended and centrality gains. The meeting attendance factor was not, however, found to be a significant predictor (linear regression) of gains in centrality, suggesting that the increases in connections among NEOSEC members took place outside of official NEOSEC functions. On the other hand, members who attended fewer than four meetings appear to show less gain than members who attended more meetings. This observation is suggestive of attendance having a consistent relationship with centrality gains, though again, the relationship fell short of statistical significance.

Roughly 40% of the responding organizations participated in grant-funded work over the time period since the beginning of NEOSEC. To examine the relationship between funded activities and network change, we created a scale ("composite funding") with organizations receiving two "points" for each directly funded program

and one "point" for each grant for which the organization served as an unfunded resource partner. Little correlation was found between funding levels and centrality gains (0.282, not significant, Pearson).

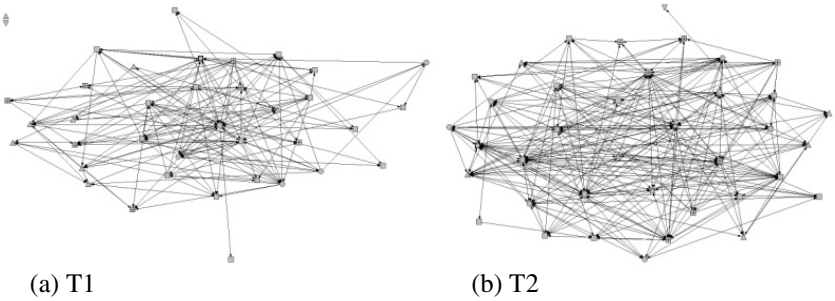
2.3 Network-Level Metrics

In this section we present findings from the analyses of the pre-post effects in the overall NEOSEC network. In this case we measured density, centralization and hierarchy for both prior to and after the formation of the NEOSEC network.

Table 4. General network-level metrics

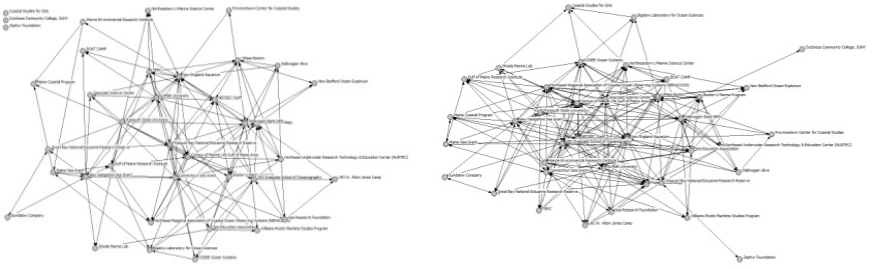
Metric	Time 1 (T1)	Time 2 (T2)
Density	1.49	2.26
Centralization	36.29%	31.16%
Hierarchy	0.05	0.00

In Table 4, *Density*, the number and strength of ties compared to the number and strength of possible ties, can be quite sensitive to context. What might be seen as dense in one context (e.g., a law firm) might be viewed as sparse in another (e.g., a family reunion). For NEOSEC, the density values show a robust increase on the order of 52%, and while one might interpret these values in different ways, the observed increase is sizeable. Centralization values, the degree to which the network exhibits "hub and spoke" structure, decreased only slightly, indicating that nodes with higher degree distribution remain so as the complexity and frequency of interactions increases: overall the network retains a rather centralized structure. Conversely, the virtual disappearance of hierarchy indicates the lack of a "pecking order" or a general equivalence among the organizations, in which a more peripheral organization can easily reach a more core organization (see Figure 1).



(a) T1 (b) T2

Fig. 1. NEOSEC Network by state. T1(a) Note that there is some clustering and a distinct set of core organizations in the center. T2(b) reveals a continued clustering, but no significant (Kruskal-Wallis) between-group differences in terms of centrality gains. An increase in overall density and complexity of ties is evident, but previously peripheral organizations now have increased ties to more central ones.



(a) T1 (b) T2

Fig. 2. NEOSEC Collaboration network by organization. T1(a) depicts the overall structure with nodes shaped by organization. Note some clustering but about half of nodes have <6 connections and none >16. T2(b) reveals a higher degree of clustering and centrality gains with a few nodes having dramatically increased levels of collaboration ties (>19) to other organizations.

3 Conclusions

The NEOSEC network has grown significantly in its overall cohesion in the patterns of relations among member organizations, and in the ego network density of individual organizations. Network density values show a sharp increase and hierarchy values (referring to non-reciprocated ties) for all practical purposes disappeared. Centralization values indicate strong network actors before NEOSEC and that they remain so. Change in the coreness values was not significant from T1 to T2, indicating that organizations that were central before NEOSEC remain so, even as the network becomes more complex.

Dyadic measures indicate significant increases with very good effect sizes, meaning that organizations have more and stronger ties with other organizations that are in turn better connected with others. This underscores the conclusion that the network grew in ways that involved all network members. The ego network density values increased sharply as well, though we found no significant relations here to the state in which the member operates, the levels of funding, or the number of NEOSEC meetings attended.

As illustrated by Figure 2, degree distribution for a small number of centralized nodes rises noticeably between T2 and T2. An artifact of the development of the NEOSEC network is a possible trend toward a scale-free structure, although additional time slices would be needed to corroborate this trend.

4 Discussion

The growth and development of educational networks depend on a host of internal and external parameters. Critical components appear to be: increased face-to-face interactions and resulting knowledge, shared goals and vision, paid staffing, and opportunities for collaboration.

Examination of NEOSEC through the lens of network analysis revealed that it has evolved from a loosely affiliated set of organizations into a tightly knit network. No single identifiable factor appeared that drove NEOSEC members to develop their relations with others, yet the NEOSEC network has effectively built overall regional capacity for expanding ocean literacy in New England. This network responds to opportunity and has a high degree of trust, which has enhanced its sustainability. The observed, meshed cohesion creates support structures for taking on complex projects that take advantage of each actor's capabilities.

The network analysis revealed important findings about the growth of the network, findings that could not have been gathered with comparable rigor in any other way. While measured ties are not the only ties that exist between actors, these ties appear to be important to network functioning and are critical indicators of network vitality. The implications of findings are currently under consideration, but a cursory look at the data reveal relative activity of individuals or groups and possible tactics for planning activities, adjusting strategies, and meeting overall educational goals.

By evaluating performance, patterns and dynamics in a small, focused, but diverse set of networks of different scales within the education arena, our goal is to use network analyses like this one to uncover strategies to effectively build sustainable education networks. Network analysis offers an approach to study the scaling of educational network structures that may be well suited to identifying bottlenecks in network structure and serve as a diagnostic tool for optimizing network function to achieve learning goals. It supports an informed process of reflection and new lines of inquiry in the structuring and evaluation of education networks. The ultimate goal is to develop network diagnostic capacity to aid in the development of networks of engaged ocean scientists and educators - a powerful force in support of learning and discovery.

References

1. Austin, J.: The Collaboration Challenge: How Nonprofits and Businesses Succeed Through Strategic Alliances. Jossey-Bass, San Francisco (2000)
2. Cross, R., Borgatti, S., Parker, A.: Making Invisible Work Visible: Using Social Network Analysis to Support Strategic Collaboration. California Management Review 44(2), 25 (2002)
3. Durland, M., Fredericks, K.: An Introduction to Social Network Analysis. New Directions for Evaluation 5 (2005)
4. Himmelman, A.: Collaboration for a Change: Definitions, Decision-Making Models, Roles, and Collaboration Process Guide. Himmelman Consulting, Minneapolis (2002)
5. Holdren, J., Lander, E., Varmus, H.: Prepare and Inspire: K-12 Education in Science, Technology, Engineering and Math (STEM) for America's Future. President's Council of Advisors on Science and Technology, Washington, D.C (2010)

Socio-dynamic Discrete Choice on Networks in Space: Impact of Initial Conditions, Network Size and Connectivity on Emergent Outcomes in a Simple Nested Logit Model

Elenna R. Dugundji and László Gulyás

Abstract. The reported research treats interactions between agents and generated feedback dynamics in the adoption of various transportation mode alternatives. We consider a simple nested logit model where an agent's choice is directly influenced by the percentages of the agent's neighbors and socio-economic peers making each choice, and which accounts for common unobserved attributes of the choice alternatives in the error structure. We explicitly address non-global interactions within several hypothesized social and spatial network structures. Discrete choice estimation results controlling heterogeneous individual preferences are embedded in a multi-agent based simulation model in order to observe the evolution of choice behavior over time with socio-dynamic feedback due to the network effects. For the particular simple model under study, we find the impact of initial conditions on the emergent long-run behavioral outcomes is dependent on network size and network connectivity. We conclude highlighting limitations of our present study and recommendations for future work.

1 Introduction

Suppose you have the possibility to choose to adopt one of a number of different behaviors or to choose to buy one of a number of different products. Moreover, suppose the choice is multi-dimensional or more generally, that there are common

Elenna R. Dugundji
Universiteit van Amsterdam, Amsterdam, Netherlands
email: e.r.dugundji@uva.nl

László Gulyás
AITIA International Inc., Budapest, Hungary
email: lgulyas@aitia.ai

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 149–160.
springerlink.com

unobserved attributes of the choice alternatives. A classic approach to statistical prediction in such a situation given an observed sample of decision-making agents in a population is the *nested logit* model, pioneered by Ben-Akiva [1] in the context of transportation demand modeling. Now suppose your choice is influenced by your individual perception of the choices made by your neighbors, colleagues and/or socioeconomic peers. Brock and Durlauf [2] have noted: There has yet to be any analysis of (nested logit) models... when self-consistency is imposed on the expected group choice percentages. Such an analysis should provide a number of interesting results. It is our aim to fill this gap. We present an application of the model to transportation mode choice using pseudo-panel microdata in the greater Amsterdam region, combining econometric estimation with computational techniques from multi-agent based simulation. This paper extends earlier work by the authors [3] by exploring effects of various hypothesized treatments of which decision-makers influence each other defined on the basis of spatial proximity and socioeconomic group.

2 Model Assumptions

Discrete choice theory allows prediction based on computed individual choice probabilities for heterogeneous agents' evaluation of alternatives. For the nested logit model we assume a sample of N decision-making agents indexed $(1,\ldots,n,\ldots N)$ each faced with a single choice among mutually exclusive elemental alternatives i in the choice subset C_n of some universal choice set C. The choice set C_n faced by agent n is partitioned into M mutually exclusive and collectively exhaustive "nests" C_{mn} of elemental alternatives which are assumed to be correlated. In general the composite choice set C_n will vary in size and content across agents: not all elemental alternatives i in the universal choice set may be available to all agents. The overall correlation structure of alternatives is however assumed to be the same across agents, aside from availability. A key feature of the nested logit model is thus that the symmetry of choice behavior is inherently broken the assumed correlations among elemental alternatives. For detailed specification the interested reader is referred to Ben-Akiva and Lerman [4] and earlier work by the authors [3].

The research reported here explores interactions between a decision-maker and the aggregate actions of other decision-makers proximally situated in a sociogeographic network. We use a priori beliefs about the social and/or spatial dimension of interactions to formulate the connectivity of the network. In the case study to be discussed, we have rich socioeconomic data for each respondent as well as the geographic location of each respondent's residence and work location. This allows us to define aggregate interactions by grouping agents into geographic neighborhoods or into socioeconomic groups where the influence is assumed to be more likely. In the simplest case, these groups are assumed to be mutually exclusive and collectively exhaustive. That is each agent belongs to one and only one group. The agent is assumed to be influenced by the average choice behavior of his or her group, and the influence by other groups is assumed to be negligible. At a global

level, the picture is a fragmented or disconnected network of *clustered groups*. If we are interested in equilibrium behavior, the consequences of such an assumption are important: there is no transmission of influence across groups, and the global picture is a weighted average behavior of the separate clusters. Thus we also consider cases with *overlapping groups*, with agents for example connected by social group as well as by residential district, or by postcode regions of residence and work location. This leads to a giant cluster for the empirical examples under consideration, with the important implication that influence can spread throughout the entire population.

3 Case Study

The data used in this paper originates from travel questionnaires administered by the Municipality of Amsterdam Agency for Infrastructure, Traffic and Transport (dIVV) in Amsterdam and a neighboring suburban municipality to the south, Amstelveen. The data set made available by the dIVV is a subset of the full modal split database, containing only direct home-work trips and direct work-home trips where the purpose of the trip at the non-home location is classified as either "work" or "business." The data received includes records of trips where respondents have indicated one of the following transportation mode choices: external system public transit or internal system public transit (23,7% mode share); bicycle or moped/motorcycle (26,7% mode share); car driver or car passenger (49,6% mode share). The final data sample used in the case study contains 2913 decision-making agents. Raw socio-economic variables available for use in the model include among other agent characteristics: income category, education level, age. Geographical location of the home and non-home locations are given in the data in terms of the centroid of a traffic analysis zone (TAZ). The dIVV considers 381 TAZ centroids in Amsterdam and 48 TAZ centroids in Amstelveen, with a total of 933 TAZs in the whole of the Netherlands. Using standard GIS software, the centroids are mapped onto postcode zones and municipal districts.

3.1 Fully Connected Network: Initial Conditions and Size Effects

Although we are fundamentally interested in non-global interactions, we start our modeling endeavor by first considering a fully connected network with global interactions. The reason for this is that when the model includes "self-loops", that is, each agent counts it own choice in evaluating the choices made by reference agents, the steady-state solutions of the socio-dynamic system can be solved analytically as derived in Dugundji [5], since the agents are perfectly homogenous in this special case. Such an analytical benchmark is useful for verification of our programming implementation of the multi-agent based model to confirm that we get expected results under known conditions in parameter space. Furthermore, the

benchmark can help us later to interpret emergent outcomes as we change the parameter settings step-by-step away from the known analytical case. By studying the simulation results for the fully-connected network under controlled conditions varying the initial starting mode shares and network size, it can help us gain insight in subsequently understanding the behavior of the system with hypothesized sociogeographic networks.

We estimate a simple nested logit model on the basis of the sample data. The only observed explanatory variable in the model is the network interaction variable. Unobserved heterogeneity across the transportation mode choice alternatives is captured by nesting the alternatives that are assumed to be correlated. Since we have only three elemental choices in this case study, there are only three possible nesting structures, namely public transit nested with bicycle, public transit nested with car, and bicycle nested with car. Estimation of the three successive nested logit models have shown the second nesting structure to be most significant in terms of loglikelihood ratio test and in terms of the *t*-test on the nest coefficient. The unobserved heterogeneity might represent here for example individual preference for a "motorized" transportation mode. The estimation results for this model are given Table 1. In a typical empirical application we would usually consider additionally other explanatory variables in the specification of the utility function, including individual-specific socioeconomic characteristics of the commuters (eg. gender) as well as individual-specific attributes of the choice alternatives (eg. travel time), and the availability of alternatives (eg. while Amsterdam and Amstelveen are well served by public transit, not everyone in the might be able to commute by public transit if there is no transit service at their work destination). We defer this detailed study for future research. Our goal with the estimation here is not an analysis to inform policy, but rather to generate empirically plausible parameters to study the theoretical behavior of the system. We deliberately restrict our consideration in this paper to the simple model in order well understand the fundamental behavior of the simple model first before proceeding to an even more complex situation. This way we can focus on understanding the network effect in the nested logit model without confounding the contributions to the long-run results.

Table 1. Estimation results for simple nested logit model with fully connected network [a]

Variable	Coefficient Estimate	Standard Error	t-Statistic
Share of respondents in the sample choosing each mode	2.76	0.16	1.78 (against 0)
Scale parameter for transit-car nest	1.03	0.05	0.67 (against 1)

a) Null log-likelihood: -3200; Final log-likelihood –3035; Likelihood ratio test: 331.

Using the approach described in Dugundji [5], we find that there are five equilibrium solutions for the long-run behavior of this simple nested logit model with sociodynamic feedback with global interactions for the particular estimated

parameter values in Table 1. Three of these solutions are stable and two of these solutions are unstable. See Table 2. Due to the symmetry of the system whereby transit and car are nested together, at any mode share value for which there is a solution for transit, there will be a dual solution with an analogous mode share value for car, and vice versa.

Table 2. Analytical equilibrium solutions for simple model with fully connected network

Solution Nr.	Stability	Bicycle Mode Share	Transit Mode Share	Car Mode Share
1	(most) stable	0.700	0.150	0.150
2	stable	0.158	0.143	0.698
3	stable	0.158	0.698	0.143
4	saddle node	0.267	0.237	0.496
5	saddle node	0.267	0.496	0.237

Using the Repast modeling platform (http://repast.sourceforge.net), we create a computational version of this model that will allow us to experiment with different hypothetical scenarios that can either be derived from the sample data, or tweaked by the modeler accordingly to study variation. There are three aspects that we will consider in this paper: initial starting mode shares, network size and network connectivity. In order not to confound the effects of the utility parameters in the econometric estimation with the time-varying evolution of the mode shares, we use the estimated coefficient values in Table 1 for all multi-agent based simulations in this paper. We defer the re-estimation of parameters based on different network structures for future research.

Example time series results for the mode shares with a fully connected network under different scenarios are shown in Fig. 1. Each run is allowed to iterate for 600,000 time steps. This is approximately 200 revisions of choices with asynchronous decision-making for the network of 2913 agents. The black times series represents the proportion of agents choosing car at any given time step, the light gray time series represents the proportion choosing bicycle, and the dark gray time series represent the proportion choosing public transit. Fig. 2 shows observed long-run outcomes at the last time step when applying different random seeds for determining the decision-making order for agents evaluating the choice distribution and updating their choice.

From Table 2 we know the most stable solution occurs with a mode share for bicycle of 0.700 and mode shares for transit and car of each 0.150. In practice we do not expect to see the saddle node solutions. Also, for initial starting mode shares as in the survey data with almost 50% car commuters, and less than 25% transit users, we might expect in practice that stable solution nr. 2 listed in Table 2 with mode share 0.698 for car and mode share 0.143 for transit will be more likely to be reached than its dual solution number 3 with the mode shares reversed. In the upper left panel of Fig. 1 we see an example time series where the stable solution nr. 2 is gradually reached. In the left panel of Fig. 2, over multiple runs we indeed consistently obtain the analytically predicted first two equilibrium solutions in Table 2.

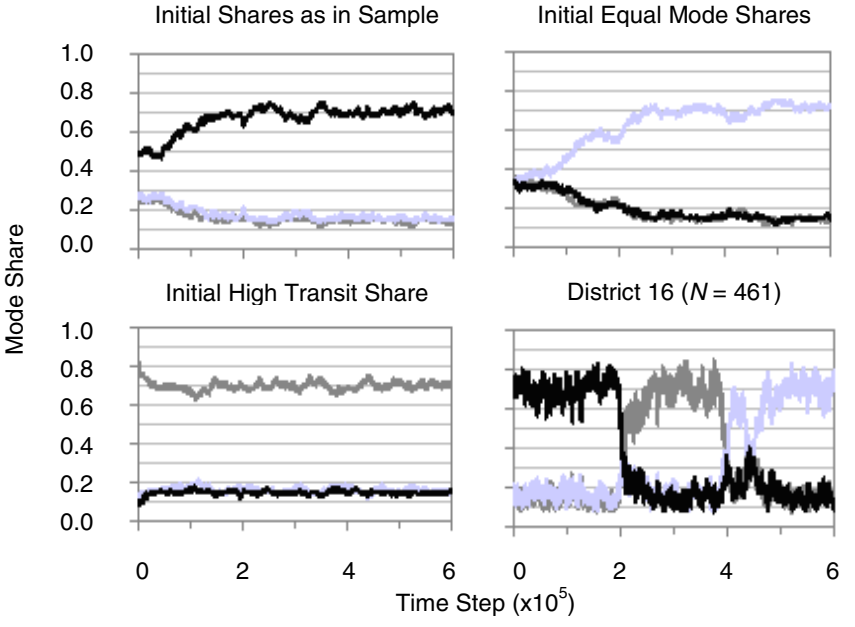


Fig. 1. Example time series for nested logit model with fully-connected network under different initial conditions ($N = 2913$, except for District 16)

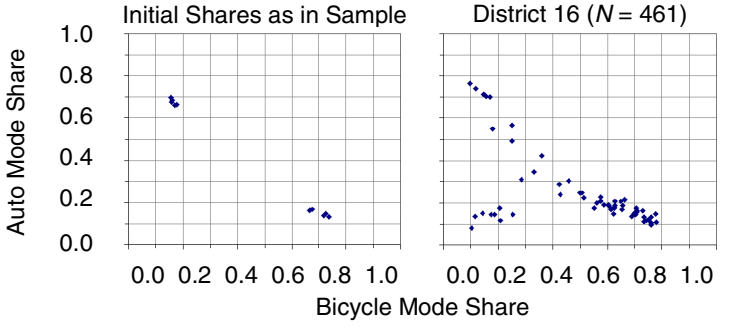


Fig. 2. Observed mode shares at t = 600,000 with different random seeds for determining the agent decision-making order with the nested logit model on a fully-connected network. Initial conditions have no significant effect on long-run outcomes under volatile dynamics.

Next, we test a hypothetical case where the initial mode shares are not determined from the survey, but are tweaked so that the starting mode shares are equal, ie. all one-third. In this non-biased case, we might expect the most stable equilibrium to be dominant. In the example time series in the upper right panel of Fig. 1 we see initially ambivalent average behavior, but once the runaway effect is established, the series proceeds to the stable solution and then stays there. Over

multiple runs with different random seeds determining the decision-making order we find that all runs went to the dominant equilibrium.

Then, we consider a hypothetical case where the initial mode share for public transit is very high (80%). The example time series in the lower left panel of Fig. 1 moves easily to the stable solution nr. 3 listed in Table 2, and stays there. Over multiple runs with different random seeds, we find that all runs are locked in at this steady state.

Finally we consider the effect of the size of the network on the long-run behavior. Since our sociographic networks with clustered groups can be expected to show the weighted average behavior of the separate clusters, it is useful to see how a separate cluster behaves. In our case study, a separate cluster is assumed by design to be a fully connected network of a subset of the total number of agents. The time series in the lower right panel of Fig. 1 shows an example for District 16 (Amstelveen South) where N = 461. Here the initial mode share for car within the district is particularly high (66,2%). However, instead of finding a long-run lock-in at stable solution nr. 2 listed in Table 2, because the dynamics become more volatile with the lesser number of agents we see that the time series cycles between all three stable equilibria in Table 2, and all modes have an opportunity in turn to become dominant.

The reason for the volatility with the smaller network size has to do with the assumption of each agent's choice being influenced here by the percentage of agents the reference group making each choice. Since each cluster by definition here contains only a subset of the total number of agents in the data sample, the influence of an individual agent updating a choice is larger within the fully connected cluster than when considering the entire data sample being fully connected. This relatively larger jump in mode share for a particular mode alternative as a given agent updates its choice within a smaller reference group, gives the possibility to jump out of a particular steady state and move to another one. In the right panel of Fig.3, when considering the behavior over multiple runs with different random seeds determining the order in which agents make decisions, we find a picture of emergent outcomes scattered across the three stable equilibria, transitioning through the region of the saddle node equilibria. Furthermore the most stable solution nr. 1 is notably dominant over all runs, despite the initial conditions of starting with high car mode share in the District.

In summary, thus far we have seen: 1) lock-in at the analytically predicted stable steady-states, 2) manifestation of the analytically predicted most stable equilibrium being dominant, and 3) for a fully connected network with smaller size, the larger jump in mode share as an agent updates each choice per iteration breaks the lock-in that we found when the entire sample is fully connected. Given this knowledge, we now proceed to our study of sociogeographic networks.

3.2 Sociogeographic Networks: Clusters and Overlapping Groups

We begin our consideration of sociogeographic networks with a broad classification by residential district. The districts in the Municipality of Amsterdam are meaningful entities for the purpose of our case study since they have their own

local government structures (http://www.amsterdam.nl/gemeente/stadsdelen) with their own directly elected representative aldermen. In the multi-party system in the Netherlands, the composition of the majority coalition in one district may be different than the majority coalition in another district reflecting the different local cultures associated with the districts. As a result the districts have the possibility to organize themselves in different ways and set different spending priorities. Residents identify themselves with their districts and often deliberately choose to live in a particular district, and not another district. In our case study, there are 9 districts represented in the data, ranging in size from 223 sampled respondents (District 4, Amsterdam East) to 461 sampled respondents (District 16, Amstelveen South). The mean size is 323 respondents with standard deviation 74, skewness 0.32 and kurtosis 0.19.

In order to be able to test the effect of spatial scale, by way of comparison with the network interdependence defined by residential district, we also define a smaller neighborhood region of influence on the basis of 4-digit postcode. There are 67 postcode regions represented in the sample, ranging in size from 10 sampled respondents to 161 sampled respondents. The mean size is 43 with standard deviation 32, skewness 2.1 and kurtosis 4.4. As with districts, the postcode definitions are also meaningful in that they do not have arbitrary boundaries: residents know in which postcode zone they live and the postcode zones have different reputations. The postcodes in the greater Amsterdam metropolitan region are generally defined such that there is homogeneity within a zone, and heterogeneity across zones, in terms of land uses and built environment according to the period that the zone was originally developed and/or subsequently re-developed in the incremental growth of the region over the years. Our assumption is that the postcode boundaries delineate spatial peers and that agents residing within a particular postcode have similar underlying preferences and values, thus exerting a relatively stronger influence than agents who live outside the postcode.

Next, under the assumption that respondents are influenced by the choice behavior of others in their own socioeconomic class regardless of their residential location, we define 13 socioeconomic groups using the three variables age, income and education [3]. The groups range in size from 99 sampled respondents to 385 sample respondents. The mean size is 224 respondents with standard deviation 111, skewness 0.33, and kurtosis –1.8. Our assumption is that a respondent's direct friends and colleagues are likely of a similar socioeconomic status. For the purposes of the case study, we consider here a relatively dense network of overlapping groups where agents are connected by both their residential district and socioeconomic class.

Finally, since we are considering commute behavior and the work location is known in the data set, we define postcode regions for the work locations. Here again the idea is that due to urban planning policy in the Netherlands the type of work locations and their accessibility will tend to be more homogeneous within a postcode zone than across postcodes. The socio-cultural and practical acceptance of traveling to work by bicycle, for example, may be likely to be higher in a postcode zone where many commuters already travel by bicycle. This in turn may inspire other workers who initially travel by another transportation mode, to revise

their mode choice. The mechanism may occur through various different means, such as direct communication with their colleagues, financial travel reimbursement incentives from their employers, simply being aware that colleagues commute by bicycle, or even just seeing lots of other bicycles parked outside on the street or in a bicycle parking area. Furthermore if there is a critical mass of bicycle commuters to a particular area, there is more stimulus to provide better bicycle facilities, such as covered bicycle parking and dedicated bicycle paths. Regardless of the precise underlying mechanism of the interaction, such an effect can approximately modeled in the aggregate as an agent being influenced by the proportion of other agents in their work postcode zone making a given mode choice. For the purposes of the case study, we define connectivity of interactions with an overlapping network where an agent is influenced both by the proportion of agents making a given choice in their work postcode zone and their residential postcode zone. This leads to a network which is much less dense than the scenario with overlapping residential districts and socioeconomic groups.

Using our multi-agent based model, we now explore the evolution of the choice behavior of the nested logit model with feedback defined by these sociogeographic networks. By experimental design, the network of disconnected residential district clusters and the network of overlapping residential district and social group are approximately five times more dense than the network of disconnected postcode zone clusters and the network of overlapping residential and work postcode zones, respectively. Example time series for the nested logit model with sociogeographic network interaction are shown in Fig. 3. As in the previous section, each run is allowed to iterate for 600,000 time steps; the black times series represents mode share for car, light gray represents mode share for bicycle, and dark gray represents mode share for public transit.

We first consider the two scenarios with a disconnected network of clustered groups. From our study of fully connected networks in the previous section, we know that smaller network size leads to more volatile sociodynamics in our model. Since we saw evidence of this volatility already in the largest district comprising 461 agents, we might expect that all other residential districts (with the smallest having only 223 agents) and accordingly all postcode zones (ranging in size from 161 to 10 agents) can only be more volatile. Furthermore since transmission of influence is prohibited across the separate clusters, the overall time-varying behavior of the global modal split must logically be the weighted average behavior of the mode shares within the separate clusters. In the example time series in the upper left panel of Fig. 3 we see initially persistent average behavior across the clusters with high mode share for car, but then over time the most stable equilibrium in Table 2 tends to become dominant. However, we see in the example time series that the overall bicycle mode hovers around 0.5 never reaches the mode share of 0.7 that it did in the previous section, since there is no interaction mechanism in this scenario to coordinate across clusters. With the volatility that we saw for District 16 alone in Fig. 1 and given the dispersion of long-run outcomes across different seeds that we saw for District 16 alone in Fig. 2, we might presume that is unlikely that all 9 clusters here will happen to be in the same equilibrium at the same time. While the majority of clusters may tend to be at the most

stable equilibrium in the long-run, probabilistically there will be some clusters that will be in one of the other two stable equilibria, or in the process of transitioning between equilibria.

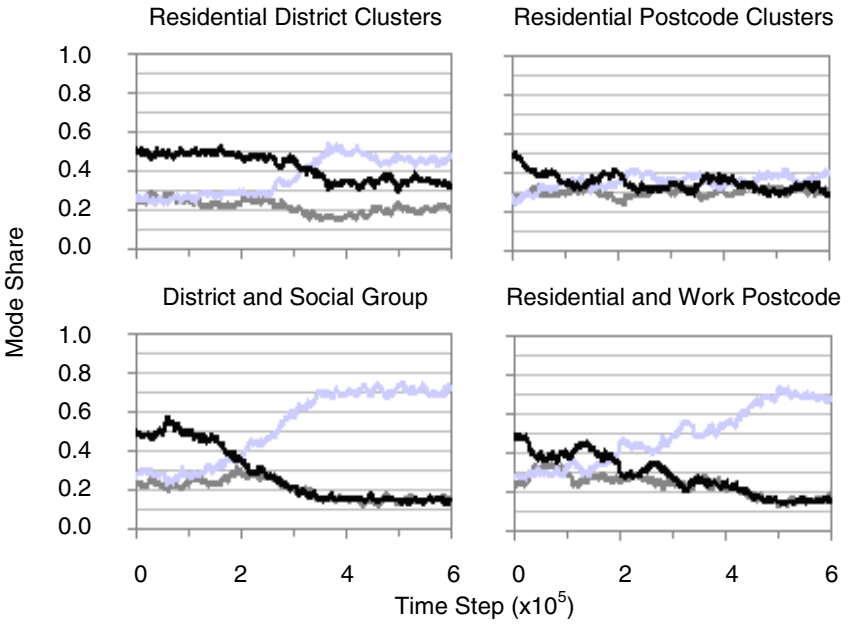


Fig. 3. Example time series for nested logit model with sociogeographic networks ($N = 2913$)

In the example time series in the upper right panel of Fig. 3, we find that the overall behavior across the 67 postcode clusters moves fairly rapidly to a roughly equal split of one-third for each mode. Here the volatility within the clusters is so high that no single mode ever stays dominant for very long and there are so many clusters that the average behavior statistically tends to be non-biased.

Finally, given this behavioral information, we proceed to understand what happens in the empirically most relevant scenarios with overlapping groups, where we have interaction within socioeconomic and spatially defined groups but there is a possibility for transmission of influence across groups. In the example time series in the lower left panel of Fig. 3 with interaction defined by overlapping residential district and social group, we find initial prominence of the high mode share for car as we did in the case of the network of disconnected residential district clusters, but here there is indeed the possibility for eventual coordination across clusters, with the entire sample locking-in to one of the stable equilibria in Table 2. In the example in Fig. 3 we find the most stable equilibrium prevailing with high mode share for bicycle; time series with other random seeds for defining the decision-making order of the agents showed stable equilibrium nr. 2 in Table 2 prevailing with high mode share for car. We never found transit mode prevailing, presumably

due to inability to overcome the initial conditions with low transit mode share. The observed long-run outcomes at time step t= 600,000 when applying different random seeds thus yields a similar picture as the left panel of Fig. 2.

The example time series in the lower right panel of Fig. 3 with interaction defined by the relatively less dense network of overlapping residential and work postcode zones is just as interesting. With the relatively small cluster sizes in this scenario, we find initial statistical tendency towards an overall non-biased split fluctuating around one-third for each mode, similar to the case of the network of disconnected residential postcode clusters. Because of the possibility for transmission of influence across clusters here though, eventually the most stable equilibrium gradually takes over as in the case we saw with the hypothetical fully connected network with initially equal mode shares. Over multiple runs with different random seeds determining the decision-making order we find that all runs indeed went to the dominant equilibrium.

In summary, we have seen: 1) global modal choice behavior in a disconnected network of clustered groups is the weighted average behavior of separate clusters; 2) smaller cluster sizes yield more volatility in our nested logit model with socio-dynamic feedback and as a result, a tendency towards an overall non-biased modal split averaged over many clusters; 3) with sufficient volatility, initial conditions have no significant effect on long-term outcomes; 4) with overlapping groups, influence can spread throughout entire sample; 5) for a giant cluster with sufficient network density and sufficient average degree, the precise connectivity of the network doesn't appear to matter in the long-run, but the initial conditions of the starting mode shares do matter; the emergent distribution of outcomes for overall modal split gives the same picture as the analytically predicted outcomes for a fully-connected network with the given initial conditions; 6) the analytically predicted most stable equilibrium ultimately prevails in connected, sparse network of overlapping clusters; initial conditions of the starting mode shares don't seem to matter here.

4 Recommendations

We have extended previous work on discrete choice with social interactions in important ways. We consider a model where an agent's choice is directly influenced by the proportions of the agent's neighbors, colleagues and/or socio-economic peers making each choice, accounting for common unobserved attributes of the choice alternatives in the error structure. We observe that different sociogeographic networks generate dramatically different dynamics and thus clearly cannot be ignored in any empirical application. Misrepresentation of the appropriate scale at which social influence occurs and of the appropriate network structure can thus yield strongly flawed policy implications when studying social feedback.

Further research is needed to explore systematically more comprehensive utility specifications, including for example the effects of availability of alternatives, alternative specific constants, agent-specific socio-demographic characteristics and agent-specific attributes of choice alternatives. Also very important for any policy

application, particularly for transportation mode choice, would be the introduction of not only positive feedback, but also negative feedback into the model to account for congestion effects in addition to agglomeration effects.

In an application of the agent-based model for policy purposes, it may furthermore be important to scale up the number of agents in the simulation to the actual relevant population size. In the domain of transportation land use planning, simulation on the basis of a realistic number of agents can be critical for understanding congestion on the transportation network. Iterative proportional fitting is an established technique in transportation modeling for generating synthetic populations [6]. An open question however is how to scale up social networks from survey data to a synthetic population. Since we have seen that network size and connectivity do indeed impact emergent outcomes of a discrete choice model with social and spatial interactions, the key importance of recent modeling efforts [7, 8] to depict and understand realistic social networks at the population level in geographic space is underscored.

Acknowledgments. The authors would like to gratefully acknowledge discussion with Harry Timmermans, Theo Arentze, Cars Hommes, Frank le Clercq, Loek Kapoen, George Kampis, József Váncza and András Márkus. Special thanks are also due to Guus Brohm and Nelly Kalfs at the Agency for Infrastructure, Traffic and Transport of the Municipality of Amsterdam, and to Willem Vermin and the High Performance Computing support team at SARA Computing and Networking Services, Amsterdam. The authors claim full responsibility for any errors.

References

1. Ben-Akiva, M.: Structure of Passenger Travel Demand Models. Ph.D. Dissertation. Dept of Civil Engineering. MIT, Cambridge, MA (1973)
2. Brock, W., Durlauf, S.: Multinomial Choice with Social Interactions. In: Blume, L., Durlauf, S. (eds.) The Economy as an Evolving Complex System III. Oxford University Press, New York (2006)
3. Dugundji, E.R., Gulyás, L.: Socio-Dynamic Discrete Choice on Networks: Impacts of Agent Heterogeneity on Emergent Equilibrium Outcomes. Environment and Planning B: Planning and Design 35(6), 1028–1054 (2008)
4. Ben-Akiva, M., Lerman, S.R.: Discrete Choice Analysis: Theory and Application to Travel Demand. MIT Press, Cambridge (1985)
5. Dugundji, E.R.: Discrete Choice on Networks: Nested Logit. In: Graduate Workshop on Computational Economics. Santa Fe Institute, Santa Fe (2003)
6. Arentze, T.A., Timmermans, H.J.P., Hofman, F.: Creating Synthetic Household Populations: Problems and Approach. Transportation Research Record 2014, 85–91 (2008)
7. Arentze, T.A., van den Berg, P., Timmermans, H.J.P.: Modeling Social Networks in Geographic space: Approach and Empirical Application. Environment and Planning A (2012)
8. Hackney, J., Kowald, M.: Exponential Random Graph Models of the Zurich Snowball Survey. Futurenet Workshop: Social Network Analysis in Transport, Manchester, UK (2011)

Tipping Points of Diehards in Social Consensus on Large Random Networks

W. Zhang, C. Lim, and B. Szymanski

Abstract. We introduce the homogeneous pair approximation to the Naming Game (NG) model, establish a six dimensional ODE for the two-word NG. Our ODE reveals how the dynamical behavior of the NG changes with respect to the average degree $< k >$ of an uncorrelated network and shows a good agreement with the numerical results. We also extend the model to the committed agent case and show the shift of the tipping point on sparse networks.

1 Introduction

The Naming Game(NG) has become a very popular model in analyzing the behaviors of social communication and consensus [1]. In this model, each node is assigned a list of names as its opinions chosen from an alphabet S. In each time step, two neighboring nodes, one listener and one speaker are randomly picked. The speaker randomly picks one name from its name list and sends it to the listener. If the name is not in the list of the listener, the listener will add this name to its list, otherwise the two communicators will achieve an agreement, i.e. both collapse their name list to this single name. The variations of this game can be classified as the "Original" (NG), "Listener Only" (LO-NG) and "Speaker Only" (SO-NG) types [2] regarding the update when the communicators make an agreement, and as the "Direct", "Reverse" and "Link-updated" types regarding the way that the two communicators are randomly picked. These variations have different behaviors but can be analyzed in the way. In this paper we mainly focus on the "Original" "Direct" version.

Mean field approach has been applied to the NG and a lot of interesting results have been obtained. They reveal the essential difference of the NG compared with other communication models, such as the voter model, in reproducing important phe-

W. Zhang · C. Lim · B. Szymanski
Department of Mathematics, Rensselaer Polytechnic Institute, 110 8th Street, Troy, New York 12180-3590, USA
e-mail: {zhangw10,limcm,szymansk}@rpi.edu

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 161–168.
springerlink.com

nomena in real world social communications. One of the most significant results is a phase transition at a critical fraction of the committed agents in the network, the tipping point [6]. Above the tipping point, the minority committed agents will persuade the majority to achieve global consensus in a time growing with the logarithm of network size, while below the tipping point, the committed agents would require the time exponential in the network size [7], so practically never, for networks of non-trivial size. However, as most applications of the mean field approximation, these theoretical predictions deviate from the simulations on complex networks especially when the network is relatively "sparse". In many studies, the dynamical behavior of the network given its average degree or the degree distribution is very important.

Recently, a so-called homogeneous pair approximation has been introduced to voter model [5], a model simpler than the NG, which improves the mean field approximation by taking account of the correlation between the nearest neighbors. Their analysis is based on the master equation of the active links, the links between nodes with different opinions. Although it shows a spurious transition point of the average degree, it captures most features of the dynamics and works very accurately on most uncorrelated networks such as ER and scale free networks.

In this paper, we apply this idea to the NG, especially the two-word NG case. Different from the voter model case, there are more than one type of active links, so we have to analyze all types of links including active and inert ones. As a consequence, instead of a one dimensional ODE in voter model case, we have a six dimensional one. We derive the equations by analyzing all possible updates in the process and write it in a matrix form with the average degree $< k >$ as a explicit parameter. The ODE clearly shows how the NG dynamics changes when $< k >$ decrease to 1, the critical value for ER network to have giant component, and converges to the mean field equations when $< k >$ grows to infinity. Then we show the good agreement between our theoretical prediction and the simulation on ER networks. Finally, we show the decrease of tipping point value in low average degree networks, i.e. we need fewer committed agents to force a global consensus in a loosely connected social network. The results of a detailed analysis of this model will be reported in another paper.

2 The Model

Consider the NG dynamics on an uncorrelated network (the presences of links are independent) together with the following assumptions which are the foundation of the homogeneous pair approximation:

1. The opinions of neighbors are correlated, while there is no extra correlation besides that through the nearest neighbor. To make this assumption clear, suppose three nodes in the network are linked as 1-2-3 (there is no link between 1 and 3), their opinions are denoted by random variables X_1,X_2,X_3 correspondingly. Therefore this assumption says: $P(X_1|X_2) \neq P(X_1)$, but $P(X_1|X_2,X_3) = P(X_1|X_2)$. This assumption is valid for all uncorrelated networks (Chung-Lu type network [4], especially the ER network).

2. The opinion of a node and its degree are mutually independent. Suppose the node index i is a random variable which picks a random node. The opinion and degree of node i, are X_i and k_i. Mathematically, this assumption means $E[k_i|X_i] =< k >$, $P(X_i|k_i) = P(X_i)$ and $P(X_i|X_j,k_i,k_j) = P(X_i|X_j)$ where j is a neighbor of i. This assumption is perfect for the networks in which every node has the same degree (regular geometry) and is also valid for the network whose degree distribution is concentrated around its average (for example, Gaussian distribution with relatively small variance or Poisson distribution with not too small $< k >$). We will show later this assumption is good enough for ER network.

In other words, the probability distribution of the neighboring opinions of a specific node is an effective field. This field is not uniform over the network but depends only on the opinion of the given node. For an uncorrelated random network with N nodes and average degree $< k >$, the number of links in this network is $M = N < k > /2$. We denote the numbers of nodes taking opinions A,B and AB as n_A, n_B, n_{AB}, their fractions as p_A,p_B, p_{AB}. We also denote the numbers of different types of links as $\mathbf{L} = [L_{A-A}, L_{A-B}, L_{A-AB}, L_{B-B}, L_{B-AB}, L_{AB-AB}]^T$, and their fractions are given by $\mathbf{l} = \mathbf{L}/M$. We take $\mathbf{L}$ or $\mathbf{l}$ as the coarse grained macrostate vector. The global mean field is given by:

$$\mathbf{p}(\mathbf{L}) = \begin{pmatrix} p_A \\ p_B \\ p_{AB} \end{pmatrix} = \frac{1}{2M} \begin{pmatrix} < k > n_A \\ < k > n_B \\ < k > n_{AB} \end{pmatrix} = \frac{1}{2M} \begin{pmatrix} 2L_{A-A} + L_{A-B} + L_{A-AB} \\ L_{A-B} + 2L_{B-B} + L_{B-AB} \\ L_{A-AB} + L_{B-AB} + 2L_{AB-AB} \end{pmatrix}.$$

Suppose X_i, X_j are the opinions of two neighboring nodes. We simply write $P(X_i = A|X_j = B)$, for example, as $P(A|B)$. We also represent the effective fields for all these types of node in terms of $\mathbf{L}$:

$$\overrightarrow{P(\cdot|A)}(\mathbf{L}) = \begin{pmatrix} P(A|A) \\ P(B|A) \\ P(AB|A) \end{pmatrix} = \frac{1}{2L_{A-A} + L_{A-B} + L_{A-AB}} \begin{pmatrix} 2L_{A-A} \\ L_{A-B} \\ L_{A-AB} \end{pmatrix}$$

$$\overrightarrow{P(\cdot|B)}(\mathbf{L}) = \begin{pmatrix} P(A|B) \\ P(B|B) \\ P(AB|B) \end{pmatrix} = \frac{1}{L_{A-B} + 2L_{B-B} + L_{B-AB}} \begin{pmatrix} L_{A-B} \\ 2L_{B-B} \\ L_{B-AB} \end{pmatrix}$$

$$\overrightarrow{P(\cdot|AB)}(\mathbf{L}) = \begin{pmatrix} P(A|AB) \\ P(B|AB) \\ P(AB|AB) \end{pmatrix} = \frac{1}{L_{A-AB} + L_{B-AB} + 2L_{AB-AB}} \begin{pmatrix} L_{A-AB} \\ L_{B-AB} \\ 2L_{AB-AB} \end{pmatrix}.$$

To establish the ODE for NG dynamics, we calculate the variable $E[\Delta\mathbf{L}|\mathbf{L}]$. It takes a long discussion to consider of all possible communications in this network. We just take one case as example: listener taking opinion A and speaker taking opinion B. The probability for this type of communication is $p_A P(B|A)$. The direct consequence of this communication is that the link between the listener and speaker changes from A-B into AB-B, so L_{A-B} decreases by 1 and L_{B-AB} increases by 1. Besides, since the listener changes from A to AB, all his other related links change. The number of

these links is on average $< k > -1$ (here we use the assumption 2, $E[k_i|X_i] =< k >$). The probabilities for each link to be A-A, A-B, A-AB before the communication is given by $\overrightarrow{P(\cdot|A)}$ (here we use assumption 1). After the communication, these links will change into AB-A, AB-B, AB-AB correspondingly and change the value of $E[\mathbf{L}]$ by

$$(< k > -1) \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \overrightarrow{P(\cdot|A)}.$$

Similarly, we analyze all types of communications according to different listener's and speaker's opinions, and sum these changes into $\Delta\mathbf{L}$ weighted by the probability that the corresponding communication happens and obtain:

$$E[\Delta\mathbf{L}|\mathbf{L}] = \frac{1}{M}[D + (< k > -1)R]\mathbf{L}$$

where D is a constant matrix, matrix R is a function of $\mathbf{L}$, given by:

$$D = \begin{pmatrix} 0 & 0 & \frac{3}{4} & 0 & 0 & \frac{1}{2} \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{3}{4} & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & 0 & -1 & 0 \\ 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & -1 \end{pmatrix}, Q_A = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad Q_B = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix},$$

$$R = \big(\mathbf{0}, \frac{1}{2}[Q_A\overrightarrow{P(\cdot|A)} + Q_B\overrightarrow{P(\cdot|B)}], Q_A[\frac{1}{4}\overrightarrow{P(\cdot|A)} - \frac{3}{4}\overrightarrow{P(\cdot|AB)}],$$
$$\mathbf{0}, Q_B[\frac{1}{4}\overrightarrow{P(\cdot|B)} - \frac{3}{4}\overrightarrow{P(\cdot|AB)}], -(Q_A + Q_B)\overrightarrow{P(\cdot|AB)}\big).$$

Then we normalize $\mathbf{L}$ by the total number of links M and normalize time by the number of nodes N:

$$\begin{aligned} \frac{d}{dt}\mathbf{l} &= \frac{N}{M}E[\Delta\mathbf{L}|\mathbf{L}] = \frac{N}{M}[D + (< k > -1)R]\mathbf{l} \\ &= 2\left[\frac{1}{< k >}D + (\frac{< k > -1}{< k >})R\right]\mathbf{l}. \end{aligned} \tag{1}$$

Now we get the ODE of $\mathbf{l}$ for the NG and $< k >$ is explicit in the formula. In the last line, the first term is linear and comes from the change of the link between the listener and the speaker. The second term is nonlinear and comes from the changes of all the related links. Under the mean field assumption, the first term does not

exist, because there is no specific "speaker" and every one receives messages from the mean field. When $< k > \to 1$, the ODE becomes:

$$\frac{d}{dt}\mathbf{l} = 2D\mathbf{l}$$

which is a linear system. When $< k > \to \infty$, the ODE becomes:

$$\frac{d}{dt}\mathbf{l} = 2R\mathbf{l}$$

If in matrix R we further require $\overrightarrow{P(\cdot|A)} = \overrightarrow{P(\cdot|B)} = \overrightarrow{P(\cdot|AB)} = \mathbf{p}$ and transform the coordinates by $\mathbf{L} \to \mathbf{p}(\mathbf{L})$, the ODE just turns back to the one we have under the mean field assumption [6].

3 Numerical Results without Committed Agents

Next we show some numerical results. Fig.1 shows the comparison between our theoretical prediction (color lines) and the simulation on ER networks (black solid lines). The dotted lines are theoretical prediction by mean field approximation. We calculate the evolution of the fractions of nodes with A, B and AB opinions respectively and show that the prediction of mean field approximation deviates from the simulation significantly while that of homogeneous pair approximation matches simulations very well.

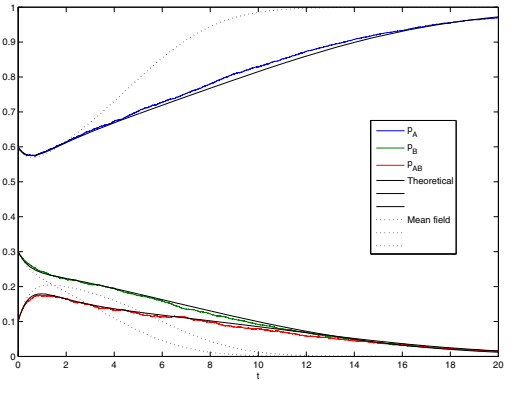


Fig. 1. Fractions of A, B and AB nodes as function of time. The three color lines are the averages of 50 runs of NG on ER network with $N = 500$ and $< k >= 5$. The black solid lines are solved from the ODE above with the same $< k >$. The black dotted lines are from the ODE using mean field assumption.

Fig.2 shows the trajectory of the macrostate mapped into two dimensional space (p_A,p_B), the black line is the trajectory predicted by the mean field approximation. We find that when $< k >$ is large enough, say 50, the homogeneous pair approximation is very close to the mean field approximation. When $< k >$ decreases, the trajectory tends to the line $p_{AB} = 1 - p_A - p_B = 0$, which means there are fewer nodes with mixed opinions than predicted by the mean field. In this situation, opinions of neighbors are highly correlated forming the "opinion blocks", and mixed opinion nodes can only appear on the boundary.

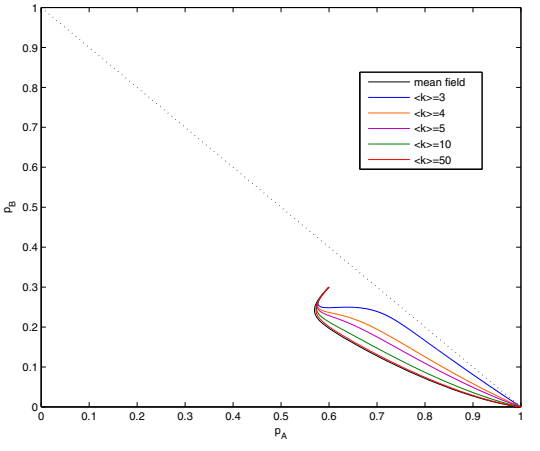


Fig. 2. The trajectories solved from the ODE with different $< k >$ mapped onto 2D macrostate space. When $< k > \to \infty$, the trajectory tends to that of the mean field equation. When $< k > \to 1$, the trajectory get close to the line $p_{AB} = 1 - p_A - p_B = 0$.

4 Committed Agents

Suppose we have p (fraction) committed agents (nodes that never change their opinions) of opinion A, and all the other nodes are initially of opinion B. Is it possible for the committed agents to persuade the others and achieve a global consensus? Previous studies found there is a critical value of p called tipping point. Above this value, it is possible and the persuasion takes a short time, while below this value, it is nearly impossible as it takes exponentially long time with respect to the system sizes.

Similar to what we did in the previous section. We derive the ODE for the macrostate, although the macrostate now contains three more dimensions. $\mathbf{L} = [L_{A-C}, L_{B-C}, L_{AB-C}, L_{A-A}, L_{A-B}, L_{A-AB}, L_{B-B}, L_{B-AB}, L_{AB-AB}]^T$, where C denotes the committed A opinion and A itself denotes the non-committed one. Hence we have a nine dimensional ODE which has the same form as equation 1, but with different details in D and R:

$$D=\begin{pmatrix}
0 & 0 & \frac{3}{4} & 0 & 0 & 0 & 0 & 0 & 0\\
0 & -\frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & \frac{1}{2} & -\frac{3}{4} & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & \frac{3}{4} & 0 & 0 & \frac{1}{2}\\
0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & \frac{1}{2} & -1 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{3}{4} & \frac{1}{2}\\
0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & -1 & 0\\
0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{4} & -1
\end{pmatrix}, Q_A=\begin{pmatrix}
-1 & 0 & 0 & 0\\
0 & 0 & 0 & 0\\
1 & 0 & 0 & 0\\
0 & -1 & 0 & 0\\
0 & 0 & -1 & 0\\
0 & 1 & 0 & -1\\
0 & 0 & 0 & 0\\
0 & 0 & 1 & 0\\
0 & 0 & 0 & 1
\end{pmatrix}, Q_B=\begin{pmatrix}
0 & 0 & 0 & 0\\
-1 & 0 & 0 & 0\\
1 & 0 & 0 & 0\\
0 & 0 & 0 & 0\\
0 & -1 & 0 & 0\\
0 & 1 & 0 & 0\\
0 & 0 & -1 & 0\\
0 & 0 & 1 & -1\\
0 & 0 & 0 & 1
\end{pmatrix},$$

$$R = (\mathbf{0}, \frac{1}{2}Q_B\overrightarrow{P(\cdot|B)}, -\frac{3}{4}Q_A\overrightarrow{P(\cdot|AB)}, \mathbf{0}, \frac{1}{2}[Q_A\overrightarrow{P(\cdot|A)} + Q_B\overrightarrow{P(\cdot|B)}],$$
$$Q_A[\frac{1}{4}\overrightarrow{P(\cdot|A)} - \frac{3}{4}\overrightarrow{P(\cdot|AB)}], \mathbf{0}, Q_B[\frac{1}{4}\overrightarrow{P(\cdot|B)} - \frac{3}{4}\overrightarrow{P(\cdot|AB)}], -(Q_A + Q_B)\overrightarrow{P(\cdot|AB)}).$$

Finally, we show the change of the tipping point with respect to the average degree $< k >$ in Fig. 3. Starting from the state that $p_B = 1 - p$, the ODE system will go to a stable state for which $p_B = p_B^*$. p_B^* is 0 if the committed agents finally achieve the global consensus. The sharp drop of each curve indicates the tipping point transition with the corresponding $< k >$. According to the figure, the tipping point shifts left when the average degree $< k >$ decreases.

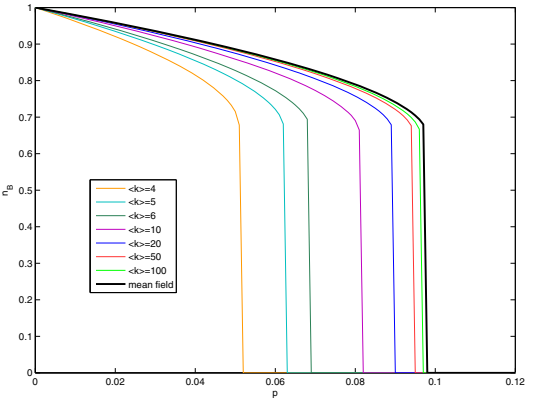


Fig. 3. Fraction of B nodes of the stable point (p_B^*) as a function of the fraction of nodes committed to A (p). The color lines consist of stable points obtained by tracking the ODE of NG on ER for a long enough time. The black lines are the stable points solved from the mean field ODE.

Acknowledgements. This work was supported in part by the Army Research Laboratory under Cooperative Agreement Number W911NF-09-2-0053, by the Army Research Office Grant No. W911NF-09-1-0254, and by the Office of Naval Research Grant No. N00014-09-1-0607. 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 U.S. Government.

References

1. Baronchelli, A., Felici, M., Loreto, V., Caglioti, E., Steels, L.: Sharp transition towards shared vocabularies in multi-agent systems. J. Stat. Mech.: Theory Exp., P06014 (2006)
2. Baronchelli, A.: Role of feedback and broadcasting in the naming game. Phys. Rev. E 83, 046103 (2011)
3. Pugliese, E., Castellano, C.: Heterogeneous pair approximation for voter models on networks. Eur. Lett. 88(5), 58004 (2009)
4. Chung, F., Lu, L.: The Average Distances in Random Graphs with Given Expected Degrees. Proceeding of National Academy of Science 99, 15879–15882 (2002)
5. Vazquez, F., Eguluz, V.M.: Analytical solution of the voter model on uncorrelated networks. New Journal of Physics 10, 063011 (2008)
6. Xie, J., Sreenivasan, S., Korniss, G., Zhang, W., Lim, C., Szymanski, B.K.: Social Consensus through the Influence of Committed Minorities. Phys. Rev. E 84, 011130 (2011)
7. Zhang, W., Lim, C., Sreenivasan, S., Xie, J., Szymanski, B.K., Korniss, G.: Social influencing and associated random walk models: Asymptotic consensus times on the complete graph. Chaos 21(2), 025115 (2011)

Modeling Annual Supreme Court Influence: The Role of Citation Practices and Judicial Tenure in Determining Precedent Network Growth

Ryan Whalen

Abstract. Using networks generated from the entire set of United States Supreme Court decision citations, this paper models yearly court influence as a function of system stability, complexity, precedent age and judicial tenure. The model demonstrates that decisions written in years when the mean judicial age is low and judges are more stable in their use of precedent, more conservative in terms of the age of precedent cited, and the yearly citation network is less complex are more likely to be cited in future years. By incorporating system endogenous variables in modeling efforts, this paper contributes to the development of complex legal systems studies, and proposes new ways to develop the field.

1 Introduction

Every year, the American Supreme Court contributes to its own body of precedent. Justices carefully craft decisions, situating them within the set of extant precedent by citing relevant prior decisions. This process generates a complex dynamic system that grows and changes from year-to-year as the Supreme Court issues more and more decisions generating ever more citation links between them.

Ostensibly, the Supreme Court's authority stems from its role as the judicial system's court of final appeal. Decisions serve as precedent, establishing the state of law for analogous disagreements in the future. This authority derives from the convention of *stare decisis*, assuring that lower courts conform to higher court precedent, while precedents stand as "good law" unless they are for some reason overturned.

Ryan Whalen
Northwestern University, 2240 Campus Drive Evanston IL
e-mail: r-whalen@northwestern.edu

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 169–176.
springerlink.com

Before decisions become precedent judges must of course write them, and while they craft these decisions judges are expected to take into account existing precedent, citing it where relevant. From 1789 to 2004 the court issued approximately 35 000 decisions with over 200 000 citations between them. If we conceive of these decisions as nodes in a network, with their citations joining them together into a "web of law" [1] we can apply network analytic techniques to assess the court's performance.

Given this web of law we can think of a court's influence in terms of how often it is cited in future years. If the court writes important decisions that go on to influence future deliberations, that court will receive more citations than a court which – for whatever reason – writes decisions that are less important in future years. Past research has examined the court in terms of constitutional eras [2,3], interest group activity [4], judicial ideology [5,6], and from a host of other perspectives. The models that these works develop and rely upon tend to use factors exogenous to the precedent system to explain and understand court behavior.

There has been little work that has attempted to model court citation influence year-by-year that incorporates variables both endogenous and exogenous to the citation system. This study fills that gap by modeling court citation patterns from 1800 to 1990 as a function of variables drawn both from the citation system itself and from exogenous historical variation.

System-level analysis of court citations contributes to a growing literature that attempts to apply artificial intelligence to legal analysis [7,8] as well as an increasingly popular and capable field of legal citation analysis [9-11]. This study is amongst the first to use the record of complex Supreme Court behavior to construct new variables that help to explain the system's functioning.

2 The Measures

The citation data used [12] comes from a set provided by Lexis-Nexis and used originally in Fowler *et al*'s [13] analysis of precedent centrality measures. It includes complete data on citations between Supreme Court cases from 1789 to 2005. The data started as a full-network edge list which was then parsed into a complete network. Subgraphs were then generated for every year including all of the decisions written that year and the citations for each.

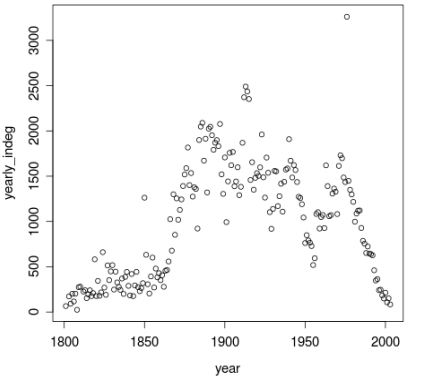


Fig. 1. Yearly in citations

Yearly citations. The dependent variable used below is the total number of citations received by decisions written in each year. Because of the central importance of precedent to the legal system, the number of citations that decisions written in each year go on to receive is a useful proxy for a year's influence. Years

during which many important decisions are written will go on to garner more citations as those decisions are deemed relevant in future years, and thus exert more influence on legal development.

The dependent variable plot (figure 1) shows a fairly low rate of citation for years prior to the civil war. Following the war, we see a sharp increase in the number of citations received per year. By start of the 20th century, the change has largely leveled off and each year receives somewhere between around one and two thousand citations. The mean number of citations received per year is 945 (sd 655), with a few outlying years. For instance, 1976 is situated well above the curve. In this instance the outlying behavior is caused by one particularly influential case - Gregg v. Georgia – that is often cited in reply to the many death penalty appeals that the Court receives.

There is a natural downward curve in more recent years as decisions written during this time period have had fewer opportunities to attract citations. Due to this consideration, the analysis below uses a subset of the data, excluding years prior to 1800 – which saw relatively little Court activity – and years after 1990, which have yet to reach citation maturity.

Precedent age. Mean precedent age was calculated by taking the mean of the age of all precedents cited in a given year. Throughout this paper, I will refer to this mean citation age as the *observed precedent age*. The observed age suggests how conservative or progressive a court is in terms of what precedent it cites – with older precedents suggesting a more conservative court and *vice versa*. However, it is difficult to compare years to one another because, as years go by, there are more old cases for judges to cite and they grow older every year. This creates a natural tendency for observed age to increase over time.

To control for this I also calculated the mean age of all extant precedent for each year. This variable, referred to below as *expected random age,* tells us what mean age we would expect if judges made citations at random. The difference between *expected random age* and *observed precedent age* provides a more nuanced perspective on how conservative or progressive a given court's citation patterns are.

Mean citation age across all years is 18.95 years (sd 7.95), while the mean expected random age is 35.56 years (sd 21.11). The mean difference between expected random age and observed citation age is 16.25 years (sd = 14.63). Looking at the observed mean citation age (figure 2) shows an unstable period prior to the civil war that is similar to what we observe in both the yearly citation and stability plots. Prior to the civil war, age steadily increases until leveling off around the mid 19th century at which point mean age fluctuates between around 20 and 35 years. Meanwhile, we see random age steadily increase and move further and further from the observed age

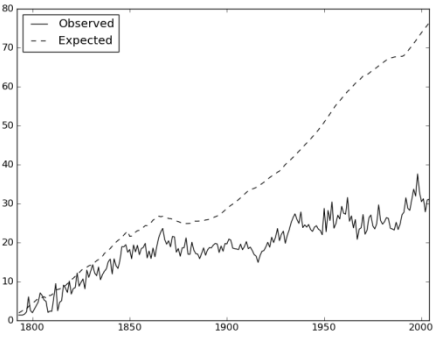


Fig. 2. Citation age

curve. Prior to the mid 19th century, the mean age of Supreme Court citations remained very close to what we would expect to see if citations were made at random. As years go by, and especially after the start of the 20th century, Supreme Court citation age begins to diverge more and more from what we would expect to see given random citations.

Citation stability. Examining the contents of yearly citation subgraphs and comparing them to one another allows us to include measures of system stability in our model. We can consider the set of precedent used in a year as a court's precedent repertoire. We can measure how stable court repertoires are by comparing them to those used in preceding years. Some courts use very similar repertoires to those used in previous years, whereas others use sets of precedent that, for the most part, have not been used in recent years. Periods of changing repertoire denote either changes in the content of cases that the Supreme Court hears or changes in the body of precedent that the Supreme Court feels is good law.

To measure precedent stability we can calculate the proportion of precedent cited in any given year that was also cited in the preceding 5 years. To determine this we simply divide the total number of unique citations in each year by the number of those citations that were also used in the previous five year period. At a value of 1, this *stability* variable tells us that all of the precedent used in a given year was also used at some point during the previous 5 years. Similarly, a 0.5 *stability* level shows that half of a year's precedent was also used at some point during the previous 5 years.

Mean stability across years is 0.52 (sd 0.18). The stability plot (figure 3) shows that prior to the civil war, the court was significantly less stable – and more variable in its stability - in the set of precedent it used. Following the civil war, precedent stability levels off and tends to vary within a narrower range, with anywhere from 40-70% of precedent used in a given year also used in the preceding 5 years. This curve suggests behavior counter to what one would expect. During the Court's early years there was less precedent available for Supreme Court justices to cite. We would therefore expect that – all else being equal – these early years would exhibit *more* stability than the later years during which justices had a much larger body of precedent to draw on. However, we observe just the opposite, with stability increasing concomitantly with the available body of precedent.

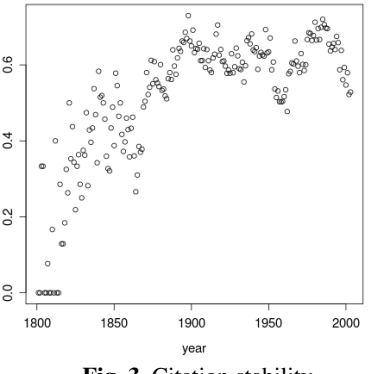


Fig. 3. Citation stability

Components. Yearly subgraphs tend to be made up of many disconnected components. In most situations the citations between case A and the body of precedent it cites and case B and the body of precedent it cites will form two distinct graph components. However, if case A cites case B or *vice versa* or the two cases share precedent, their ego networks will join to form one larger component. We can thus

measure the number of components in each yearly subgraph and use this measurement to infer how connected a year's precedent is. When there are many subcomponents in a given year, cases tend to be isolated from one another, each addressing its own body of precedent. However, when there are fewer components, the decisions within a year are more interrelated and – in a sense – more complex as they are more likely to depend on and interact with one another. The analysis below includes a *components* variable calculated by measuring the number of connected components in each yearly subgraph.

The number of *components* in the yearly citation subgraphs (figure 4) shows a striking inverted-U shaped curve (mean=32.8, sd=21.46). As the 19th century progressed, the Court's yearly citation networks consisted of an ever-growing number of components, until just before the turn of the century when component number peaked and subsequently declined for the duration of the 20th century. Substantively, this means that during the 19th century, yearly decisions became separated into distinct silos, each relying individually on its own body of precedent. During the 20th century we see the reverse, where each year's decisions become more related and are much more likely to rely on one another's findings and share precedent.

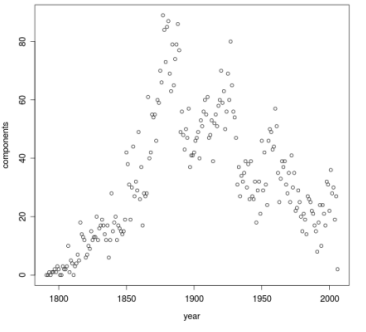


Fig. 4. Graph components

Judicial experience. There are numerous reasons to believe that judicial Supreme Court tenure is significantly related to court influence. Scholars have long noted "freshman effects" [14] [15] for Supreme Court justices. When we measure mean Supreme Court judicial experience we are in a sense measuring how "fresh" an entire court is. Courts with lower mean experience are likely to have worked together less than other courts and they lack the presence of more experienced justices who might serve as stabilizing factors within the court.

To measure judicial experience, the U.S. Supreme Court Justices database [16] was used to calculate mean judicial tenure for each year included in the model. Examining mean judicial tenure on a yearly basis shows quite clear court eras, during which the court builds up judicial experience before one or more particularly long serving justices leave the bench to be replaced by newcomers. Mean judicial tenure across all years is 11.71 (sd 3.48).

Cases: In addition to the above measures, the number of cases written in each year is included as a control variable.

Modeling yearly citations. While the descriptives above provide an interesting perspective on Supreme Court history, this paper's chief priority is to test whether or not variables endogenous to the citation system serve as meaningful predictors of eventual court influence. In order to do so, OLS regression was used

to model yearly citations as a function of stability, judicial experience, the difference between random expected age and real age and the number of decisions written that year. The results (table 1) are discussed below.

Table 1. Dependent Variable = Yearly in citations

	Estimate	St. Error	Beta Coef.	p-value
Intercept	284.21	119.94		0.019
Stability	709.11	222.57	0.188	0.0017
Age_Diff	-8.48	2.86	-0.172	0.0035
Judicial Tenure	-29.12	7.50	-0.144	0.0015
Components	5.95	1.59	0.197	<0.001
Cases	3.97	0.44	0.651	< 0.001
				Adj . R^2=0.75

3 Discussion

Overall the model accounts for a relatively high proportion of variability in yearly citations (adj. R^2=0.75). All of the predictor variables, and the cases control variable, are significant predictors of yearly citations.

Stability. The positive coefficient demonstrates that years which use bodies of precedent similar to those used in the five preceding years are more likely to attract citations. That is to say, the less stable courts are with the body of precedent they use and the more they diverge from precedents that have recently been cited the less influence they have in future years.

Citation age. The negative citation age coefficient suggests that as the difference between mean citation age and expected random age increases (i.e. as courts stray from random expected age by citing more recent decisions) the court becomes less likely to attract citations in future years.

Components. The number of components a year's subgraph has and the number of citations that year goes on to garner are positively related. This could perhaps reflect a preference for less complexity *within* a year's precedent network. The *citation age* and *stability* findings above showed an aversion to change as measured from some baseline established *outside* the yearly subgraph. On the other hand, the *components* finding demonstrates that fewer relationships and less complexity *within* a year's decision network are also related to the number of citations a year's decisions will garner.

Judicial tenure. As the collective experience of a court grows, the decisions they write become less likely to attract future citations. Much of the literature on judicial tenure suggests that judges are more moderate early in their time on the bench. Perhaps this leads to more moderate decisions for courts that are made up of disproportionately short-tenure justices, and perhaps these moderate decisions are more palatable to future justices. Alternately, this effect could be related to the phenomenon of recent case preference that we observe in the citation age plot

(figure 4). It is conceivable that older, more experienced judges are more in touch with older jurisprudence, whereas younger, less experienced judges could be more in touch with newer case law, especially that which they helped create.

Cases. This control variable shows a significant positive relationship with yearly citations, and moreover a relatively large effect size. This is unsurprising as years during which more decisions were written are, all else being equal, more likely to attract citations.

The whole model. Stepping back from an examination of each variable's place in the model, we see a model that is itself significant, not only statistically but also methodologically. While scholars have long advocated for an increased focus on empirical studies of the legal system [i.e. 17], there has as yet been relatively few legal citation analyses. Most of the research done prior to this study has been descriptive in nature, and – to the author's best knowledge – none have used system endogenous variables as elements in an analytic model.

Court evolution. Another strength of this analysis is its ability to provide us with insight into how the Court's behavior has changed over time. Most of our variables demonstrate an "establishment" period prior to the civil war. During this period, court behavior had yet to reach a level of relative stability, showing more variability from year to year. Following the civil war, and especially after the start of the 20th century, we see a court that behaves much differently than it had in its early years. We see much less fluctuation in the set of precedent used, an increasing preference for more recent precedent and much more stability in the number of citations each year goes on to receive.

4 Conclusion

This study demonstrates that we can use citation networks to analyze Supreme Court influence. It shows that precedent stability, citation age, the number of components in a year's citation subgraph and judicial tenure are all significantly related to the number of citations a year's decisions will go on to garner. Years with less experienced Supreme Court justices at the bench, that are stable in regards to the body of precedent they cite, consistent when citing from across the age spectrum of available precedent, and relatively uncomplicated in terms of how many relationships exist between decisions are more likely to attract citations in future years. However, this study's most important contributions are not the substantial conclusions arising from its analysis. Rather, its contribution to the development of a new type of legal analysis variable - derived from measurements endogenous to the precedent citation system – and the demonstration that these variables are meaningful predictors of system behavior, will hopefully inspire similar studies in the future.

Acknowledgments. The author would like to acknowledge the support of National Science Foundation grants: OCI-0904356 and IIS-0838564.

References

1. Thomas, H.F.: The Web of Law. San Diego L Rev. 44, 309 (2007)
2. Ackerman, B.A.: We the people. Belknap Press of Harvard University Press, Cambridge (1991)
3. Burnham, W.D.: Constitutional Moments and Punctuated Equilibria: A Political Scientist Confronts Bruce Ackerman's We the People. The Yale Law Journal 108(8), 2237–2277 (1999)
4. Hansford, T.G., Johnson, K.: Interests and Institutions: The Causes and Consequences of Organized Interest Activity at the US Supreme Court. In: Annual Meeting ofthe American Political Science Association, Boston, MA, pp. 28–31 (2008)
5. Epstein, L., Andrew, M.D., Quinn, K.M., Segal, J.A.: Ideological Drift Among Supreme Court Justices. Northwestern University Law Review 101 (2007)
6. Epstein, L., Segal, J.A., Westerland, C.: Increasing Importance of Ideology in the Nomination and Confirmation of Supreme Court Justices. The Drake Law Review 56, 609 (2007)
7. von der Lieth Gardner, A.: An artificial intelligence approach to legal reasoning (1987)
8. Bench-Capon, T.: Argument in artificial intelligence and law. Artificial Intelligence and Law 5(4), 249–261 (1997)
9. Landes, W.M., Lessig, L., Solimine, M.E.: Judicial influence: A citation analysis of federal courts of appeals judges. The Journal of Legal Studies 27(2), 271–332 (1998)
10. Fowler, J.H., Jeon, S.: The authority of Supreme Court precedent. Social Networks 30(1), 16–30 (2008)
11. Leicht, E.A., Clarkson, G., Shedden, K., Newman, M.E.J.: Large-scale structure of time evolving citation networks. The European Physical Journal B-Condensed Matter and Complex Systems 59(1), 75–83 (2007)
12. Fowler, J.H., Jeon, S.: Supreme Court Citation Network Data (2010), `http://jhfowler.ucsd.edu/judicial.html` (accessed October 7, 2011)
13. Fowler, J.H., Johnson, T.R., Spriggs, J.F., Jeon, S., Wahlbeck, P.J.: Network Analysis and the Law: Measuring the Legal Importance of Precedents at the U.S. Supreme Court. Political Analysis 15(3), 324–346 (2007), doi:10.1093/pan/mpm011
14. Hagle, T.M.: Freshman Effects for Supreme Court Justices. American Journal of Political Science, 1142–1157 (1993)
15. Wood, S.L., Keith, L.C., Lanier, D.N., Ogundele, A.: Acclimation Effects for Supreme Court Justices: A Cross-Validation, 1888-1940. American Journal of Political Science, 690–697 (1998)
16. Epstein, L., Walker, T.G., Staudt, N., Hendrickson, S., Roberts, J.: The U.S. Supreme Court Justices Database. Northwestern University School of Law (2010), `http://epstein.usc.edu/research/justicesdata.html` (accessed November 17, 2011)
17. Ruhl, J.B.: Complexity theory as a paradigm for the dynamical law-and-society system: A wake-up call for legal reductionism and the modern administrative state. Duke Law Journal 45, 849 (1995)

The Effect of Citations to Collaboration Networks

Pramod Divakarmurthy and Ronaldo Menezes

Abstract. In this paper we investigate collaborations in computer science based on the Association for Computing and Machinery (ACM) Digital Library dataset. We have constructed two types of network of collaborations one based on all publications and a second that only considers publications with at least one citation. We compare and measure the metrics for both the networks and we show that there are slight structural changes and significant fluctuations in the ranking of authors.

1 Introduction

A social network represents a set of individuals with relationships among them [11]. In network terms, individuals are represented as nodes and relationships represented with ties (edges) between these individuals. An edge can represent a friendship, or any type of relationship between the individuals linked. In this work, we assume the relationship is based on authors' collaborations in a research paper. We focus on Computer Science collaboration and hence our study is based on the publication data available at Association for Computing and Machinery (ACM) Digital Library which is arguably the most complete database of computer-related publications.

The study of collaboration patterns in concerned with co-authorship; researchers who are connected to one another when they co-authored a paper. The assumption for the social relationship is that if individuals have worked together in a paper then one can safely assume that they know each other. Once the networks are built, Social Network Analysis (SNA) techniques are used to better understand the structure of collaborations between computer science authors.

Studying scientific collaborations networks as well as how papers cite others (citation networks) has become increasingly important. Understanding these networks

Pramod Divakarmurthy
BioComplex Laboratory, Department of Computer Science, Florida Institute of Technology, Melbourne, FL, USA
e-mail: {pdivakar2011,rmenezes}@my.fit.edu

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 177–185.
springerlink.com

help us better understand how scientific discoveries and innovations are communicated within the scientific community. Thousands of research papers are published every year. In most cases, citation counts are used to measure the impact of scientific work in the scientific community. A study by Garfield et al. [8] in 1992, showed that there is a high correlation between citations and Nobel Prize winners. However, there are works that are never cited. This alone makes us wonder what is the effect of these uncited works to the studies done in collaboration networks.

Music studies have always used ”citations”. In music, composers are credited when their song is recorded. Songs that are written but not recorded are not listed in the music databases and can be considered as non-existent. Using the same concept, we have considered the number of citations of each paper in this study. Cited papers imply that the knowledge generated from the collaborations has been used in further studies. Hence, we generate two types of networks of authors: *(i)* one “traditional” where all papers are considered, and *(ii)* citation-based one which considers the citations of paper. In the later network, only papers with one or more citations are used to generate network of authors, while in the former all papers, regardless of their citations counts, are considered. We measure and discuss the characteristics of both the networks and rank authors on both networks.

2 Related Works

Network science is a topic that has gained increasing importance in recent years. However, the study of scientific collaboration has been around for quite some time. For instance, the concept of Erdös number has permeated the mathematical research community for more than thirty years [4]. Erdös number is a measure of the shortest path (geodesic distance) between an author and the well-known Hungarian mathematician Paul Erdös in a co-authorship network.

Recently, studies on large-scale collaboration networks by Newman [10] in the fields of bio-medicine, physics and mathematics, demonstrated that biology is highly collaborative field when compared to mathematics and physics, a result that reflects the laboratory and experimentation needs. Barabási et al. [2] studied the evolution in time of collaboration networks in neuroscience and mathematics. The results showed that collaboration network tend to be scale-free, and that the network evolution is governed by preferential attachment. A similar type of study was performed on the scientific collaboration of the Pacific Asia Conference on Information Systems (PACIS) [6]; the network contained a significantly large main component and it exhibited small-world characteristics. Using visualization techniques and SNA metrics they revealed the structural characteristics of PACIS community and were able to identify influential members. More recent studies have also shown that collaborations are usually local: same University, city, or county and comparatively fewer individuals from different countries [7].

3 Networks and Metrics

A graph is a collection of nodes and edges (connections between nodes). From the point of view of network sciences, a graph is considered to be a network when nodes represent real-world objects and edges represent the link between them. Nodes or edges may have a variety of properties associated with them. For instance, in a social network nodes may represent people and may include different properties such as nationality, gender and age; edges may represent a friendship or other kind of relationship and can have attributes such as the strength of the relationship.

One common type of graph is a *k-partite graph* (called a *bipartite graph* when $k = 2$), in which the vertices are partitioned into k-disjoint subsets, and each edge connects vertices in distinct partitions. Hence a bipartite graph is one in which the nodes can be divided into two sets U and V so that every edge in the graph connects a node in V with a node in U. For instance, a network made of actors-movies may contain nodes representing actors and nodes representing the movies. Bipartite networks are important for us because in our study we have a bipartite network of papers and authors, although we concentrate on the author projection.

The properties of a given network can be described at two levels, global and individual node properties. Global graph metrics describes the characteristics of the entire network, for example the graph's diameter, mean node distance, number of components, cliques, clusters, small-world phenomena, etc. Individual properties relate to the analysis of the properties of network nodes, such as centrality, degree, and position in a cluster. The status of an node is usually expressed in terms of its centrality, i.e. a measure of how central the node is to the network graph. Central nodes have a higher degree of influence in the network. There are however many variations of centralities. The *Degree centrality* of a node is defined as the total number of edges that are adjacent to the node; it measures how many connections authors have to their immediate neighbors in the network. *Closeness centrality* focuses on how close an author is to all other authors in the network. Authors may be well connected to their immediate neighbors but they can be part of an isolated clique. Although, such a node is locally well connected, its closeness centrality may be low. Authors with high closeness scores are likely to receive information more quickly than others. The *Betweenness centrality* of a node determines how often the node is found on the shortest path between any pair of nodes in the network.

The degree distribution express the probability, $p(k)$, that a node in the network will have k connections. It has been observed that in many real networks [9] their degree distribution roughly follows a power law as given by Equation 1, where, c and λ are constants. For most of the real networks $2 \leq \lambda \leq 3$.

$$p(k) = ck^{-\lambda} \tag{1}$$

An important characteristic of some network is the *clustering coefficient*, which is a measure of the ration in which nodes in a graph tend to cluster together. In

co-authorship network, clustering coefficient indicates the how much a node's collaborator has written a paper with one of its other collaborators. The clustering coefficient , C_i, of that node is given by Equation 2, where, m_i is the number of links between the k_i neighbors of i; the clustering coefficient of the entire network is just the average of all C_i over the number of nodes in the network n. As a transitivity measure, clustering is more applicable to social networks but it is also used to identify small-world networks [12] which are expected to have high clustering and short average path lengths.

$$C_i = \frac{2m_i}{k_i(k_i - 1)} \tag{2}$$

4 Building Social Networks from Collaborations

The Association for Computing and Machinery (ACM) is the primary society for computer science professionals. As part of their services, they include a Digital Library which indexes many scientific journals, magazines, conference papers, and books in computer science. In order to perform our study we have gathered information about publications available in the digital library and constructed a dataset. The dataset includes works available in the ACM Digital Library from 1951 to 2011, although the core of the dataset is from 1981 to 2010. Using a Web Crawler, we extracted the information and processed the bibliographic data available for each paper found. Information such as published year, title, authors, citations and subject classification were stored as part of our dataset. We also extracted information of authors who had published a paper by 2010. Additional information about the authors such as, name of author, his affiliation (mentioned in his work), number of papers published, number of citations and publications years were also extracted from website. After all was done our dataset included 62,758 authors who published about 233,464 papers over a span of approximately 60 years (1951-2011).

5 Experiments and Results

Our first analysis looks at what kind of network we are working with. Table 1 shows that the networks we study here can clearly be characterized as small world. The clustering coefficient of the network is very high and indicates that the network is organized in groups of highly collaborative individuals with few connections to the outside of the group. Yet, these outside connections do exist and lead to short paths between nodes, where short is defined as $\ell \approx logn$. Table 1 shows the characteristics of the ACM networks compared to another network available in the literature [5]. The ACM networks display a higher clustering meaning this network has a higher collaborations that form triads (cliques of degree 3). Furthermore, the power-law characteristics is within the expected values for real networks with $\lambda = 2.28$. None of the works mentioned in Section 2 have taken the citations of the papers into

consideration. We generated what we call the ACM Citations network (ACM-C) considering the citations of every paper. Note this is not a network of papers and their citations but a collaboration network which considers the collaboration only for cited papers. There are 141,604 cited papers in our ACM dataset. Table 1 shows a comparison of ACM-C and ACM networks with another real network.

Table 1. Network Statistics. Comparison of the ACM and ACM-C with the measurements taken from the Film-Actors network [1]

	ACM	ACM-C	Film Actor
Nodes (n)	62,758	50,614	449,913
Links (m)	340,962	225,191	25,516,482
Mean Degree (z)	10.86	8.89	113.43
Exponent Power Law (λ)	2.3	2.3	2.3
Average Clustering Coeff. (C)	0.60	0.64	0.20
Average Path Length (ℓ)	4.99	5.47	3.48

Since we are interested in identifying the top collaborators from ACM and ACM-C networks, we decided to rank them. Hubs (nodes with high degree) are important but they represent the level of "collaborativeness" of a researcher at the present time. Yet, one may want to have a more predictive measure for the researchers. Closeness centrality expands the definition of degree centrality by focusing on how close an author is to all other authors in the network and hence may represent the "potential collaborativeness" of the researcher; one may not have a lot of collaborators but the structure of the network makes him a prime candidate to acquire new collaborations. This measure is very important since it may indicate the authors who are in the best position to improve their connectivity (degree centrality). One of the known characteristics of social networks is that triads tend to form from triple; if A knows B who in turn knows C, there is high probability that A will get to know C. Given this property, it should be clear that nodes with high closeness should in general have a higher probability of acquiring new collaborators, hence the "potential collaborativeness".

Table 2 shows the top 10 authors in both networks used here. Many of the authors who are considered hubs in full ACM network remain as hubs in the ACM-C network. However some of them do not, and this is what makes the citation study a better representation; Rick Rand (IBM) appears in the ACM network because it has a high degree. However his connections come from works that have not been cited. We claim that ACM-C work is a better snapshot of "collaborativeness" in the computer science field. We note that rankings for closeness changed significantly from ACM to ACM-C, meaning that the "potential collaborativeness";ACM-C ranks shows "potential collaborativeness" to cited authors instead of all authors.

The results in Table 2 may still have problems when it comes to hubs because a person may have high degree by collaborating with many people who have no collaborators themselves. Since the use of rank approaches such as Pagerank is not

Table 2. Ranking of authors according to ("collaborativeness") and closeness centrality ("potential collaborativeness"). Note that the list of authors for degree remain quite stable (shown in **bold**) when we consider only citations. However the closeness rank changes significantly. This may be an indication that the study could be improved if we worked on the core of the network as in Table 3.

	ACM Network		ACM-C Network	
rank	degree	closeness	degree	closeness
1	**Jack Dongarra**	Martin Jung	**Jack Dongarra**	Ian Petersen
2	**Mingqiang Li**	Norbert Oster	**Alberto Vincentelli**	Andrey Savkin
3	**Ian Foster**	Sven Sohnlein	**Ian Foster**	Robin Evans
4	**Manish Gupta**	Florin Pinte	**Manish Gupta**	Y. Feng
5	**Alberto Vincentelli**	R. Lacher	**Mingqiang Li**	Gordana Felic
6	**Noga Alon**	S. Vasupongayya	**Noga Alon**	Zongru Liu
7	**Luca Benini**	Benjoe Juliano	**Luca Benini**	C. Liu
8	David Maier	David Munson	Ewa Deelman	Praveen Nadagouda
9	**Hector Garcia-Molina**	Stefano Andriani	**Hector Garcia-Molina**	Chien Ta Minh
10	Rick Rand	Daniele Menon	Derek Lieber	Tim Walsh

appropriate in undirected networks, we decided to concentrate on the *core of the network*. We filtered the network by removing edges below a certain threshold, followed by the removal of nodes with degree zero. We increased the threshold until we were left with about 10,000 authors (or about 20% of the network nodes). We then extracted giant component (largest connected subgraph) of the network. The experiments in Table 3 shows the results of the experiments with the core network for ACM and ACM-C.

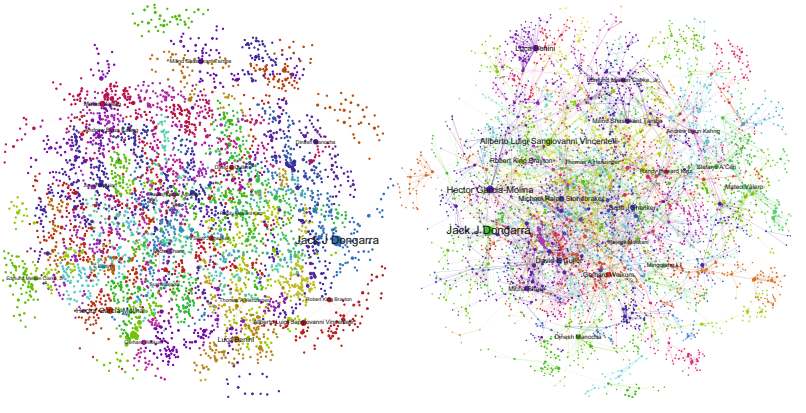
Table 3. Ranking of authors according to ("collaborativeness") and closeness centrality ("potential collaborativeness"). Note that the list of authors for degree is still quite stable when we consider only citations. However we notice that closeness now also remains stable. Authors that appear in two or more rankings are shown in **bold**.

	ACM Network		ACM-C Network	
rank	degree	closeness	degree	closeness
1	**Jack Dongarra**	**Scott Shenker**	**Jack Dongarra**	**Scott Shenker**
2	**Hector Garcia-Molina**	**Hari Balakrishnan**	**Hector Garcia-Molina**	**Christos Papadimitriou**
3	**Luca Benini**	**David Culler**	**Alberto Vincentelli**	**Prabhakar Raghavan**
4	Mateo Valero	**Joseph Hellerstein**	**Luca Benini**	Jeffrey Ullman
5	**Alberto Vincentelli**	**Christos Papadimitriou**	**David Culler**	**Hari Balakrishnan**
6	Andrew Byun Kahng	Ion Stoica	**Michael Stonebraker**	Mihalis Yannakakis
7	Milind Tambe	**Prabhakar Raghavan**	Robert Brayton	Rajeev Motwani
8	**Micha Sharir**	Thomas Anderson	Gerhard Weikum	**Randy Katz**
9	**David Culler**	**Randy Katz**	**Micha Sharir**	**Michael Stonebraker**
10	Thomas Henzinger	Li Feng Zhang	**Scott Shenker**	**Joseph Hellerstein**

Tables 2 and 3 shed light on many issues related to rank of authors. First, the quality of the dataset is essencial to the ranking approach. Although this might sound obvious it may be overlooked because one may think that quality is quantity. We

can see that many of the hubs in Table 2 are not present in Table 3. This is because they were hubs with many connections but their connections were weak (not well established). Once we reduced the network by removing sporadic collaborations their positions in the network were corrected. Second, it is hard to measure "potential collaborativeness". As with anything related to predicting the future, we need to see if the list provided in Table 3 for the ACM-C network confirms to be true in a few years. However we know that social networks tend to form triads and we believe that authors with high closeness are the ones with a high potential because they can form collaborations with many already-established researchers. Last, we can safely assume that computer scientists such as Jack Dongarra (University of Tennessee), Hector Garcia-Molina (Stanford University), Alberto Vincentelli (UC Berkeley) and Luca Benini (University of Bologna) are among the most prolific in computer science; they are hubs, their connections tend to be established (long-term collaborations), and their work is very well cited. Looking at the closeness we immediately see that Scott Shenker (UC Berkeley) appears to have the highest potential to acquire important (cited) collaborations; similar argument can be made for Christos Papadimitriou (UC Berkeley) and Prabhakar Raghavan (Yahoo Labs).

Last, Figure 1 depicts the filtered networks. The size of the node represents the degree of collaboration of those nodes in the network. The color of the node represents the community it belongs to based on the modularity algorithm [3]. Here we use communities just for visualization purposes but it is important to know that the authors forming the communities changes dramatically. We are currently working on a research paper to discuss community formation in these networks.



(a) ACM Network **(b)** ACM-C Network

Fig. 1. Visualization of ACM and ACM-C network. The size of the nodes represents the degree of collaboration of authors and the color represents the community to which they belong. The edge thickness represents the number of collaborations between two authors.

6 Conclusion

In this paper, we studied the network of collaborations from the Association for Computing and Machinery (ACM) dataset. We have constructed a network of collaboration in which nodes are authors who are linked to other authors they have collaborated in a paper we then changed the collaboration rule to include collaborations at least one citation. We compared the metrics for both the network and showed the visualization of the networks. We have ranked the authors based on the number of collaborations in one network and compared their ranking with the other network. Similarly, we have ranked authors based on closeness centrality since it represents a potential metric for "collaborativeness". We showed that there are fluctuations in the ranking when the citations of paper is taken into consideration.

We are collecting more data to make the dataset more complete since currently the core of our dataset is from 1981 to 2010. We will continue to work on the current dataset on many fronts. In particular we want to understand the evolution of the two networks we work on in this paper. One issue that can affect the results we presented relate to the fact that recent publications generally have low citations at first but with time the citations count of a good paper increases gradually, so the removal of early works for lack of citations may be a little unfair. We are currently working on a way to normalize the citations so that early works have a better chance of being considered.

References

1. Amaral, L., Scala, A., Barthélémy, M., Stanley, H.: Classes of small-world networks. Proceedings of the National Academy of Sciences 97(21), 11,149 (2000)
2. Barabasi, A., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., Vicsek, T.: Evolution of the social network of scientific collaborations. Physica A: Statistical Mechanics and its Applications 311(3-4), 590–614 (2002)
3. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment (10), P10,008 (2008), http://stacks.iop.org/1742-5468/2008/i=10/a=P10008
4. Castro, R.D., Grossman, J.W.: Famous trails to paul erdös. Mathematical Intelligencer 21, 51–63 (1999)
5. Chen, Q., Chang, H., Govindan, R., Jamin, S.: The origin of power laws in internet topologies revisited. In: Proceedings of INFOCOM 2002, Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 2, pp. 608–617. IEEE (2002)
6. Cheong, F., Corbitt, B.: A social network analysis of the co-authorship network of the pacific asia conference on information systems from 1993 to 2008. In: Proceedings of Pacific Asia Conference on Information Systems (PACIS 2009), Hyderabad, India (2009)
7. Divakarmurthy, P., Biswas, P., Menezes, R.: A temporal analysis of geographical distances in computer science collaborations. In: IEEE International Conference on Social Computing, pp. 657–660 (2011)

8. Garfield, E., Welljams-Dorof, A.: Of nobel class: A citation perspective on high impact research authors. Atheoretical Medicine 13, 117–135 (1992)
9. Newman, M.: The structure and function of complex networks. SIAM Review, 167–256 (2003)
10. Newman, M.: Co-authorship network and pattern of scientific collaboration. Proceedings of National Academy of Sciences 101, 5200–5205 (2004)
11. Wassernman, S., Faust, K.: Social network analysis: Methods and applications, structural analysis in social sciences. Cambridge University Press, New York City (1994)
12. Watts, D., Strogatz, S.: Collective dynamics of small-world networks. Nature 393(6684), 440–442 (1998)

Network Analysis of Software Repositories: Identifying Subject Matter Experts

Andrew Dittrich, Mehmet Hadi Gunes, and Sergiu Dascalu

Abstract. A software developer joining a large software project faces a steep learning curve before they are able to make real contributions. One challenge is finding the subject matter experts who can answer questions about a specific area of the software or to review changes. This is especially true of large projects with many modules and a large number of authors. In this paper, we describe a method to model a software project as a network using information mined from the project's version control repository, and demonstrate how network analysis techniques can be used to identify the key authors and subject matter experts. We investigate metrics that can be gathered using network analysis, such as which groups of authors typically work together, and how closely knit the developers are on a project. We analyze several specific projects to demonstrate the applicability of these techniques and several hundred projects to show general trends.

1 Introduction

A new developer starting on a large has a lot to learn before they can be a productive member of the team. The project contains many different modules, each of which can be complex on its own. Typically, a junior developer will turn to a more senior developer to ask questions, and to gain insight into the overall architecture of a project. However, it can be difficult to identify experts for a particular area. A good candidate to start with is the person who last modified a file in a module, but this person may have just fixed a formatting problem or a compiler warning, and might not be the best person to ask.

Identifying the most experienced author for a specific area of the project is also a problem for project managers. If a bug is found in a specific module of a large software project, then ideally, the most experienced developer in that area of the project

Andrew Dittrich · Mehmet Hadi Gunes · Sergiu Dascalu
University of Nevada, Reno
e-mail: andy.dittrich@gmail.com, {mgunes, dascalus}@cse.unr.edu

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 187–198.
springerlink.com

should be assigned to fix it. Unfortunately, there is not an easy way to identify that individual. If the manager has been working on this project for a while, then they most likely have the experience to know who the key developer is in this area. Alternatively, they can survey the team members to find someone who is familiar with the area of the code in question.

A project manager may also be interested in how the development team works together. If each developer works on a separate part of the project, and there is no overlap in responsibilities, then there is increased organizational risk from team members leaving the organization. A manager can mitigate this risk by analyzing which members work together and organizing the team such that there is more overlapping knowledge [12]. This risk is difficult to quantify, as there are limited methods for measuring team cohesiveness.

Researchers have investigated collaborative networks to understand different aspects of collaborations [8]. This paper proposes modeling the version control repository as a network, and applying network analysis techniques to identify the key authors for the project and to measure team cohesiveness.

The next section discusses related work. Section 3 discusses how data can be gathered from a source control repository. Section 4 discusses how network analysis techniques are applied. Sections 5 and 6 discuss the results of this analysis on some specific projects, and general trends resulting from the analysis of a few hundred projects. Section 7 analyzes the results. Section 8 concludes the article and suggests future research in this area.

2 Related Work

There are many metrics that can be used to analyze a software project, but there are very few metrics to identify key authors. Commonly used metrics include defect rate, complexity, test coverage, and productivity [10]. These metrics are rarely used to judge a specific author. Associating software metrics with specific authors can cause authors to feel threatened, and is not recognized as a best practice in industry [13]. Hence, typical software metrics are not available to solve this problem.

Other techniques have been developed to identify individuals familiar with specific areas of software. One such method is described by Linstead et.al. [6]. This method searches the source code for keywords or topics, and associates authors with the topics based on the history contained in the revision control repository. This method is able to identify an author who is familiar with a particular topic in the source code. Based on their results, this method is effective in identifying subject matter experts for specific areas of code. However, this method does not consider which authors are the core developers for the overall project, and does not take advantage of the existing relationships between authors that are available in the version control repository.

Another network analysis method is described by Lopez-Fernandez et.al. [7]. This method mined open source version control repositories to identify networks of authors and gain insight into the overall structure of a group of developers. The

approach connects two authors if they have contributed to the same module, and produces an author network. General data is then gathered from this network in order to characterize the project overall.

Huang et.al. describe a similar analysis technique where an author network is created using data from a source control repository [3]. Authors are connected if they have worked on a file in the same directory. The resulting author graph is analyzed using distance centrality to separate the network into kernel and peripheral developers.

Several articles have been written on methods for gathering data from source control repositories. Voinea et.al. describe a framework for querying CVS repositories, parsing the data, and analyzing it. They also propose a method to visualize the resulting data to highlight patterns in the development of a project, such as changes in the development team over time [14] [15]. Kagdi et.al. describe a method to recover the ordered sequence of changed items in a Subversion repository using several heuristics, and a method for analyzing the results [5]. These advanced repository mining techniques were not required for this study. The data needed from a repository can be easily obtained from a log file and converted into a graph for further analysis, as described in the next section.

Several studies have been performed to identify connections between members of networks. Extensive analysis of authors of academic papers has been performed to identify relationships between authors and author groups [9]. This analysis takes a similar approach by linking authors that worked on the same paper, and using this information to create an author network. This differs slightly from the software collaboration. Coauthorship of an academic paper consists of multiple authors working together at the same time, whereas two developers may work on the same source code at separate times with less collaboration between authors.

3 Gathering the Data

The first step in this analysis is to create a bipartite graph that links each unique author with the files that they changed. One set of vertices in the graph are authors and the other set are files. An edge is created for each author that changes a file. For a basic analysis, each edge has equal weighting.

The primary source for this data is a log file produced from the version control system. In this case, projects using Subversion were analyzed, and an xml-format log was produced containing details on every modification to every file in the repository. Other projects using other version control systems such as Mercurial, Git, CVS, or Perforce could have also been used.

The log data was analyzed to produce a list of author-file pairs for each author that made a change to each file. This was interpreted as an edge list for a graph that represents the repository, i.e., a bipartite graph where one type of vertex represents a file and another type of vertex represents an author. A subsection of the bipartite graph produced for the open source Audacity project is shown in Figure 1a.

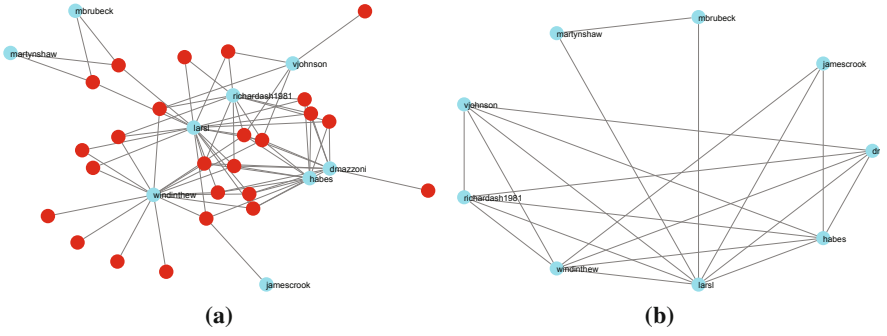


Fig. 1. Graph representations of the Audacity project. (a) A portion of the bipartite graph representing the Audacity project. Authors are blue and files are red. (b) The author graph resulting from the projection of the bipartite graph.

In some cases, only a subset of the repository needs to be analyzed. To accomplish this, the input data can be filtered to get more specific results using several methods. This can be done by analyzing a log for only one section of a repository, e.g., a single folder or module. The data could also be filtered by file name filters, e.g., '*.cpp'.

Another problem is with the initial addition of files. The author that adds a file will be associated with that file, even if they are not an expert in that area. To avoid this, the initial addition of files can be ignored, and only modifications are considered in the analysis.

Filtering the data by the date may also be necessary for long-term projects. Over many years of development, authors will tend to make connections to each other by working on the same file. This may give the incorrect impression that a development team works closely together while certain individuals might have never met. Filtering the data by a specific time period will avoid this problem.

4 Analyzing the Data

The graph that was produced by analyzing the data is a bipartite graph with edges between authors and the files that they modified. This can be projected into two undirected one-mode graphs that show the relationships between authors and the relationships between files separately. The one-mode author graph produced from this projection has a vertex for each author, and an edge connects two authors if they made changes to the same file. This one-mode graph represents the network of connections between authors. For instance, the author graph resulting from the projection of the graph in Figure 1a is shown in Figure 1b.

The core developers for a project can be identified by analyzing the author graph. These are the authors that are the most connected in the author graph. If an author is

well connected, then it indicates that they have worked on many different files with many different authors, and most likely have a wide range of knowledge in the area.

To measure how well connected an author is, we can check the centrality of the author. Measuring the degree of the author is a straightforward way to measure. However, this ignores the degree of the other authors to which this author is connected. If an author has connections to others with many connections, then this can indicate that the author works with other important authors, and should have a higher weight. Another measure of centrality that takes this into consideration is the eigenvector centrality. Measuring the eigenvector centrality for an author is a good indication of how well connected this author is in the author network, and can be used as a proxy to find the experts in this area.

It is important to know which authors typically work together on a software project. This is very useful for a project manager when assigning resources to a specific project. Authors who have a history of working well together tend to make a more productive team than those who don't. Hence, identifying these authors might be beneficial. One way to do this is to identify the communities of authors in the author graph. There are several algorithms for doing this. The algorithms that gave the best results in our analysis were the greedy method, the modularity maximization method, and the spinglass method.

The greedy method is a very simple algorithm that runs in $O(n \log^2 n)$ time, and is well suited for extremely large networks [9]. This algorithm is implemented in iGraph's community_fastgreedy method [4]. The projects analyzed in this study had between 4 and 158 authors, so the simplicity of the greedy algorithm was not necessary, and more complex algorithms could be explored.

The modularity maximization method is discussed in Newman [9]. This method breaks the network into communities such that the total modularity of the network is maximized. This algorithm is also implemented in iGraph's community-leading-eigenvector method [4]. This algorithm resulted in many small communities. Hence, it may be useful if identifying small teams of programmers to work together or for pair programming.

The spinglass method is a complex algorithm that simulates the cooling of a hot system into a grounded state [9]. It associates negative modularity with the energy of an infinite range spin glass and attempts to minimize the energy of the system to find communities [11]. This algorithm is implemented in iGraph's community-spinglass method, as well [4]. This method produced a few large communities in the projects analyzed, and seemed to give the best results among all methods.

The communities determined by either of the methods can be used to set up the optimal team structure for a project by selecting people with a collaboration history for new projects. Alternatively, a project manager could pick people from different communities to encourage cross-team cooperation.

The author graph can be analyzed to determine how the authors work together. Ideally, each author would be connected to each other author. This would indicate that every author had worked together with every other author, and there are at least two authors familiar with every file. So the risk of losing a key employee would be mitigated because there is always a backup who is familiar with the code.

Table 1. Top 10 authors as measured by centrality metric

Audacity		Subversion		Super TuxKart	
1.000	richardash1981	1.000	cmpilato	1.000	cosmosninja
0.971	dmazzoni	0.998	maxb	0.976	hikerstk
0.969	llucius	0.997	kfogel	0.955	auria
0.968	vjohnson	0.995	hwright	0.924	mbjornstk
0.964	jamescrook	0.995	dlr	0.896	coz
0.964	msmeyer	0.994	blair	0.880	hiker
0.954	mchinen	0.992	julianfoad	0.805	grumbel
0.953	windinthew	0.991	brane	0.791	thebohemian
0.951	martynshaw	0.991	ehu	0.791	scifly
0.947	mbrubeck	0.989	sussman	0.744	donconso

This can be measured by the transitivity or clustering of the author graph. A high clustering coefficient indicates that many authors are connected, and a low clustering coefficient indicates that authors typically work alone.

5 Results for Specific Projects

The analysis methods described above were applied to three specific open source projects, namely, Audacity, Subversion, and Super TuxKart.

Audacity is an open source audio editing program. The Audacity project was first hosted on SourceForge in May of 2000, and has 60 unique authors, 9450 unique files, and 24,377 modifications connecting them. The core developers identified by the analysis techniques described above are shown in Table 1. The results were confirmed based on developer credits available on the Audacity website, and indicates that this technique can identify the core developers of the project.

Communities of developers in the Audacity project were identified using the spinglass technique as in Figure 2. It is difficult to verify that these communities are accurate without knowledge of the developers or experience working on this project.

The Subversion project started using Subversion for source control (self-hosting) in August of 2001 [1], so there is an extensive history consisting of 158 unique authors, 6752 unique files, and 92,775 modifications connecting them. The core developers identified are shown in Table 1. Again, these results were confirmed using information available on the Subversion website.

Three communities of developers were identified using the spinglass algorithm to analyze the entire subversion project. Due to the lengthy history of the project, the results are difficult to interpret. Over such a long time, it is likely that many developers would develop connections to many others. For example, if a single file has a history of 20 revisions, then an author could potentially make 20 connections when this file is changed. This leads to a highly connected author graph without clearly defined communities.

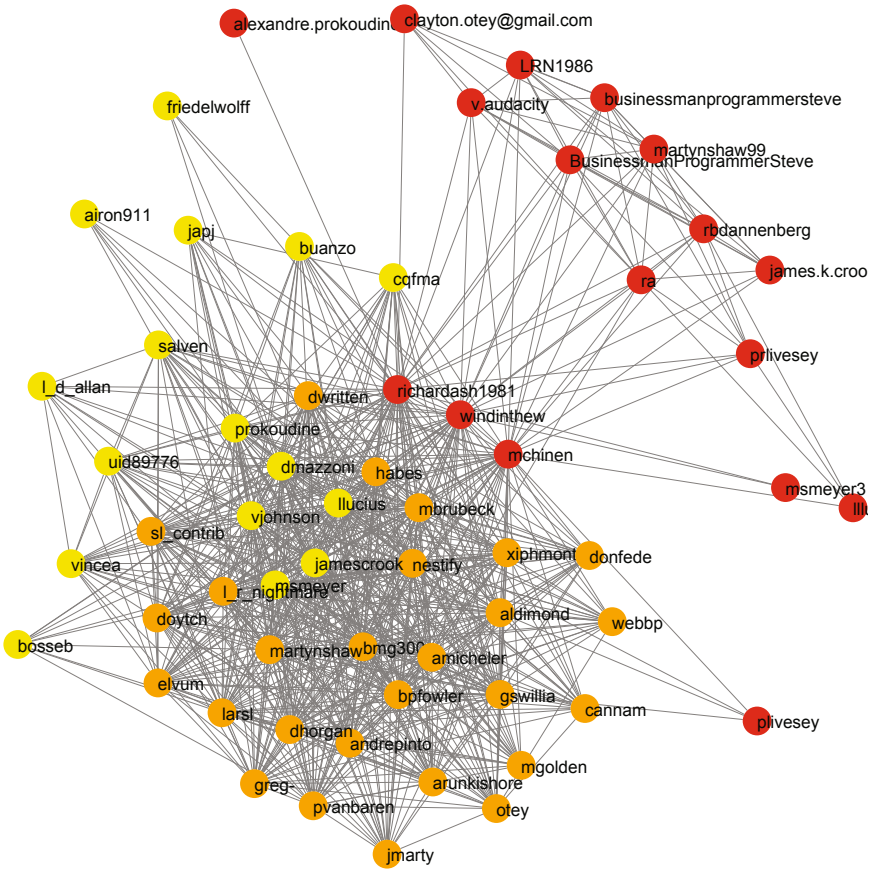


Fig. 2. Communities of authors in the Audacity project

Super TuxKart is an open source multiplayer racing game. It was based on TuxKart, and became Super TuxKart in 2006. The project consists of 36 unique authors and 12,597 unique files. The core developers identified by the analysis techniques described above are shown in Table 1. These results were verified based on the list of core developers listed on the Super TuxKart website, as well.

6 General Results

This analysis technique was automated and applied to a large number of projects. These results can be used as a guideline to see how a specific project's metrics compare to projects in general.

Open source projects from SourceForge were analyzed in bulk. The top 10,000 projects as measured by weekly downloads were gathered. Of these 10,000 projects, 5,031 used a Subversion repository. 1,063 of these repositories had no data available

and were discarded. Many of the remaining projects had only a few authors, which would not give meaningful results. In order to get meaningful data about the relationships between authors, a minimum of 15 authors was chosen. This eliminated 3,665 projects, which left 303 projects for this analysis. This included projects with up to 158 authors and between 377 and 192,121 files.

Each project was analyzed to identify the core developers, author communities, and clustering coefficient for the author graph. The clustering coefficient was of the most interest. The distribution of the clustering coefficient for the author graphs in Figure 3 reveals that most projects have a clustering between 0.7 and 0.9. Audacity had a clustering of 0.783, Subversion had a clustering of 0.880, and Super TuxKart had a clustering of 0.626.

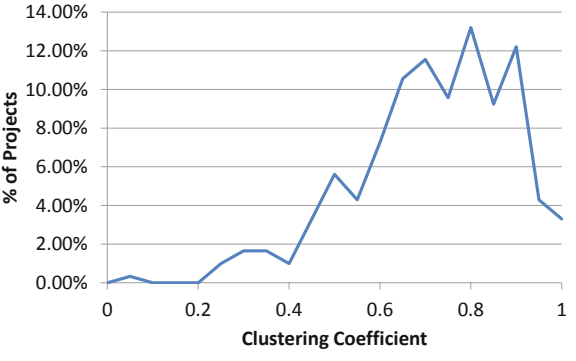


Fig. 3. Clustering coefficient distribution of all projects

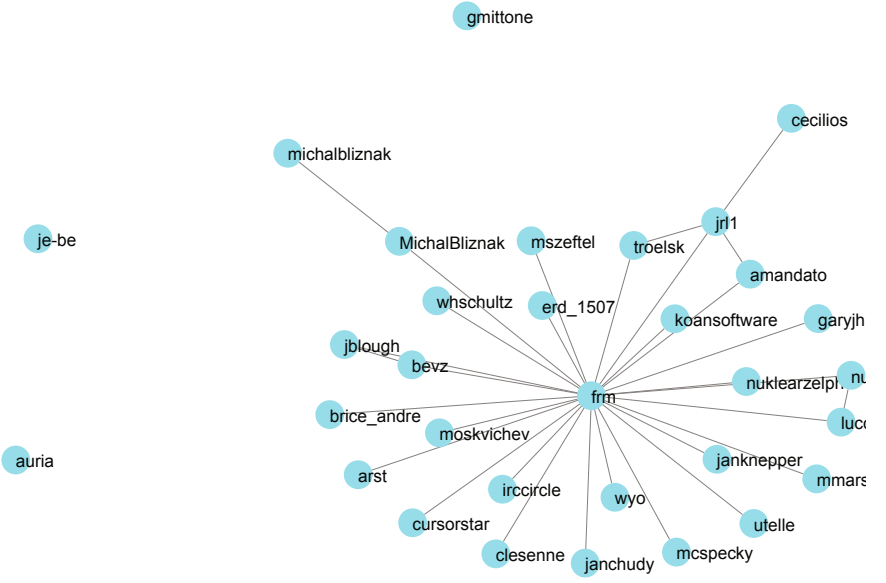
The project with the lowest clustering was wxCode with a clustering of 0.036 and its author graph is shown in Figure 4a. This project is a collection of add-on components and libraries for use with wxWidgets. Each component is separately maintained by a different author, which explains why the authors in this project typically don't work together.

The project with the second lowest clustering was Axiom 3D Engine with a clustering of 0.208 and its author graph is shown in Figure 4b. This project is a cross platform 3D rendering engine, and has 17 authors for 20,890 files. The low clustering coefficient indicates that the authors typically don't work together, which makes sense considering that there are a few authors and many files.

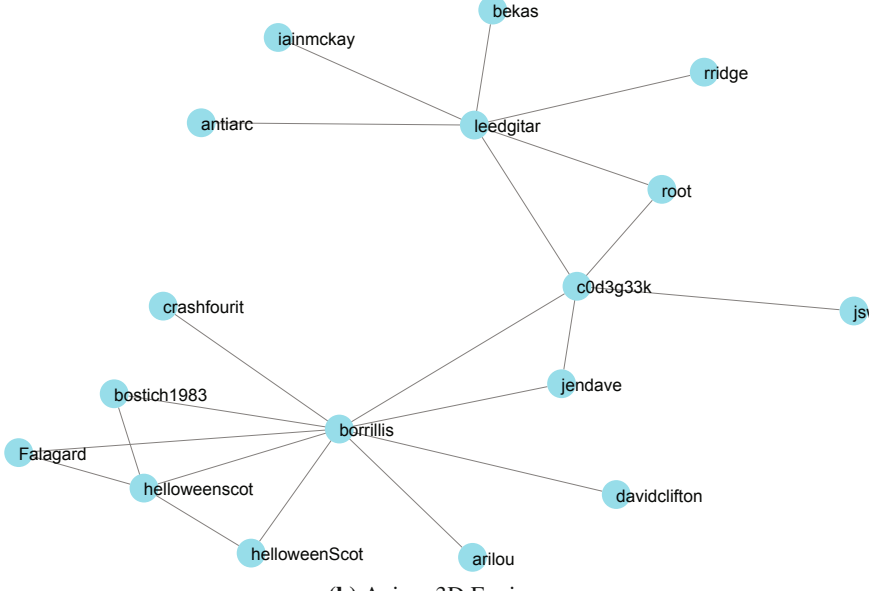
The project with the highest clustering was pkgbuild with a clustering of 0.991 and its author graph is in Figure 5a. This project is a tool for building Solaris SVr4 or IPS packages, and has 70 unique authors for 4,145 unique files. Another example of a project with a high clustering coefficient is MegaMek with a clustering of 0.980 and its author graph is in Figure 5b. This project is an online version of the BattleTech board game, and has 31 unique authors for 10,735 unique files.

7 Analysis

There are several things that can impact the results from this analysis. Our method assumes that the changes made by each author are relevant to the file being modified.

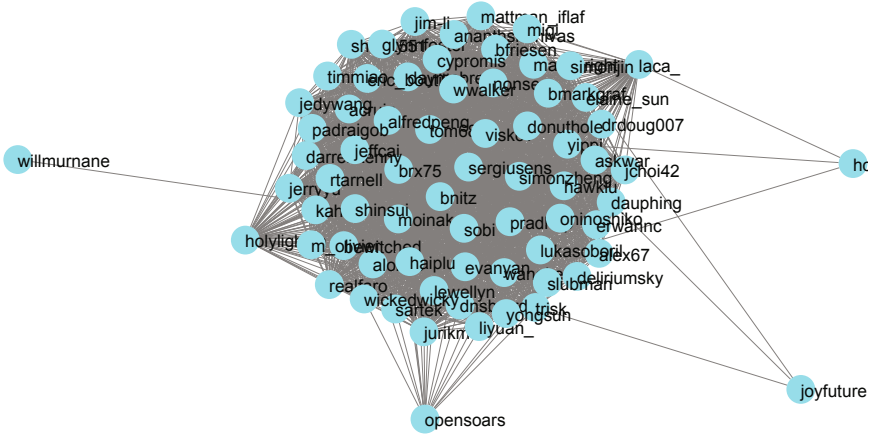


(a) wxCode

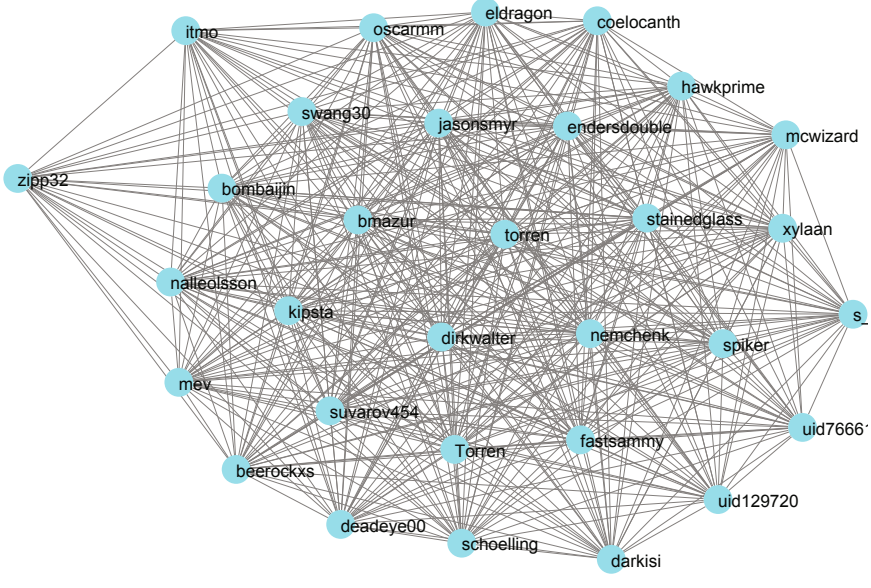


(b) Axiom 3D Engine

Fig. 4. Author graphs with the lowest clustering coefficient. (a) wxCode project with a clustering of 0.036. (b) Axiom 3D Engine project with a clustering of 0.208.



(a) pkgbuild



(b) MegaMek

Fig. 5. Author graphs with the highest clustering coefficient. (a) pkgbuild project with a clustering of 0.991. (c) MegaMek project with a clustering of 0.980

This is not always true. An author could make a change to the formatting of a file or correct a typo in a comment. This author would then be linked to that file and all of the authors who had modified it previously. This should not be a common occurrence, but it has the potential to affect the results of the analysis and make some authors seem to be core developers when they really are not.

A developer's personal software practice could also skew the results. If an author makes many small changes to a file, then their connection to that file will have a higher weight than an author who makes one large change. If edge weights are ignored, then this is not a problem, but that could also skew the results towards authors who made a single change.

Another potential problem is anonymous contributions or multiple online identities for the same person. The projects analyzed in this study did not allow anonymous users to change the source code. If a project did allow anonymous users, then there would be a disproportionate number of changes associated with the anonymous user, and this anonymous user could appear to be a core developer, even though it represents many unique individuals in reality. To protect against this, each anonymous user should be considered as a separate developer [2]. Similarly, one person may have several different online identities that are used to make changes, which could prevent this person from being identified as a core developer, or even put that person in multiple developer communities.

8 Conclusion and Future Work

It can be difficult to identify the subject matter experts for a software project or module within a project. Several techniques have been explored in the past to extract software metrics from a version control repository, and each is specific to the data being sought. This paper describes a network analysis technique that can be used to accurately identify the core developers for a specific software project, and measure how often the developers work together on the same area of code. The analysis was performed on 303 open source projects. Specific details were presented for 3 of these projects, and the general trends were identified based on the analysis of 303 projects. The accuracy of this analysis was confirmed based on credits and other information available on the project websites. Information related to communities of authors within a project was difficult to verify.

The information gathered from this analysis is useful for a new developer in order to identify subject matter experts to answer their questions, and for a project manager when assigning resources. The clustering coefficient of the author graph is a useful indicator for a project manager. If the clustering is too low, then there may be increased risk of key team members leaving the organization. The distribution of clustering coefficients of all projects can be used by a project manager as a basis of comparison.

This analysis should be expanded in the future to attempt to improve the accuracy of the results and to obtain more insight into the project structure. One area that can be explored is how the data is filtered. This study allowed a user to filter the data by file name and to exclude the original addition of files to the repository. The ability to filter by a time period would be useful to limit the analysis for long-lived projects. Other filtering techniques could be developed to limit the analysis to only a certain set of authors, or files containing certain text.

Another area of future research is analyzing how the author graph changes over time. As new developers start work on the project, how do they get incorporated into the author network, and how to older developers transition away from a central role? This may offer insight into the team dynamics for a project and indicate how accepting they are of new developers.

References

1. Collins-Sussman, B., Fitzpatrick, B.W., Pilato, C.M.: Version control with subversion (2007), http://svnbook.red-bean.com/en/1.4/index.html
2. Howison, J., Crowston, K.: The perils and pitfalls of mining sourceforge. In: Proceedings of the International Workshop on Mining Software Repositories (MSR 2004), pp. 7–11 (2004)
3. Huang, S.K., Liu, K.M.: Mining version histories to verify the learning process of legitimate peripheral participants. In: Proceedings of the 2005 International Workshop on Mining Software Repositories, MSR 2005, pp. 1–5. ACM, New York (2005)
4. The igraph website (2010), http://igraph.sourceforge.net/
5. Kagdi, H., Yusuf, S., Maletic, J.I.: Mining sequences of changed-files from version histories. In: Proceedings of the 2006 International Workshop on Mining Software Repositories, MSR 2006, pp. 47–53. ACM, New York (2006)
6. Linstead, E., Rigor, P., Bajracharya, S., Lopes, C., Baldi, P.: Mining eclipse developer contributions via author-topic models. In: Fourth International Workshop on Mining Software Repositories, ICSE Workshops, MSR 2007, p. 30 (2007), doi:10.1109/MSR.2007.20
7. Lopez-Fernandez, L., Robles, G., Gonzalez-Barahona, J.M.: Applying social network analysis to the information in cvs repositories. In: Proceedings of 26th International Conference on Software Engineering, ICSE 2004 (2004), doi:10.1109/ICSE.2004.1317529
8. Newman, M.E.J.: Coauthorship networks and patterns of scientific collaboration. Proceedings of the National Academy of Sciences of the United States of America 101(suppl. 1), 5200–5205 (2004), http://www.pnas.org/content/101/suppl.1/5200.abstract, doi:10.1073/pnas.0307545100
9. Newman, M.E.J.: Networks an Introduction. Oxford University Press, New York (2010)
10. Ordonez, M., Haddad, H.: The state of metrics in software industry. In: Fifth International Conference on Information Technology: New Generations, ITNG 2008, pp. 453–458 (2008), doi:10.1109/ITNG.2008.106
11. Reichardt, J., Bornholdt, S.: Statistical mechanics of community detection. Phys. Rev. E 74(1), 016,110 (2006), doi:10.1103/PhysRevE.74.016110
12. Sommerville, I.: Software Engineering, 8th edn. Addison-Wesley, Harlow (2007)
13. Umarji, M., Shull, F.: Measuring developers: Aligning perspectives and other best practices. IEEE Software 26(6), 92–94 (2009), doi:10.1109/MS.2009.180
14. Voinea, L., Telea, A.: Mining software repositories with cvsgrab. In: Proceedings of the 2006 International Workshop on Mining Software Repositories, MSR 2006, pp. 167–168. ACM, New York (2006)
15. Voinea, L., Telea, A.: An open framework for cvs repository querying, analysis and visualization. In: Proceedings of the 2006 International Workshop on Mining Software Repositories, MSR 2006, pp. 33–39. ACM, New York (2006)

The Social Structure of Organ Transplantation in the United States

Srividhya Venugopal, Evan Stoner, Martin Cadeiras, and Ronaldo Menezes

Abstract. As of today, 110,629 Americans are waiting for an organ transplant yet in 2010 only 28,664 people received organ transplants. This fact alone demonstrates that the country is facing a shortage of organs. Numbers such as these make it absolutely clear that we need to be looking for improvements in the organ allocation system in the USA. Before one starts proposing new allocation systems, it is crucial to understand the structure of the current system. In spite of availability of data on transplants, to our knowledge, no proper analysis has been done using the data. This paper looks at this data and what it may reveal about the allocation process currently in place. In order to structure the data we used techniques from network sciences to create a network of locations (henceforth called a geographical social network) representing all the transplants in the USA since 1987 where nodes represent states in the USA. This "social structure" is then analyzed using techniques from network sciences to bring clarity to the organ donation process.

1 Introduction

End-stage organ failure is a major public health concern. With limited treatment alternatives, transplantation has become the best option for people with

Srividhya Venugopal · Evan Stoner · Ronaldo Menezes
BioComplex Laboratory, Department of Computer Sciences,
Florida Institute of Technology, Melbourne, Florida, USA
e-mail: {svenugopal2010,stoner2010,rmenezes}@my.fit.edu

Martin Cadeiras
Division of Cardiology, David Geffen School of Medicine,
University of California Los Angeles, California, USA
e-mail: mcadeiras@gmail.com

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 199–206.
springerlink.com

failing organs. Consequently, maintaining an adequate supply of donor organs has become the main goal of transplantation programs. While the number of organ donors has increased by less than 5% per year on average, demand has grown by almost 20%. Currently, more than 100,000 Americans are waiting for an organ transplant, but unfortunately less than one-third will receive an organ killing nearly 20 people each day[11].

Based on this data, it is evident that organ transplantation becomes a life saving option only for a minority. The long list of people waiting for an organ also suggests that there is an urgent need to reformulate the way organs are allocated and identify possible alternatives to the current system.

How to best reduce this deficit and make transplants available to the largest-possible population has been the question of many researchers and organizations dedicated to the study of organ transplantation. Most studies have concentrated in understanding the effect of allocation policies to consequently modify the existing rules. With the availability of transplantation data, one way to proceed could be not only to analyze more data, but to improve the way we extract *knowledge* from the existing data. Thus, in this paper, we construct a geographical social network (GSN) where nodes represents states, based on the donor-recipient relationship and use techniques from network sciences to analyze the network formed for many organs.

An important concept we adopt in this paper is imagining an organ as a commodity that flows between patients and communities of disparate social, biological and geographical backgrounds and spreads in response to the growing demand of patients suffering from end-stage organ failure. Keeping this in mind, we model the sharing of this commodity as a network.

The study of the structure of the network underlying the organ transplant system is a robust methodology that has been exploited over the past decade in many science fields [3]. In this paper we apply network concepts to the field of organ transplantation to help us answer questions related to the structure and allocation of organs in the USA.

Common to understanding networks is their derivation from information collected about a system. In the United States, the United Network for Organ Sharing (UNOS) has been created to support, coordinate, and promote solid organ transplantation, following the rapid expansion of the field as transplantations become safer. The study of the information contained in the UNOS database by applying concepts of social networks is expected to follow principles that are well established and fundamental to systems such as those studies described for the spread of obesity [5] and infectious diseases [6].

This paper focuses on the following question: Are organs being kept locally whenever possible? For this we use community analysis on the GSNs we generate from the UNOS database as a way to reveal structure (or lack thereof) in the system.

2 Related Work

In 1954, the Organ Procurement and Transplantation Network (OPTN) was established to maintain a national registry for organ matching and to develop allocation policies. Subsequently, UNOS was formed to develop and operate a system to allocate organs to potential recipients all over United States. Organ procurement organizations (OPOs) are responsible for obtaining and allocating organs for transplantation. UNOS has divided the USA into 11 geographical regions to aid in the handling of organs. These regions differ from each other with respect to size and population. Currently there are 69 OPOs [10] in the 11 national regions. Figure 1 depicts the 11 regions.

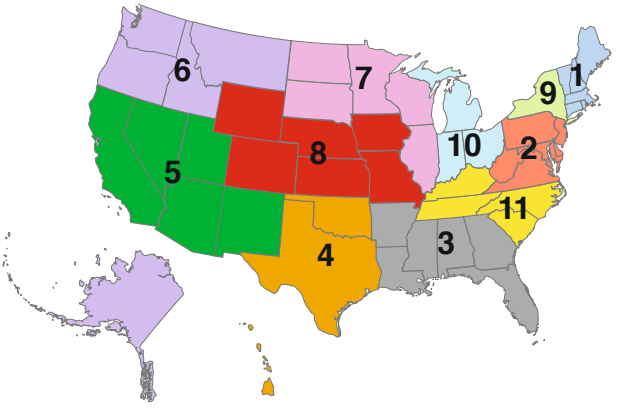


Fig. 1. UNOS divides the United States into 11 geographic regions

The major criteria used for allocating organs to patients are the severity levels and geographical location. Organs are allocated to local OPOs based on the status levels, and only when the severity levels within the local OPOs are exhausted the organs are allocated to the next regional OPOs and lastly to the OPOs at the national level. Patients will be assigned to three different categories based on the predicted mortality. The organs are allocated based on these status levels and are first allocated within the local OPO and only after the local OPOs are exhausted, the organ is allocated to adjacent OPOs. This locality in allocation has a good reason to exist: most organs cannot wait long periods to be transplanted because they degrade in quality—organs such as kidney can only withstand up to 36-48 hours, beyond which the organ can no longer be used for transplantation.

An important issue regarding the fairness of the allocation is the disparities in the amount of the time patients wait for organs. Evidence suggests that minorities and the poor may have limited access to organs [7]. African Americans have to wait twice as long as Caucasians to receive a kidney transplant [1], which might be attributed to a shortage of African American donors. Also,

due to the high cost of health insurance, for many minority patients Medicaid is the only available health insurance. Medicaid policy varies from state to state and does not cover the cost of kidney transplantation completely [8]. All these factors clearly indicate the existence of inequalities in organ allocation system that need to be considered while analyzing the network. The reality is that the organ transplant system represents a collection of relationships between diseases, comorbidities, donors, recipients, healthcare facilities and cultural and economical backgrounds.

In the past few years, our understanding of networks has undergone a revolution because of the emergence of a new array of theoretical tools and techniques for mapping out real networks. The growing interest in interconnectedness showed that networks can be identified for all aspects of human health [2].

3 Geographical Social Network Structure

The essence of network science is to define concepts and measures to characterize the topology of real world networks. Although these concepts have been used in the past for social network analysis, they have recently been used in other areas of science [9].

We introduce the concept of a *(GSN)* of organ transplants. This structure assumes the states are nodes which are linked when a transplant occurs between the states. Note that this still captures the social aspect of organ donation since the transplantation used for the relationship between two geographical locations is a social relation between two individuals.

In this paper we use community detection to find tightly connected group of states. Many algorithms have been proposed to identify communities, but here we use the fast algorithm proposed by Blondel et al. [4].

3.1 Building the GSNs

When building a network, it is very important to carefully choose the objects that the nodes will represent, and the relationships that the edges will capture. The relationship (edge) seems very clear: in the process of organ donation we have a donor and a recipient, so the relationship is this link between the donor and the recipient. However, the use of people as nodes yields a structure with no interesting features. Therefore, we decided to, instead of linking people, link in a geographical network the location of their residence.

In order to build the GSN, we use the dataset provided by UNOS which contains information on all transplants performed in the United States that were reported to OPTN since October 1, 1987. It includes both deceased and living-donor transplants for heart, intestine, kidney, liver, lung, and pancreas. When we assign states as nodes we are likely to find that two nodes can be connected more than once. In network terms, the GSNs are weighted networks that is, if

the number of transplantations between California and Arizona is 20, then the weight of the edge connecting the nodes representing these states is 20.

The transplantation is a directed relationship: the organ goes *from* someone *to* someone else. In this paper we mostly disregard direction and use an undirected version of the network. In most cases, we are trying to find nodes involved in the process independent of the directionality. In Section 4 we discuss that, as future work, we intend to use the directionality to understand how the network changes when considering other dimensions in the dataset such as ethnic groups, education level, and others.

3.2 Analyzing the GSNs

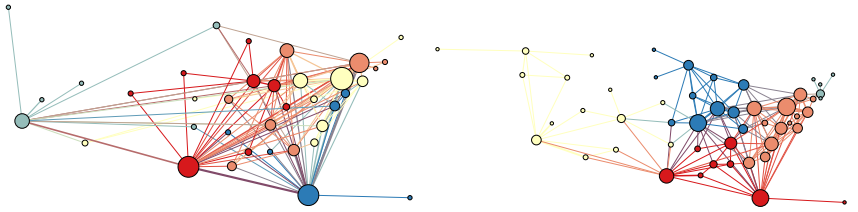
Our first analysis relates to the community formation in the state-level GSN. We have generated a GSN at the state level for each of the six organs available in the dataset and looked at their community divisions. Recall that since the relationships should prioritize smaller distances (according to current organ donations policies), we should be able to see the communities correlating quite well to regions of the United States.

In network sciences it is common to show topological characteristics of the network. However for the state-level study, the network in itself does not display interesting characteristics given its density. Given that states are connected to one another when a transplant occurs between two people from those states, it is likely that the GSNs are almost fully connected. Table 1 shows some information about the full GSN networks.

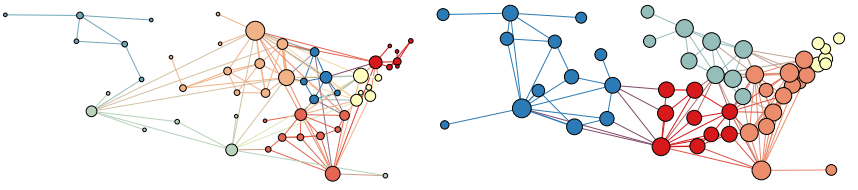
Table 1. All networks below have 56 nodes which include the 50 states in the Unites states plus its territories. Graph density is a measure of the number of edges in the network in relation to the max number of edges it could have (in this case 1,540).

GSN	Edges	Graph Density	Communities
Intestine	519	0.337	11
Lung	882	0.573	4
Pancreas	940	0.610	7
Heart	980	0.636	3
Liver	1,169	0.759	4
Kidney	1,224	0.866	4

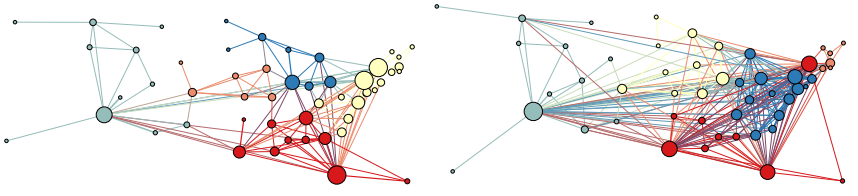
Table 1 shows that the GSNs are very dense. The density value represent the fraction of edges that exist in the graph in relation to the maximum number of edges that could exist. Note that the networks for organs with more transplants are the most dense ones (heart, liver and kidney). Given that they approximate fully connected networks, they do not present interesting features.



(a) GSN for Intestine with $w^+ = 4$ (b) GSN for Lung with $w^+ = 20$



(c) GSN for Pancreas with $w^+ = 20$. (d) GSN for Heart with $w^+ = 40$



(e) GSN for Liver with $w^+ = 50$ (f) GSN for Kidney with $w^+ = 80$

Fig. 2. Community formation of state-level GSNs represented by color. Note the differences between some of the organs. In these networks, the 48 contiguous states are placed relative to their true geographical locations, while Hawaii, Alaska and the territories are not

It is common in the literature to apply an edge threshold to a network in an attempt to reduce the number of edges, making it easier to review network features. Figure 2 depicts the GSNs that have been filtered to remove edges below certain thresholds. The simple explanation for the use of thresholds is that in many instances relationships should be considered as important if they are strong. In other words, if every node is linked to every other node via an edge with a weight at least w, then w does not express a significant relationship and we should therefore only consider edges with a weight greater than w. We can choose any number greater than w as our filtering threshold, which we call w^+. In Figure 2 we use different values of w^+ depending on the structure of the initial, unfiltered network. We find each networks w^+ by trying different values until the community structure becomes clear. The edge-weight threshold is applied empirically for each of the network so that the formation of communities reveals meaningful information.

What we can observe in Figure 2 are significant differences between the geographical organization of the communities. Take for instance Figure 2(d)

which shows five communities for heart transplantation. Although the specific number of communities is not relevant in this study, what we can see is that the communities have very well defined borders that correlate to areas of the United States and more specifically the UNOS regions. We already expected to have less regions in the community because some UNOS regions are very small and near each other, like regions 9 and 10 in Figure 1.

Now contrast Figure 2(a) with Figure 2(d). Without much effort one can clearly see that the geographical divisions between communities is better for the heart GSN than for the intestine GSN. This is an interesting finding because the community analysis of the network reveals that the "distance" aspect of organ policies are respected better for in heart transplantation than intestine transplantation. In Figure 2 we see the GSNs for heart, pancreas, lung, and liver with the best organized communities, while intestine and kidney GSNs demonstrate need for improvement. The GSN for kidney is actually okay except for a node in the northeast of the United States (representing the state of New York) which is part of the southeast community (including Florida, Texas, Puerto Rico, and others). In the GSN for pancreas transplants we see Puerto Rico linked to a community in the southwest of the country rather than staying with the more natural community of states in the southeast.

4 Conclusion and Future Work

In this paper we have shown that the use of network sciences can significantly help us understand and identify problems in the process of organ allocation. Our approach in this paper was to concentrate on what we call a geographical social network (GSN) because the distance traveled by organs is crucial to the health of the organ—they should travel small distances whenever possible. Using community analysis on the GSNs we have demonstrated that the organ allocation process differs amongst the organs and can be further optimized for certain organs such as intestine and kidney. We also show that for the other organs considered in this study the community analysis reveal a structure that is very similar to the regional divisions implemented by UNOS, which means the organs are already being kept as local as possible.

A few examples of data that can help us further develop this study include information about ethnical groups, education level, cause of death, and many others attributes that may shed light on factors that influence the process. In this paper we have already started to identify possible issues related to ethnic groups. We believe the way forward in this study is to understand the differences of the characteristics found here (communities) for different groups divided by ethnicity, education level, income level, religious beliefs, etc. This paper and our future work can help us understand the process of organ allocation in a more detailed way and lead to changes in the system that could benefit thousands of people every year.

Acknowledgements. This work was supported in part by the Health Resources and Services Administration contract 234-2005-370011C. The content is the responsibility of the authors alone and does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

References

1. Alexander, G.C., Sehgal, A.R.: Barriers to cadaveric renal transplantation among blacks, women and the poor. American Medical Association (1998)
2. Barabási, A.-L.: Network theory–the emergence of the creative enterprise. Science 308(5722), 639–641 (2005)
3. Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. Science 286(5439), 509 (1999)
4. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment (10), P10008 (2008)
5. Christakis, N.A., Fowler, J.H.: The spread of obesity in a large social network over 32 years. New England Journal of Medicine 357(4), 370–379 (2007)
6. Gardy, J.L., Johnston, J.C., Sui, S.J.H., Cook, V.J., Shah, L., Brodkin, E., Rempel, S., Moore, R., Zhao, Y., Holt, R., Varhol, R., Birol, I., Lem, M., Sharma, M.K., Elwood, K., Jones, S.J., Brinkman, F.S., Brunham, R.C., Tang, P.: Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. New England Journal of Medicine 364(8), 730–739 (2011)
7. Gibbons, R.D., Duan, N., Meltzer, D., Pope, A., Penhoet, E.D., Dubler, N.N., Francis, C., Gill, B., Guinan, E., Henderson, M., Ildstad, S.T., King, P.A., Martinez-Maldonado, M., Mclain, G.E., Murray, J., Nelkin, D., Spellman, M.W., Pitluck, S.: Waiting for organ transplantation: results of an analysis by an institute of medicine committee. Biostatistics 4(2), 207–222 (2003)
8. Kasiske, B.L., Neylan, J.F., Riggio, R.R., Danovitch, G.M., Kahana, L., Alexander, S.R., White, M.G.: The effect of race on access and outcome in transplantation. New England Journal of Medicine 324(5), 302–307 (1991)
9. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45(2), 167–256 (2003)
10. Pritsker, A., Martin, D., Reust, J., Wagner, M., Wilson, J., Kuhl, M., Roberts, J., Daily, O., Harper, A., Edwards, E., Bennett, L., Burdick, J., Allen, M.: Organ transplantation policy evaluation. IEEE Computer Society Washington (1995)
11. U.S. Department of Health and Human Services. 2009 Annual report of the u.s. organ procurement and transplantation network and the scientific registry of transplant recipients: Transplant data 1999-2008. Technical report, Health Resources and Services Administration, Healthcare Systems Bureau, Division of Transplantation (2009)

A Novel Framework for Complex Networks and Chronic Diseases*

Philippe J. Giabbanelli

Abstract. Complex networks have provided a wealth of information regarding infectious diseases, for example by understanding how the network structure impacts the basic reproduction number or immunization strategies. However, researchers have struggled to translate this knowledge to chronic diseases, where social networks are at play but broad societal factors also have an important role. This translation is becoming urgent given the increasing prevalence, and the escalating healthcare costs, of conditions such as obesity. In this paper, we provide a mathematical framework that enables researchers to represent both the network and societal aspects of chronic disease, thereby facilitating this translation effort. Our framework uses Complex Networks to represent the population, where influences between neighboring nodes are modelled through Fuzzy Cognitive Maps that account for societal effects. Applying our framework to real-world cases, possibly through processes such as Group Model Building, may facilitate the better direction of policy towards the management of chronic diseases.

1 Introduction

Research in complex networks has resulted in tremendous advances to our understanding of *infectious* diseases. Selected examples include understanding how the network's structure impacts on the basic reproduction number [1, 2] (*i.e.*, will the disease die out or become an epidemic in the presence of given properties?),

Philippe J. Giabbanelli
MoCSSy Program, Interdisciplinary Research in the Mathematical and Computational Sciences (IRMACS) Centre; and the Dept. of Biomedical Physiology, Simon Fraser University, Canada
e-mail: giabba@sfu.ca

* Research funded by the Canadian Institutes of Health Research (MT-10574). We are indebted to D.T. Finegood, D. Grace and V.K. Mago for having shared ideas that were to later shape this work. We thank R. Dorrell for his thorough feedback, and D. Grace for his suggestions.

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 207–215.
springerlink.com

mitigates the influence of the starting point [3], or can be capitalized upon to design better immunization strategies [4]. However, a growing number of countries are now facing the burden of *chronic* diseases. For example, the prevalence of obesity in the United States has steadily increased over the past decades, to reach an alarming two third of adults being overweight or obese [5]. This condition is detrimental to individuals' health [6, 7] and also to the country's economy since costs have been estimated from 860 to 960 billion US dollars by 2030 given the current trends [8]. While peers contribute to one's obesity status [9], similarly to infectious diseases, there is also a strong contribution of factors such as social determinants. Thus, complex networks can be useful to represent the contributions of peers, but cannot be directly applied as the disease is not purely infectious.

In this paper, we introduce a novel framework that capitalizes on complex networks to represent peers' effects in chronic diseases, while encompassing broader, societal determinants using Fuzzy Cognitive Maps. This framework has several benefits. Firstly, practitioners may have had a limited training regarding chronic diseases during medical schools [10], and their models often adopt a clinical perspective centered on individuals' physiology. Newer models have either addressed the population structure [11] (*e.g.*, Figures 1 and 2 for obesity), or taken a sociological lens to seek out the social 'root causes', but could not provide practitioners with a comprehensive picture that would complement their clinical knowledge. Our framework makes the creation of such a comprehensive view possible by integrating both network aspects and sociological causes. Furthermore, it is able to carry out predictions, whereas numerous sociological models are solely conceptual. Secondly, a working model constructed from our framework can be populated with information derived from experts' knowledge, via processes such as Group Model Building. Indeed, objective and/or quantitative data may not be readily available, and our framework can capitalize on qualitative, subjective assessments that are more commonly found in sociological approaches. Finally, there is a need to translate the complex networks knowledge on epidemics into chronic diseases, where similar theoretical issues abound but have yet to be addressed (*e.g.*, can public policies leverage social networks to mitigate the obesity epidemic? when training a community, which voices would have the stronger impact based on social ties?).

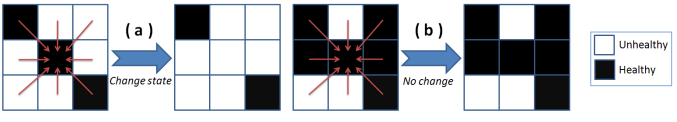


Fig. 1. Rush and colleagues provided an early model of obesity centered on peers [12]. They used a cellular automaton, where each person is a cell whose colour indicates whether they are healthy (black) or unhealthy (white). A person interacts with the 8 neighbours surrounding him (red arrows), whose states are aggregated and compared to a given threshold. If the result is greater than the threshold then the person is assigned that dominant state. For a threshold of 5, in (a) the central healthy person will become unhealthy due to having 6 unhealthy friends, and in (b) the person has too few unhealthy friends to change state. *Color online*.

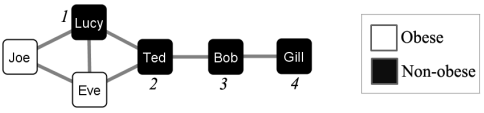


Fig. 2. Similarly to Rush *et al.* [12], Bahr and colleagues considered that an individual's state changes based on peers [13]. Taking a network perspective, they used more realistic population structures than the grid defined by a cellular automaton. However, their rules were entirely based on the idea of obesity as 'infectious': in this example, if a node is assigned the state found in at least the majority of neighbors, then the whole population would become obese at time steps labeled next to the nodes. For example, Lucy has 2 obese friends and 1 non-obese so she will be obese; the same will then apply for Ted.

In Section 2, we informally introduce how our framework *couples complex networks* with *Fuzzy Cognitive Maps* (FCMs). The underlying mathematics are then formally described in Section 3. Finally, we discuss the practical applications of this model, in terms of future research as well as strategies for action.

2 Fuzzy Cognitive Maps and Complex Networks

2.1 Fuzzy Cognitive Maps

An individual can be at increased risk of chronic diseases based on social determinants such as income, education, or gender. Data regarding the relationships between such factors and a disease outcome is often uncertain and/or vague. For example, reports can be conflicting even for a similar population make-up, and experts may disagree with one judging a relationship to be 'medium' whereas another sees it as 'high'. However, decisions such as public policies still have to be made based on this knowledge. Fuzzy Set Theory is precisely designed for this situation [14]:

> Fuzzy set theory resembles human reasoning under approximate information and inaccurate data to generate decisions under uncertain environments. It is designed to mathematically represent uncertainty and vagueness, and to provide formalized tools for dealing with imprecision in real-world problems.

Consider a group of six experts asked to estimate the impact of 'education' on a disease outcome. Following a classical process [15], each would choose linguistic terms such as 'very low', 'low', 'medium', 'high', and 'very high'. Perceptions of what constitutes, for example, a 'medium' relationship differs amongst individuals. What some individuals may call 'medium' would be the same as what others would designate as 'high'. Fuzzy set theory accounts for this issue via membership functions: a term such as 'medium' is not associated to one specific value but rather to a range, and this range can overlap with those used by other terms. If four experts

declare the relationship to be 'very high' while the two others chose it to be 'high', then their opinions can be summarized in the form of IF-THEN rules:

R_1: IF (*Education* is ON) THEN *Disease outcome* is HIGH (positively) (2/6)
R_2: IF (*Education* is ON) THEN *Disease outcome* is VERY HIGH (positively) (4/6)

These rules summarize that 2 out of 6 opinions predicted a high causality, and 4 out of 6 opinions opted for a very high causality. Using these rules and membership functions, one can derive a specific value (*i.e.*, by using an aggregation method, a defuzzification method, and an inference mechanism)[1] [16].

This process can be repeated to estimate the impact of different relationships based on opinions expressed in Group Model Building or in the literature. These relationships can then be articulated to form a network called a Fuzzy Cognitive Map (FCM), introduced by Kosko in 1986 [17]. An FCM is composed of *nodes* representing domain concepts (*e.g.*, 'obesity', 'education', 'income') linked by *edges* representing causal relationships. Concepts take value in the interval [0, 1] where 0 represents that the concept does not hold for a given individual (*i.e.*, it is *false*) and 1 that it certainly holds (*i.e.*, it is true). Edges are labeled as positive or negative to indicate that the target concept respectively increases or decreases with the source concept. Edges take weights in the interval [0, 1], and their weights are obtained by the aforementioned process. Most importantly, the FCM changes over time: it updates the value of concepts, and can thus make predictions (see next Section for the formal process). FCMs have been used extensively [18], including in settings where errors in predictions would have disastrous consequences, such as calculating the dose for radiotherapy treatment [19]. Thus, they offer a robust approach to integrate uncertain or inaccurate knowledge, as is often found when examining how factors contribute to a chronic disease.

2.2 Complex Networks

Social networks are a significant factor in health [21]. This was popularized [22, 23] due to an article from Christakis and Fowler [24]. While the article's methodology was recently criticized [25], its key conclusion about the importance of peers influences is well supported [9]. This was illustrated by the model from Rush *et al.* [12] (Figure 1), later improved by Bahr and colleagues [13] who, instead of constraining interactions to take place with nearest neighbors in a grid, used networks (Figure 2) and various rules for changing state. While thinking of obesity as directly 'contagious' was possibly useful in the early development of models, a heavy reliance on that metaphor ignores entirely social determinants and can lead to making inadequate public health recommendations. This effort was thus pushed forward through a model in which individuals do not influence each other directly [26], but instead influence shared social activities such as (the level of) physical activity or diet patterns (*i.e.*, energy intake). Whether an individual changes behaviour then depends on the

[1] Commonly used tools include the Fuzzy Toolbox from Matlab.

combination of peers' and environmental influences. Changes in behaviour further participate to changes in weight through an approximation of body metabolism. In this model, social determinants start to appear under the umbrella of 'environmental factors', but much is still required to elucidate which determinants matter and how they interact. Therefore, we propose a theoretical framework that couples complex networks with FCMs in order to express both peers' and societal influences, while capitalizing on the strength of techniques valued for these aspects separately.

2.3 Coupling Complex Networks and Fuzzy Cognitive Maps

Section 2.1 showed how Fuzzy Cognitive maps can be valuable to represent social determinants, and Section 2.2 highlighted the role of complex networks in health. In this Section, we explain how our framework couples both techniques.

Initialization. Given a chronic disease such as obesity, assume that an FCM has been constructed and a social network has been created[2]. For each node of the social network, we will create one instance of the FCM (Figure 3). For each of these instances, the values of concepts will be drawn from specific probability distributions. For example, in Figure 3 we need to provide an initial value for 'Exercise'. In the United States, a suitable probability distribution would be a normal distribution with a mean of 1.53 (level for a sedentary individual [31]), a standard deviation of 0.1 (since most individuals are sedentary [32]), and a range of 1.4 to 4.7 (based on data from the Food and Agriculture Organization [31].

Matching. In the FCM, we identify the concepts that are *influenced* by peers, and those (not necessarily distinct) that are *influencing* peers. Then, we establish a matching (*i.e.*, a weighted bipartite graph) linking influenced concepts to influencing concepts and indicating the strength. How exactly one concept influences another is problem-specific, and is a classical issue studied as 'culture diffusion' in anthropology. For example, in Axelrod's approach, the strength of the interaction depends on the cultural similarity between two individuals, and the outcome is to exchange the value on a concept where the individuals differ [33]. A smoother approach, similar to [26], would be to consider that *if* the difference between one and peers' is significant enough then an influence can be conveyed, and will modify one's concepts by a given percentage, similarly to the approach in [20].

Simulation. The overall process, formally specified in Section 3, consists of (i) applying the influence of peers on each individual's FCM (as discussed above), and (ii) applying the inference engine of the FCM for each individual. The first task

[2] If no real-world social structure can be used, we would recommend to generate a small-world network since the last decade showed this property to hold often in social networks [27]. The experiments can also be carried on several synthetic populations, in which case one could account for the presence of other population-wide properties such as the scale-free distribution of degrees. Generating such networks can be achieved using an array of models from statistical mechanics or graph theory (*e.g.*, [28, 29, 30]).

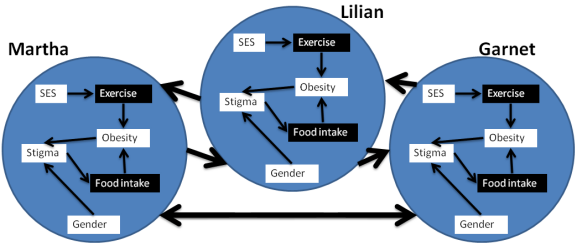


Fig. 3. The initialization provides each individual with an instance of the FCM. Some concepts of the FCM can be influenced by peers (in black) while others (in white) cannot. This example shows that one's exercise and food intake can be influenced by peers, whereas the socio-economic status (SES) is not perceived as being influenced by peers. The simulation will evolve the influenced factors based on peers' FCMs, and then apply the inference engine to evolve each FCM.

must be carried on in parallel in order for all individuals to change at the same time (similarly to updating the values of a cellular automaton). In the second task, the FCMs are independent and the operations can be carried on in any order.

3 Formal Framework

The formal definitions follow the order introduced in Section 2. Definition 1 formalizes an FCM (Section 2.1). Definition 2 formalizes the social network (Section 2.2). The coupling (Section 2.3) firstly requires us to define influenced (in black in Figure 3) and influencing factors, in Definition 3. Then, the matching between these factors is provided by Definition 4 and the influences carried on by the matching are formalized in Definition 5. Finally, Algorithm 1 states how the population evolves at a given time step, and its key lines are discussed.

Definition 1. *A Fuzzy Cognitive Map (FCM) is formalized as a 4-tuple* (V,E,W,M) *where the vertices* V *are the concepts, the directed edges* E *are the relationships between concepts, the matrix of weights* $W = (w_{i,j})_{,1\leq i\leq |V|,1\leq j\leq |V|} \in \mathbb{R}$ *represents the strengths of these relationships, and the vector* $M = (m_i)_{,1\leq i\leq |V|} \in [0,1]$ *associates a value to each vertex.*

Definition 2. *The population is formalized as a directed unweighted graph* $G = (\mathbb{V},\mathbb{E})$ *where the vertices* $\mathbb{V}$ *are the individuals and the edges* $\mathbb{E}$ *are their social ties. Each vertex* $v \in \mathbb{V}$ *contains a Fuzzy Cognitive Map, denoted by* v_{FCM}.

Definition 3. *The set of concepts of an FCM that are influenced by peers is denoted by* $\alpha(V) \subseteq V$, *while concepts influencing another FCM are denoted by* $\beta(V) \subseteq V$.

Definition 4. *A matching function* $F : (a \in \alpha(V)) \mapsto \{b \in \beta(V)\}$ *determines, for each influenced concept, the set of concepts that influence it*[3].

Definition 5. *The function* $\gamma : (\mathbb{R}, \mathbb{R}, \mathbb{R}) \mapsto \mathbb{R}$ *takes an influenced concept, an influencing concept, the weight of the relationship between the two, and determines the change that should be applied on the influenced concept.*

Algorithm 1. Evolves the population

Require: The population has been initialized, the FCMs have starting values

1: *//Applies the social influences in parallel*
2: **for** $i \in \mathbb{V}$ **do**
3: **for** $j \in \mathbb{V} | (j,i) \in \mathbb{E}$ **do**
4: *//for each neighbor j influencing a person i*
5: **for** $a \in \alpha(i_{FCM})$ **do**
6: *//for each concept influenced by peers*
7: **for** $b \in F(a)$ **do**
8: *//for each influencing concept*
9: $M(a) \leftarrow M(a) + \gamma(M(a), M(b), W((a,b)))$ *//updates the value of concept a*
10: *//Evolves each individual's FCM until it stabilizes*
11: **for** $i \in \mathbb{V}$ **do**
12: **while** i_{FCM} does not stabilize **do**
13: **for** $g \in V(i_{FCM})$ **do**
14: $M_g(t+1) = f(M_g(t) + \sum_{h \in V(g_{FCM}), h \neq g} M_h(t) \times W_{h,g})$ *//updates each concept*

Social influences are all applied at the same time so the operation must be done in parallel (Section 2.3). Therefore, the new value of $M(a)$ (line 9) can be buffered, and all new values updated after the main loop (line 10). For $|\mathbb{V}|$ individual FCMs with $|\alpha(V)|$ influenced concepts each, this takes a space of $|\mathbb{V}| \times |\alpha(V)|$. Less space can be consumed if an individual is updated as soon as all his neighbors have been processed. The FCM's stabilization condition (line 12) commonly consists of updating all concepts until few *target* concepts (*i.e.*, those giving the predictions) change by less than a specified amount. Line 14 is the standard update of an FCM [34], where f is a threshold function that bounds the new concept's value to be in the interval [0, 1]. This function can be, for instance, a sigmoid function such as the hyperbolic function $f(x) = \tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. Some concepts of an FCM are constant (*e.g.*, they have no incoming edges) and they should not be updated. This can be solved either by restoring these concepts to their previous value following line 14, or selecting only those concepts that satisfy problem-specific requirements (line 13).

[3] In Section 2.3, we assumed that influencing and influenced concepts are the same in all FCMs. That is, $\forall a, b \in \mathbb{V}, \alpha(a_{FCM}) = \alpha(b_{FCM}), \beta(a_{FCM}) = \beta(b_{FCM})$. This simplifies the definition of F by assuming that influencing factors do not depend on which peer is selected. However, Algorithm 1 does not depend on this assumption. Therefore, if one needs a model in which the factors influencing individuals vary among peers, the matching function F should be generalized to also depend on the peer; that is, $F : (a \in \alpha(V_{i \in \mathbb{V}}), j \in \mathbb{V}) \mapsto \{b \in \beta(V_j)\}$.

4 Conclusion

Complex networks have provided a wealth of information regarding infectious diseases. However, the community has struggled to translate this knowledge to chronic diseases, where one is influenced by peers but also strongly by social determinants. This translation is becoming urgent given the increasing burden of chronic diseases such as obesity. In this paper, we have developed a new mathematical framework based on complex networks and Fuzzy Cognitive Maps.

Our framework facilitates the translation of the theory of complex networks and epidemics into chronic diseases, and it also benefits settings such as group model building. Indeed, groups have gathered to better understand chronic diseases by creating maps[4]. This demonstrates that groups can identify key factors and their relationships, but the end product often remains at a conceptual stage and does not have predictive power. Our framework enables these groups to push forward their process into mathematical models, while only requiring (i) a synthesis of participants' knowledge on the strength of relationship, (ii) the identification of influenced and influencing factors, (iii) a synthesis of the relationship between influencing and influenced factors. Our use of fuzzy logic simplifies the process by enabling participants to describe strengths via linguistic terms such as 'Very high' or 'medium'.

Our framework comes with limitations that should be the object of future theoretical research. Each node of our networks is influenced by a set of factors, whose concepts can change over time while their relationships are considered to be constant. However, there is evidence that these relationships change, which could prompt for dynamic networks depending on the time scale under consideration. The specification of the dynamicity could be informed by a life course perspective, since it shows how the factors influencing a node change as its role evolves (*e.g.*, [35]). Finally, setting up a virtual population requires providing initial values for the factors influencing each individual (Section 2.3-Initialization). For a given scenario, data may be available for *each* factor, but factors are rarely independent from each other. For example, assuming statistical independence between being black, a woman, and living with obesity, may be very inaccurate when the individual's experience is precisely shape by being a black obese woman altogether. This issue is not specific to our framework but abounds in the way we often conceptualize populations [36], and requires advances in data collection tools [37].

References

1. Pastor-Satorras, R., Vespignani, A.: Phys. Rev. E 65(3), 036104 (2002)
2. Piccardi, C., Casagrandi, R.: Phys. Rev. E 77(2), 026113 (2008)
3. Crepey, P., Alvarez, F.P., Barthelemy, M.: Phys. Rev. E 73(4), 046131 (2006)
4. Boily, M., Asghar, Z., Garske, T., Ghani, A., Poulin, R.: Math. Popul. Stud. 14, 237 (2007)

[4] For the latest reports, see (online) *Group Model Building* from Peter S. Hovmand and Bishop George White, Institute for Systems Science and Health, Pittsburgh May 22-27, 2011.

5. Flegal, K., Carroll, M., Ogden, C., Curtin, L.: JAMA 303, 235 (2010)
6. Sturmer, T., Gunther, K.P., Brenner, H.: Journal of Clinical Epidemiology 53, 307 (2000)
7. Sturm, R.: Health Affairs 21(2), 245 (2002)
8. Wang, Y., Beydoun, M.A., Liang, L., Caballero, B., Kumanyika, S.K.: Obesity 16, 2323 (2008)
9. Hammond, R.A.: Current Opinion in Endocrinology, Diabetes and Obesity 17, 467 (2010)
10. Block, J.P., DeSalvo, K.B., Fisher, W.P.: Preventive Medicine 36, 669 (2003)
11. Bedoya, D., Matteson, C., Finegood, D.: Obesity Reviews 11(suppl. 1), 471 (2010)
12. Rush, W., Biltz, G., et al.: Tech. rep. (New England Complex Systems Institute) (2003)
13. Bahr, D., Browning, R., Wyatt, H., Hill, J.: Obesity 17(4), 723 (2009)
14. Li, Z.: Introduction. In: Fuzzy Chaotic Systems, pp. 1–11. Springer (2006)
15. Zadeh, L.A.: Information and Control 8, 338 (1965)
16. Stylios, C.D., Groumpos, P.P.: Journal of Intelligent and Fuzzy Systems 8(2), 83 (2000)
17. Kosko, B.: International Journal of Man-Machine Studies 24, 65 (1986)
18. Stylios, C.D., Georgopoulos, V.C., et al.: Applied Soft Computing 8, 1243 (2008)
19. Papageorgiou, E., Spyridonos, P., et al.: Applied Soft Computing 8(1), 820 (2008)
20. Giabbanelli, P.J., Tornsney-Weir, T., Mago, V.K.: Applied Soft Systems (2012), under revisions
21. Smith, K., Christakis, N.: Annual Review of Sociology 34, 405 (2008)
22. Kolata, G.: The New York Times (2007)
23. Thompson, C.: The New York Times (2009)
24. Christakis, N.A., Fowler, J.H.: New England Journal of Medicine 357(4), 370 (2007)
25. Lyons, R.: Statistics, Politics, and Policy 2(1) (2011)
26. Giabbanelli, P.J., Alimadad, A., et al.: Obesity Reviews 11(suppl. 1), 65 (2010)
27. Schnettler, S.: Social networks 21, 165 (2009)
28. Comellas, F., Ozon, J., Peters, J.: Information Processing Letters 76(1-2), 83 (2000)
29. Giabbanelli, P.: Advances in Complex Systems 14(6), 853 (2011)
30. Watts, D., Strogatz, S.: Nature 393, 440 (1998)
31. Food and Agriculture Organization, Report, joint FAO/WHO/UNU expert consultation (2004)
32. Booth, F., Chakravarthy, M.: Presidents Council on Physical Fitness and Sports Research Digest 3(16) (2002)
33. Axelrod, R.: Journal of Conflict Resolution 41(2), 203 (1997)
34. Stach, W., Kurgan, L., Pedrycz, W., Reformat, M.: Fuzzy Sets and Systems 153(3), 371 (2005)
35. Blane, D.: The Life Course, the Social Gradient, and Health. In: Social Determinants of Health. Oxford University Press (2006)
36. Wheaton, W., Cajka, J., Chasteen, B., et al.: RTI Press Publication MR-0010-0905 (2009)
37. Bowleg, L.: Sex Roles 59, 312 (2008)

Centrality and Network Analysis in a Natural Perturbed Ecosystem

Gilberto C. Pereira, Fatima F. Santos, and Nelson F.F. Ebecken

Abstract. The aim of this work is to gain knowledge on the interactions between the chlorophyll-a and nine meroplankton larvae of epibenthonic fauna. The studied case is the Arraial do Cabo upwelling system, Southeastern of Brazil, which provides different environmental conditions. To assess this information a network approach based in probability estimative was used. Comparisons among the generated graphs are made in the light of different water masses, application of Shannon biodiversity index, and the closeness and betweenness centralities measurements. Our results show the main pattern among different water masses and how the core organisms belonging to the network skeleton are correlated to the main environmental variable. We conclude that the approach of complex networks is a promising tool for environmental diagnostic.

Keywords: Coastal upwelling, Ecological networks, Plankton interactions, Environmental analysis.

1 Introduction

The Brazilian coast presents a large variety of ecosystems but little is known about its biodiversity, degree of connectivity and behavioral patterns. Nowadays, these systems are subjected to a large number of anthropogenic pressures without either knowing the load processing ability of the biological networks and its structural stability. These issues make any initiative in coastal management hard and complex for decision making. One of the major problems is the habitat change, destruction or loss (Halpern et al., 2007; 2008). Despite their adaptive character (Levin and Lubchenco 2008) and often redundant linkages, marine ecosystems are vulnerable to rapid changes in diversity and function (Palumbi et al. 2008). The widespread decline of species, habitats, and ecosystem function have led to calls

Gilberto Carvalho Pereira · Fatima F. Santos · Nelson F.F. Ebecken
Federal University of Rio de Janerio-Brazil
e-mail: {gcp,fatima,nelson}@ntt.ufrj.br

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 217–224.
springerlink.com

for ecosystem based management (EBM) as a solution for what ails the oceans (USCOP 2004, SEMIEA, 2004). Recent legislative instruments have been approved worldwide addressing the need to assess the ecological status (Borja et al. 2008). In this way, many initiatives can be found in literature (Simonini et al., 2009; Pereira et al., 2008). According Norberg (2004), environmental factors regulate biodiversity through species sorting processes. Species distributions in communities affect ecosystem processes and environmental factors. These dynamics are determined by the traits of species in the community. The question of how changes in biodiversity will affect the ecosystem functioning, the so-called biodiversity-ecosystem function (BEF) debate, is clearly not easy to answer. However, it has long been recognized that species interact in ecosystems with other species and with abiotic factors in many ways, of which pairwise interactions are only one possibility (Hutchinson, 1959). In this context, Raffaelli (2006) argued that a system approach is necessary to address issues involving changes in biodiversity and function of natural ecosystems. Although pairwise interactions have always had a key role in ecology, a new focus on complex networks has been placed (Dunne et al., 2002). Several studies have shown how the structural characteristics of complex networks are related to their stability and dynamic (Huxel and McCann, 1998, Albert et. al, 2000, Strogatz, 2001, Kolasa, 2005). The application of centralities indices to the network component can identify the keystone species (Jórdan et al., 2006, Libralato, et al., 2006), and the role they play in a network (Gonzalez et, al, 2009). So, the aim of this paper is (i) to use a network approach to investigate differences between different water masses based on graphs generated from chlorophyll-a, merozooplankton larvae of epibenthic fauna and some environmental parameters; (ii) establish differences in biodiversity; (iii) apply the closeness and betweenness centralities measures in order to determine the positional importance of each specie or node; (iv) identify which set of n node belongs to the core skeleton of the network, (v) examine how the core organisms are correlated to the main environmental variable.

2 Material and Methods

2.1 Studied Area

The studied plankton community is found in a small (45 km2), shallow (10m depth), wind-driven and upwelling-influenced Anjos Bay, which is formed by Cabo Frio Island (23 S, 42 W) in the state of Rio de Janeiro, southeastern Brazil.

Dominant E¬NE winds are influenced by tropical maritime anticyclones due to the Coriolis Effect and Ekman transport, which shunt nutrient-depleted surface water (Brazil Current) offshore (Castelao and Barth, 2006). This water body is followed by up-flowing, nutrient-rich (12 μMLNO3-N), deeper South Atlantic Central Water (SACW), which comes from around 200–300m depth and reaches the surface sporadically. An inverse pattern can be caused by S-SW winds because cold fronts drive the oligotrophic Brazil Current (<1 μM-LNO3-N) toward the coast. As SACW is heated in the euphotic layer, nitrate declines more rapidly than

phosphate, and the N/P ratio declines (Pereira et al., 2009). These processes generate different habitat conditions that influence at the same time changes in community and trophic structure (Pereira and Ebecken, 2009).

2.2 Available Data

The available data (Table 1), is a matrix of 18 variables and 512 samples concerning to a weekly harvested medium-term time series (10 years) of physical, chemical and biological gradients coming from November of 1994 to December of 2005.

Table 1. Basic Statistics of the available data

Variables	**Min**	**Max**	**Mean**	**Std dev**
Temperature (^oC)	15,88	29,40	22,66	1,84
Salinity (g/L)	32,13	39,78	35,82	0,86
Oxygen (O_2)(mg/L)	2,58	8,79	5,29	0,49
Phosphate (PO_4) (µg/l)	0,00	3,69	0,26	0,21
Nitrite (NO_2) (µg/l)	0,00	0,64	0,08	8,08
Nitrate(No_3) (µg/l)	-0,09	10,19	0,68	0,95
Ammonium (NH_4) (µg/l)	0,07	7,85	1,26	0,88
PH	6,39	10,44	8,13	0,42
Chlorophyll-a (mg/m^3)	0	11,94	0,99	1,17
Cirripedia (Org/m^3)	0	3641	205	355
Mytilidae (Org/m^3)	0	2636	92	170
Decapoda (Org/m^3)	0	437	20	35
Polychaeta (Org/m^3)	0	1683	20	89
Ostreidae (Org/m^3)	0	1132	31	91
Cypris (Org/m^3)	0	5192	21	248
Ascidiacea (Org/m^3)	0	1115	14	66
Isognomon (Org/m^3)	0	2342	31	161
Bryozoa (Org/m^3)	0	101	2	6

The physical and chemical variables demonstrate the hydrologic variability of the environment as a function of interchangeable periods of upwelling and downwelling events. The water mass identification was made through temperature and salinity gradients according Pereira *et al.,* (2008). The biological variables are the chlorophyll-a as estimation of phytoplankton biomass, a single food resource, and 9 merozooplankton taxa representing consumers. Every variable were categorized into five classes: zero which means no occurrence, low, mean, high and extremely high.

2.3 Network Generation

In this paper we present an approach for the discovery of community structure in networks with only a single type of vertex (although they represent biotic and abiotic variables we considered them the same type) and a single type of undirected and unweighted edge, although generalizations to more complicated network types are possible. Our divisive algorithm focus is not on removing the edges between vertex pairs with low similarity, but on finding edges with the highest values of occurrences, i.e. we focus on finding community structure based on the values of the edges and not on the attributes of the vertices, as is more usual

Each variable is a vertex or node in a graph whose edges represent the interaction between them. We were not interested to know only which variables each population interacts with, but to measure the simultaneity intensity of this interaction To quantify these interactions, it was considered the probability of presence of variable Bi given the presence of a variable Bj, and is thus a measure of the statistical association between Bi and Bj. represented by P (Bi|Bj) which measures the strength of the association between Bi and Bj. As P (Bi|Bj) does not take into account statistical confidence, we considered the equation 1, proposed by Stephens et. al. (2009), which also measures the degree of confidence one can have in the statistical association between Bi and Bj relative to the null hypothesis, P(Bi), that the distribution of Bi is independent of Bj and distributed with this probability over the region of interest (in our study, just one geographic position).

$$e\,(Bij \mid Bj) = NBj\,(P(\,Bi \mid Bj) - P(\,Bi))\,/\,(NBj\,(P(\,Bi)\,(1 - P(Bi)))\,½ \qquad (1)$$

Essentially, it is a one-sided binomial test where the null hypothesis is that the distribution of Bi is random over the collected data. The sum of the values of the edges was considered to identify the network structure. Thus, in our proposal, an edge is considered part of a sub-network if it connects a vertex pair in an amount equal to a defined threshold. Naturally, the vertex pair connected by this edge is also part of this sub-network. So, as an example, if a vertex is part of a 10-threshold sub-network, it is connected by an edge with value "10" to at least one other vertex. The approach we take to identify the structure of the network follows roughly these lines. Thus, the general form of our network structure finding algorithm is as follows:

1. Calculate the value of each interaction with equation 1.
2. Calculate the value of the sum of the edge (for each pair of vertices) in the network.
3. Calculate the frequency distribution of the edges over the values (each frequency distribution class is equal 1).
4. Identify the values that correspond to 25%, 50% and 75% of the frequency distribution as a reference for first pre-division of the edges (and respective vertex pair) into sub-networks. Naturally, a vertex is categorized in a frequency distribution class according the highest value of its edges.

5. Categorize each edge (and the vertex pair connected by it) in one of the 4 intervals defined above.
6. Identify the edges with the highest values (above 75% interval) and considered them part of the core network.

For graphical representation, the interactions whose values are higher than the value that corresponds to75% of distribution (empirical threshold) appear reinforced. In fact, this threshold was set based on the average summed to one fold of the standard deviation of all water masses. On this way, the nodes connected by these stronger values of interactions will be considered as belonging to the skeletons of such networks.

2.4 Applied Indices

The Shannon-Wiener index was applied to each network to access differences in biodiversity. It was computed as:

$$H' \quad -\sum_{i=1}^{S}(p_i \ln p_i) - [(S-1)/2N] \tag{2}$$

where S is the total number of species, called species richness; N the total number of individuals. pi is the relative abundance of each species i, calculated as the proportion of individuals of a given species to the total number of individuals in the community: ni/N. In order to establish the positional importance of each node we applied two common measures of centrality: closeness (CC) and betweenness (CB). The former, is based on the total distance between one vertex and all other, such that large distances yield low centrality values. In the network theory, it is defined as the mean geodesic distance between a vertex v and all other vertices reachable from it such as:

$$\frac{\sum_{t:V\setminus v} d_G(v,t)}{n-1} \tag{3}$$

On the other hand, betweenness is a centrality measure of a vertex within a graph so that vertices that occur on shorter paths between others have higher betweenness than those that do not. For a graph G:=(V,E) with n vertices, the betweenness CB(v) for vertex v is computed as follows:

1. For each pair of vertices (s,t), compute all shortest paths between them.
2. For each pair of vertices (s,t), determine the fraction of shortest paths that pass through the vertex in question (here, vertex v).
3. Sum this fraction over all pairs of vertices (s,t). Such that (Shivaram, 2005):

3 Results and Discussion

The use of temperature and salinity data enabled us to identify different water masses such that the Coastal/Tropical mixing type corresponds to 44.80% of the occurrences followed by the Tropical water of Brazil current (25,57%), Coastal water (22.40%), SACW/Coastal (3.17%), SACW/Tropical (1.36%) and SACW with only 0.90%. However, as previously reported by Pereira et. al.(2008), it was found that 1.81% of the examples do not belong to any of these ranges, suggesting another class of water, identified here as "New". Table 2 presents the results of the applied index related to each node and water masses.

Table 2. The Shannon-Wiener, Closeness and Betweenness indices of the studied populations for each water mass

	Coastal Water	Coastal tropical	Tropical	Acas	Acas Coastal	Acas Tropical
Shannon Index	1.94	2.0	1.97	1.47	1.63	1.49
Closeness and Betweeness Centralities (CC-CB)						
Biotic Variables						
Ascidiacea	0.9-0	0.69-0	0.9-0	0-0	0.64-0	0.47-0
Bryozoa	0.9-0	0.75-0.02	1-0.02	0-0	0.64-0	0-0
Cirripedia	1-0.03	0.9-0.04	1-0.02	0.53-0	0.75-0.02	0.7-0.06
ClorofA	1-0.03	1-0.27	1-0.02	0.7-0.25	1-0.28	0.8-0.23
Cypris	0.69-1	0.53-0	0.69-0	0-0	0.56-0	0.43-0
Decapoda	1-0.03	0.82-0.02	1-0.02	0.38-0	0.9-0.19	0.62-0.01
Isognomon	0.82-0	0.69-0	0.9-0	0.38-0	0.56-0	0-0
Mytilidae	1-0.03	0.82-0.03	1-0.02	0.53-0	0.75-0.02	0.56-0
Ostreida	1-003	0.82-0.02	0.9-0	0.53-0	0.75-0.04	0.62-0.01
Polychaeta	0.82-0	0.69-0	0.9-0	0.53-0	0.6-0	0.56-0

The highest biodiversity (2) occurs in the mixing of Coastal/Tropical Water Mass, while the smallest (1,47) was verified in the SACW. The centrality values provide us a good evaluation about the positional importance of these populations or nodes in each of water mass. The graph topology of the Coastal/Tropical Water Mass can be constructed. It will show firstly the occurrence of the lowest (L) values of these variables indicating the oligotrophic condition of this water mass. It is also possible to see that chlorophyll-a is strongly and preferably associated to the ammonium (NH4) followed by phosphate (PO4), nitrate (NO3) and nitrite (NO2) respectively. The PO4 importance to chlorophyll in this system has been previously highlighted in (Pereira et al. 2009). This graph would explicit the occurrence of Cirripedia, Mytilidae and Decapoda as the main consumers. The chlorophyll-a and these three groups of consumers are present in the most of water mass (data not show) indicating they represent the skeleton of the biological network at the

studied site. Differences were detected by the presence of mean values of chlorophylla(M) in SACW and SACW/ Tropical water mass that is the result of upwelling process and the absence of Decapoda in the class "New".

4 Conclusion

The fundamental goals underlying community ecology is to model the distribution of biota, identify their interactions patterns and understand what drives the assemblages in order to perform predictions. The biological monitoring of the marine part of coastal zone is crucial and has become a politically as well a scientifically vital task. The main contribution of this paper is to show how the representation of biological interaction could be constructed through a network approach to discriminate those of greater influence for a specific condition. It was possible to identify the core network of each water mass and their similarities.

Acknowledgments. This paper was supported by CAPES the Brazilian research agency and FAPERJ the Rio de Janeiro State research foundation.

References

1. Albert, R., Jeong, H., Barabási, A.L.: Error and attack tolerance of complex network. Nature 406, 378–382 (2000)
2. Borja, A., Dauer, D.M.: Assessing the environmental quality status in estuarine and coastal systems: comparing methodologies and indices. Ecological Indicators 8(4), 331–337 (2008)
3. Castelao, R.M., Barth, J.A.: Upwelling around Cabo Frio, Brazil: The importance of wind stress curl. Geophys. Res. Lett. 33, 3602 (2005)
4. Dunne, J.A., Williams, R.J., Martinez, N.D.: Network structure and biodiversity loss in food webs: robustness increases with connectance. Ecol. Lett. 5, 558–567 (2002)
5. Gonzalez, A.M.M., Dalsgaard, B., Olesen, J.M.: Centrality measures and the importance of generalist species in pollination network. Ecological Complexity 7, 36–43 (2010)
6. Halpern, B.S., Walbridge, S., Selkoe, K.A., Kappel, C.V., Micheli, F., D'Agrosa, C., Bruno, J.F., Casey, K.S., Ebert, C., Fox, H.E., Fujita, R., Heinemann, D., Lenihan, H.S., Madin, E.M.P., Perry, M.T., Selig, E.R., Spalding, M., Steneck, R., Watson, R.: A global map of human impact on marine ecosystems. Science 319, 948–952 (2008)
7. Halpern, B.S., Selkoe, K.A., Micheli, F., Kappel, K.: Evaluating and ranking the vulnerability of global marine ecosystems to anthropogenic threats. Conservation Biology 21, 1301–1315 (2007)
8. Huxel, G.R., Mc Cann, K.: Food Web Stability: The Influence of Trophic Flows across Habitats. The American Naturalist 152 (3), 460-469 152(3), 460–469 (1998); Jordán, F., Liu, W.-C., Davis, A.J.: Topological keystone species: measures of positional importance in food webs. Oikos 112, 535–546 (2006)
9. Kolasa, J.: Complexity, system integration, and susceptibility to change: biodiversity connection. Ecol. Complex. 2, 431–442 (2005)

10. Libralato, S., Christensen, V., Pauly, D.: A method for identi-fying keystone species in food web models. Ecological Modeling 195, 153–171 (2006)
11. Lubchenco, J.: Resilience, robustness, and marine ecosystem based management. Bioscience 58, 1–11 (2008)
12. Newman, M.E.J.: A measure of betweenness centrality based on random walks. Arxiv preprint cond-mat/0309045, pp. 1–15 (2003)
13. Norberg, J.: Biodiversity and ecosystem functioning: A complex adaptive systems approach. Limnol. Oceanogr. 49(4, part 2), 1269–1277 (2004)
14. Palumbi, S., McLeod, K.L., Grunbaum, D.: Ecosystems in action: Lessons from marine ecology about recovery, resistance, and reversibility. Bioscience 58(1), 33–42 (2008)
15. Pereira, G.C., Coutinho, R., Ebecken, N.F.F.: Data Mining for environmental analysis and diagnostic: a case study of upwelling ecosystem of Arraial do Cabo. Brazilian Journal of Oceanography 56(1), 1–12 (2008)
16. Pereira, G.C., Evsukoff, A., Ebecken, N.F.F.: Fuzzy modelling of chlorophyll production in a Brazilian upwelling system, Ecol. Model. 220, 1506–1512 (2009)
17. Pereira, G.C., Ebecken, N.F.F.: Knowledge discovering for coastal waters classification. Expert Systems with Applications 36, 8604–8609 (2009)
18. Raffaelli, D., Van den Putten, W.H., Person, L., Wardle, D.A., Petchey, O.L., Koricheva, J., Van den Heijden, M.G.A.: Multi-trophic processes and ecosystem function. In: Loreau, M. (ed.) Biodiversity and Ecosystem Functioning (2002)
19. Naeem, S., Inckhausti, P.: [SEMIEA] EU Supporting European Marine Integrated Ecosystem Assessments, p. 256. Oxford University Press, Oxford (2004)
20. Simonini, R., Grandi, V., Massamba-N'Siala, G., Iotti, M., Montanari, G.: International Council for the Exploration of the Sea, Copenhagen, Denmark
21. D: Assessing the ecological status of the North-western Adriatic Sea within the European Water Framework Directive: a comparison of Bentix, AMBI and M AMBI methods. Marine Ecology 30, 241–254 (2009)
22. Shivaram, N.: The Betweenness Centrality of Biological Networks. Thesis of the Faculty of Virginia Polytechnic Institute and State University. p. 74 (2005)
23. Stephens, C.R., Heau, J.G., Gonzalez, C., Ibarra-Cerdenã, C.N., Sanchez-Cordero, V., et al.: Using Biotic Interaction Networks for Prediction in Biodiversity and Emerging Diseases. PLoS ONE 4(5), e5725 (2009), doi:10.1371/journal.pone.0005725
24. Strogatz, S.H.: Exploring complex networks. Nature 420, 268–276 (2001)
25. [USCOP] US Commission on Ocean Policy, An Ocean Blue-print for the 21st Century. USCOP, Washington, DC (2004)

The Explanatory Power of Relations and an Application to an Economic Network

Mauricio Monsalve

Abstract. Understanding the topology of complex networks is a central concern of network science. Within this endeavor, we study the problems of building theories from the non topological attributes of linked vertices and assessing their explanatory power. We design a simple framework for building theories from the attributes of vertices and apply it to explain the topology of the Chilean shareholding network, an economic network which vertices represent firms and edges represent an ownership relation, finding that a relational theory based on financial information explained the topology of the network only in part.

1 Introduction

Understanding the topology of complex networks is a central concern of network science [1]. Some relationships are explained by the topology of a network while others are explained by the nature of the elements in the relation. Let us first consider a topologically explained relationship. Popularity produces social relations yet popularity itself is also produced by them. (See the Albert-Barabasi model [2] for example.) This is a topological effect rather than a relational one: here the network topology sets the likelihood an edge between two vertices is formed. But other relationships are better explained by the nature of the involved actors. Consider physical attraction, where one person likes another person. This relationship is better explained by the physical and behavioral traits of people. Thus, physical attraction is explained at the relational level: the likelihood an edge between two vertices exists depends solely on the non topological attributes of each vertex. However, not all relationships are better explained by either the relational or topological levels. How do we know to which extent a relationship is explained by each?

Mauricio Monsalve
The University of Iowa, Iowa City, Iowa
e-mail: mauricio-monsalve@uiowa.edu

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 225–236.
springerlink.com

Much research has been devoted to the study of the relations, under different scopes and contexts [3]. *Dyadic analysis* has been concerned with studying dyadic (relational) data, and has been closely connected to social network analysis and statistics [4, 5]. *Relational data analysis and mining* has been concerned with describing and searching relations, and has been closely connected to logic and statistics [6, 3]. And bearing the word *link* as in Web hyperlinks, the name *link analysis and mining* has often been used in Web research [3, 7, 8].

Experience has shown that the attributes of vertices should at least partially explain the topology of networks. By using the attributes of vertices to understand relations, researches have often found that similar vertices are often connected together or the opposite. These phenomena have been called *homophily* and *heterophily*, respectively [9, 10]. And when they are about similar or dissimilar degrees, then they are called *assortativity* and *dissortativity*, respectively [11].

In this work, we develop a more general methodology to elaborate theories that explain the existence of edges by taking into consideration the non topological attributes of the involved vertices, discuss how to evaluate the explanatory power of such explanations, and apply the methodology to an economic network: *the Chilean shareholding network*. A shareholding or stock ownership network is an ownership network linked by a partial ownership relation [12, 13, 14]. Shareholding networks have been studied in a number of countries [14, 15, 16, 17], and have acquired special interest because of their role in the recent financial crises [18].

The paper is organized as follows. Section 2 proposes the methodology for building relational theories. Section 3 explains how the Chilean shareholding network was obtained. Section 4 shows the construction of a relational, non topological theory to explain the Chilean shareholding network and its evaluation. Conclusions are presented in section 5.

2 The Explanatory Power of the Relational Level

We have divided vertex attributes in topological and non topological, and we have said that relationships can be explained by them. But if we want to determine to which extent relations are explained by each type of attribute, we must be careful. We must not assume that topological and non topological attributes explain different phenomena. It could be the case that a relationship is well explained by either one, and so the structure of the network. Then, the question is about determining which type of attribute is better and enough to explain the relationship.

We recommend using non topological attributes when it comes to explaining the relationship structure of a network. Observe that topological attributes can be arbitrarily complex. In particular, it is possible to uniquely identify structurally equivalent vertices in a network thorough topological attributes, for example, by making use of many different evaluations of Katz centrality. In contrast, non topological attributes are usually limited in number, and their values are often limited in variety. In particular, continuous values call for more condensed representations, like their statistical descriptions. Thus, non topological attributes are generally a good starting

point for explaining relationships in networks, since they provide a natural control for complexity.

2.1 Methodology with Probability Distributions

To test whether the topology of a network is explained by the non topological attributes of its vertices, one must find a way to explain the relations thorough non topological attributes and show that the topology of network induced by this explanation is similar to the network one is trying to explain. But testing graph similarity is a computationally hard problem: it is not known whether testing if two graphs have the same topology is in P or NP (graph isomorphism problem), and finding the largest identical portion of two graphs is NP-hard (maximum common subgraph isomorphism problem) [19, 20]. However, one workaround consists in evaluating graph similarity thorough topological measures, such as centralities [3, 21, 22]. Isomorphic graphs have exactly the same distributions of centralities, and when graphs are subject to small changes, centralities do not change much either [23], implying that similar graphs have similar centralities. Besides, they can be computed in polynomial time, often resorting to the computation of shortest paths, and several can be quickly approximated by random methods [24, 25].

We will assume that networks are generated by random processes instead of deterministic processes. Thus, we will focus on probabilistic explanations, i.e. functions that map the attributes of two vertices to the probability that there is an edge between them. The power of a probabilistic explanation can then be evaluated by sampling many networks and comparing their similarity to the original one.

Now, let us concern ourselves with building probabilistic explanations. Let $G = (V,E)$ be a directed graph, with vertex set V and directed edge set $E \subseteq V \times V$. Consider two vertices $X,Y \in V$ with attributes described in vectors $x = (x_1,x_2,...,x_n)$ and $y = (y_1,y_2,...,y_n)$, and assume that there is a probability $h(x,y)$ that there is a directed edge between them. The problem now consists in finding h. To find it, observe that an edge (x,y) (yes, as a $2n$-tuple of attributes) exists with probability $h(x,y)$ given that x and y exist. Let $f(x)$ be the distribution of the attributes of a vertex, $g(x,y)$ be the distribution of observed edges, $|V|$ be the number of vertices, and $|E|$ the number of directed edges. Then, the following relation holds:

$$|E|g(x,y) = |V|^2 f(x)f(y)h(x,y). \tag{1}$$

Thus, the previously unknown function h is found:

$$h(x,y) = \frac{|E|}{|V|^2}\frac{g(x,y)}{f(x)f(y)} = \rho\frac{g(x,y)}{f(x)f(y)}, \tag{2}$$

where ρ is the density of the directed network with loops. Note that Eq. 2 holds with both continuous and discrete probability distributions f and g. However, note that since h consists of a ratio, the distributions of g and specially f must be computed

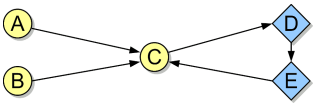


Fig. 1. Graph used in the example

carefully. If the variables are continuously distributed, then f and g must be estimated continuously, for example, by using kernel density estimators [26] or copulas.

To sample a network from these distributions, one must:

- sample $|V|$ vertices according to the probability distribution f, and
- for each vertex $X \in V$ with attributes x,
  - $\diamond$ for each vertex $Y \in V$ with attributes y,
    - $\circ$ add (X,Y) to the set of directed edges with probability $h(x,y)$.

From this sampling algorithm, it is easy to see that edges as pairs of attributes are distributed according to Eq. 2.

2.2 *Illustrative Example*

We now provide a simple, illustrative example of the methodology described in subsection 2.1. Consider a graph $G = (V,E)$ of vertices $V = \{A,B,C,D,E\}$ and directed edges $E = \{(A,C),(B,C),(C,D),(D,E),(E,C)\}$. Vertices come in two shapes: A, B, and C are circles ($\circ$), while D and E are diamonds ($\diamond$), as shown in Fig. 1. The density of the graph is $\rho = 5/25 = 1/5$.

The distribution of shapes in vertices is:

$$f(\circ) = 3/5, \quad f(\diamond) = 2/5,$$

and the distribution of shapes in edges is:

$$g(\circ,\circ) = 2/5,\ g(\circ,\diamond) = 1/5,\ g(\diamond,\circ) = 1/5,\ g(\diamond,\diamond) = 1/5.$$

Then, h is:

$$h(\circ,\circ) = (1/5)\frac{(2/5)}{(3/5)(3/5)} = 2/9, \quad h(\circ,\diamond) = (1/5)\frac{(1/5)}{(3/5)(2/5)} = 1/6,$$

$$h(\diamond,\circ) = (1/5)\frac{(1/5)}{(2/5)(3/5)} = 1/6, \quad h(\diamond,\diamond) = (1/5)\frac{(1/5)}{(2/5)(2/5)} = 1/4.$$

Assume we have sampled three graphs, and we are using their degree distributions to assess their similarity. (We define the degree of a vertex as the sum of its in and out-degrees.) After sorting them, the original degrees are $(1,1,2,2,4)$ and the sampled degrees are $(1,2,2,2,3)$, $(1,1,1,2,4)$, and $(0,2,2,3,3)$. Using the cosine to test the colinearity of these vectors, we get 0.9617, 0.9814, and 0.9231. We have not specified baselines (e.g. we could use an Erdös-Renyi random graph as a reference

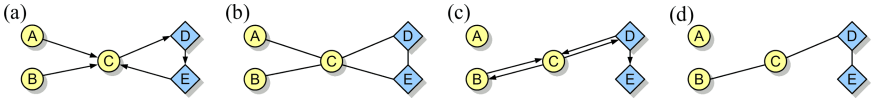


Fig. 2. Reductions to undirected graphs. The directed graph in (a) is undirected in (b) and the directed graph in (c) is undirected in (d).

for the degree distribution), but we can say that the generated graphs are do not differ much from the original one, regarding degrees. (Well, it is hard to arrive to conclusions with such a small statistical sample, but at the same time, classifying five edges in four categories ($(\circ,\circ)$, $(\circ,\diamond)$, $(\diamond,\circ)$, $(\diamond,\diamond)$) should not lead to graphs too different from the original one.)

2.3 Additional Considerations

Our first consideration regards how to apply the methodology to explaining and sampling undirected graphs. We recommend the simple solution of transforming the undirected graph $G_u = (V, E_u)$ into a directed graph $G = (V, E)$, such that $E = \{(X,Y) \in V \times V | \{X,Y\} \in G_u\}$. By applying the methodology to the directed graph G, we obtain functions f and h that explain both G and G_u thorough non topological attributes. To sample from these functions, we must introduce a relation $>_T$ on the vertices V such that $\forall X, Y \in V$, either $X >_T Y$ or $Y >_T X$. Then, to sample undirected graphs:

- sample $|V|$ vertices according to the probability distribution f, and
- for each vertex $X \in V$ with attributes x,
  - $\diamond$ for each vertex $Y \in V$ with attributes y, and such that $X >_T Y$,
    - $\circ$ add $\{X,Y\}$ to the set of undirected edges with probability $h(x,y)$.

Our second consideration regards the problem of obtaining an undirected graph from the explanation (functions f, h) of a directed graph. We will show that this is impossible. Have a directed graph $G = (V,E)$ and its functions f and h. The function h_u for the undirected graph $G_u = (V, E_u)$ where $E_u = \{\{X,Y\} | (X,Y) \in E\}$. Let $X, Y \in V$ be two vertices, and x, y be the non topological attributes of X and Y. Then, we could say that $h_u(x,y) = 1 - (1 - h(x,y))(1 - h(y,x))$. Now refer to the previous example, Fig 1. Consider $h(\diamond,\diamond)$ again, and let $h_u(\diamond,\diamond)$ be the probability two $\diamond$ vertices are connected in the undirected case. Then, $h_u(\diamond,\diamond)$ should be $1/3$ (with loops) or 1 (without loops). If we wrote the probability two different $\diamond$ vertices are connected in the directed model, we get $1 - (1 - h(\diamond,\diamond))^2 = 7/16$, which is different from either $1/3$ and 1. What happens is that the function that maps h from the directed case to the h_u of the undirected case does not exist. We can think of directed graphs G' and G'' that have $h' = h''$ yet $h'_u \neq h''_u$, proving that such function does not exist. See Fig. 2 for an example.

3 Data: The Chilean Shareholding Network

For our experiments, we considered the Chilean shareholding network. A shareholding network is a network where actors (vertices) represent firms or people and directed edges, or arcs, represent a shared ownership relationship [12, 13, 14]. For two actors X and Y, a directed edge (X,Y) means that X has some ownership on Y. The ownership relationship we model is the possession of shares or stocks. Thus, directed edges are weighted, because two actors may have different amounts of shares of a firm.

The Chilean shareholding network is monitored and regulated by the SVS, la Superintendencia de Valores y Seguros (the Superintendency of Stocks and Insurances). They keep records of which firms openly trade their shares in the stock market. For these firms, they keep data on a number of legal and financial matters. We are interested in their main shareholders (listed as the top 12) and financial statements, specially assets, equity, debt and profit.

3.1 Selecting Actors and Relations

We discarded all actors that were not in the listings of the SVS, to ensure that all of them were firms with financial statements. Thus, we discarded all natural people and several firms from the shareholding relation.

We also discarded shareholding relations with less than 5% of participation. Firms have small portions of ownership on other firms to obtain profits instead of partially controlling the decisions of the owned firms. But by considering only the top 12 shareholders, we are mostly considering relations of control. By discarding relations with less than 5% of participation, we are ensuring that we are only working with relations of control.

3.2 Matching Names to Identifiers

To construct the shareholding relation, we downloaded the list of shareholders of each firm in the listings of the SVS. These lists of shareholders were just lists of names and proportions of ownership. The relation is then constructed by matching the names of the shareholders of each firm to the names of the firms in the listings. However, the names of the shareholders were written informally, as opposed to the formal style used in the listings of firms monitored by the SVS. Moreover, the writing style of the shareholders varied from firm to firm. Therefore, we had to use a robust technique to match these names.

We matched shareholder names to firms as follows. First, we merged together all the names a firm lists (complete, fantasy, stock name) in one string. Then, we removed all special characters, most abbreviations and all connectives. We also abbreviated some words, to collapse similar words into the same. (E.g. `administrador`, `administradora`, and `administradores` were replaced by `adm`.) After this,

we transformed the string of names of each firm into a list of unique words. We counted how many times a word appeared in these lists, to keep a registry of which words identify firms better; common words do not help identifying firms while uncommon words may identify a firm completely. Finally, to identify a firm, we used the following formulas. Let S be a set of words and $\phi(w)$ be the frequency of word w in the lists (0 if not present). We then define a score for set S:

$$score(S) = \sum_{w \in S} \left(\frac{1}{1 + \phi(w)} \right)^{\gamma}.$$

Let S_X be the words in the name of a firm X and S_z be the set of words of the name of some listed shareholder. We then look for a firm X that maximizes $score(S_X \cap S_z)$. However, we accept this candidate firm X only if:

$$\frac{score(S_X \cap S_z)}{score(S_X)} \geq \theta. \tag{3}$$

We empirically adjusted the parameters γ and θ, so that the matching worked well, which was verified by inspection. We did not perform $score(S_X \cap S_z)/score(S_X \cup S_z) \geq \theta$ instead of Eq. 3 because shareholder names (S_z) were reported in various different writing styles.

After matching all firm names to shareholder names, we can represent the relations as tuples of attributes, allowing us to model the shareholding relation according to the methodology developed earlier.

3.3 Retrieving Financial Statements

Finally, we needed to retrieve the financial statements of the firms. These are the attributes of the actors in the network.

Financial statements were listed on the website of the SVS in three formats: as Web pages, eXtensible Business Reporting Language (XBRL) files, and Acrobat's Portable Document Format (PDF) files. We created parsers to retrieve the financial data displayed on Web pages and XBRL files, since they followed standardized formats. We processed PDF files manually since they often consisted of scanned images.

We retrieved financial statements from December 2009 since several firms did not report them from 2010 and on. If they were not available because firms reported late, we retrieved the ones from March 2009. If the firm was new, we retrieved the statements from December 2010.

Finally, we normalized all financial statements to Chilean Pesos (CLP) because it was the most common currency in the statements. We also retrieved only individual financial statements instead of consolidated financial statements.

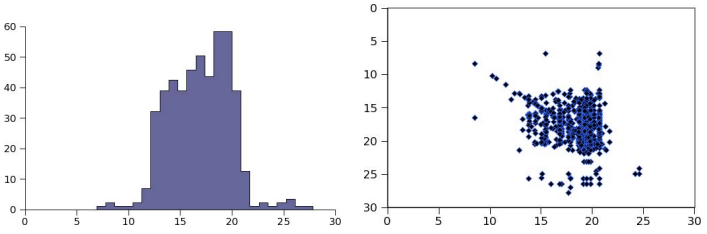


Fig. 3. Distribution of log-assets in vertices (left) and edges (right)

3.4 Resulting Network

We were only able to retrieve the financial statements of 427 firms, yet there were 568 firms in the listings. Of the 1108 relations, 787 were between firms with known financial data and solely 240 involved more than 5% of ownership. We retrieved a sparse graph.

We found that assets was the attribute with the most explanatory power. Equity and profits were strongly related to assets, and debt did not show explanatory power and had high concentration around small values, often being zero. Under the natural logarithm, we found that assets (log-assets) follow a bell shaped distribution, as shown in Fig. 3. We also plotted the edges as pairs of attributes (log-assets) in the scatter plot to the right of Fig. 3, which shows that firms are not strongly correlated thorough log-assets; correlation is about 0.24, but the points are mostly concentrated within a round cloud. (Save for a small line of loops and homophily.)

4 Experimental Evaluation

We now describe how we applied the methodology developed earlier to our data and then discuss the results obtained.

4.1 Evaluation Methodology

First, we built the probabilistic description of the network using kernel density functions [26], which distribute a probability cloud around each sample point. These probability clouds are called kernels, hence the name of the methodology. We used Gaussian kernels to transform each point into a narrow normal distribution. Let $X = \{x_1, x_2, ..., x_n\}$ be a set of observations, where each $x_i \in X$ is a vector in $\mathbf{R}^N$. Then, the kernel density estimator of the distribution of X is:

$$k(x) = \frac{1}{n}\sum_i K_b(x - x_i) = \frac{1}{n}\sum_i \frac{1}{\sqrt{2\pi}b} \exp\left(-\frac{||x - x_i||^2}{2b^2}\right), \tag{4}$$

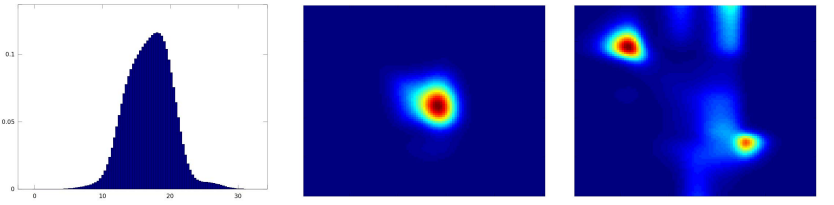


Fig. 4. Distributions f (left), g (center) and h (right). Axes are log-assets.

where K_b is a normal$(0, b)$ probability distribution function, N is the number of dimensions of the observations, and b is the *bandwidth*, the parameter that sets how wide a point is smoothed. We use the following asymptotic estimator for b:

$$b \approx 1.06\sigma n^{-1/5},$$

which approximates the optimal b as the number of observations in the sample increases. Here, σ is the standard deviation of the sample. We used it as the root mean squared distance to the average observation.

Following the observations made in subsection 3.4, we took the logarithm of the assets (which we call log-assets) before estimating the probability density functions. The results are shown in Fig. 4. Observe how f and g are smooth versions of the distributions previously plotted in Fig. 3. We did not use the correction for correlated data in the computation of g since the original observations exhibit a rounded distribution (Fig. 3).

We sampled 200 graphs according to these distributions. We used the distribution of degrees, closeness and betweeness centralities to compare the similarity between the original and sampled graphs. The comparison was performed as follows: both the centralities of the original and sampled graphs were sorted and then compared thorough correlation and cosine functions, and thorough norm ratios, to evaluate magnitudes. Norm ratios were evaluated as $|cent_{sample}|/|cent_{original}|$.

4.2 Experimental Results

Results are shown in Fig. 5. We can first see that comparisons thorough correlations and cosines are nearly identical, meaning that sampled centralities are proportional to the original ones. For degree centrality, we can see that the actual sampled graphs have smoother degree distributions than the original graph. This can be explained because we did not add the topological restriction of weights of the network. In particular, in-degree centrality is limited (requiring at least 5% of ownership limits in-degree to be at most 20), as opposed to out-degree. This was not modeled by the probability distribution. At any rate, we see that the sampled degree distributions have similar shape to the original one. Closeness centrality is also underestimated

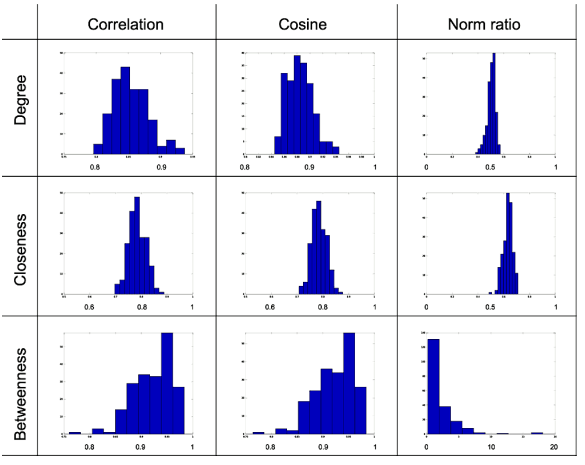


Fig. 5. Comparison of sampled centralities to the original ones

but bears a correlation of nearly 0.8. Betweenness centrality behaves quite differently. Correlations between the sampled and the original graphs are skewed and close to 1. In terms of magnitude, they are often close to 1, but the tendency is to overestimate the original betweenness centrality.

Deviations in closeness and betweenness centralities may arise from the greater sparsity and decentralization of the sampled graphs. Recall that closeness centrality measures the distances of the vertices to the rest of the network and that betweenness centrality measures the ratio of shortest paths that goes thorough each vertex. If the network is more sparse, distances become shorter because the connected components are smaller, so closeness centrality must become smaller. Regarding betweenness centrality, note that all vertices' centralities must sum 1. To increase the norm of such a vector, all it is necessary is to make its components more even. And this is what happens when a network becomes sparser and/or decentralized: the ratios of shortest routes thorough vertices become more similar. And so, betweenness centrality becomes larger in norm, which is what we observed in the experiments.

5 Conclusions

We developed a methodology to explain relations between vertices according to their attributes and applied it to an economic network. The methodology, which we summarize in Fig. 6, is very simple to use in practice, but is to be used with caution. While it is primarily aimed to directed graphs, we also discussed how to apply it to undirected graphs. We have not discussed how to choose graph similarity measures and how to interpret them. This is left for future work.

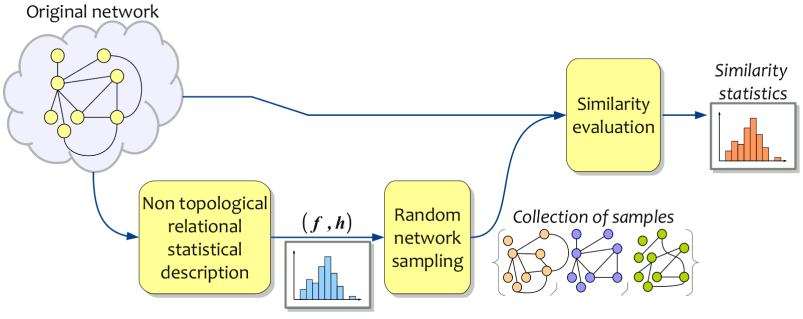


Fig. 6. Outline of the methodology

Note that the problem we consider is related to the purpose of doing research in network science: choosing the right level of abstraction. There is science at the macro level (macroeconomics, macrosociology, &c) and science at the micro level (microeconomics, microsociology, &c), and network science appeared as the bridge between both levels. But we are yet to identify the levels of abstraction to work with within the network abstraction.

Acknowledgements. This research is a continuation of the author's MSc Thesis, which was partially supported by grant Fondecyt 1070348 [28].

References

1. Barabasi, A.L.: Scale-free networks: a decade and beyond. Science 325(5939), 412–413 (2009)
2. Barabasi, A.L., Albert, R.: Emergence of scaling in random networks. Science 286(5439), 509–512 (1999)
3. Getoor, L., Diehl, C.P.: Link mining: a survey. SIGKDD Explor. Newsl. 7(2), 3–12 (2005)
4. Kenny, D.A., Kashy, D.A., Cook, W.L.: Dyadic data analysis. The Guilford Press, NY (2006)
5. Mizruchi, M.S., Marquis, C.: Egocentric, sociocentric or dyadic? Identifying the appropriate level of analysis in the study of organizational networks. Soc. Netw. 28(3), 187–208 (2006)
6. Domingos, P.: Prospects and challenges for multi-relational data mining. SIGKDD Explor. Newsl. 5(1), 80–83 (2003)
7. Getoor, L.: Link mining: a new data mining challenge. SIGKDD Explor. Newsl. 5(1), 84–89 (2003)
8. Popescul, A., Popescul, R., Ungar, L.H.: Statistical relational learning for link prediction. In: Proc. of the Workshop Learn Stat Model Relat Data, IJCAI (2003)
9. McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a Feather: Homophily in Social Networks. Annual Rev. Sociol. 27, 415–444 (2001)
10. Patil, A.N.: Homophily based link prediction in social networks. Tech paper. Stony Brook (2009)

11. Newman, M.E.J.: Assortative Mixing in Networks. Phys. Rev. Lett. 89, 208701 (2002)
12. Garlaschelli, D., Battiston, S., Castri, M., Servedio, V.D.P., Caldarelli, G.: The scale-free topology of market investments. Physica A: Stat. Mech. Appl. 350(2-4), 491–499 (2005)
13. Battiston, S., Glattfelder, J.B., Garlaschelli, D., Lillo, F., Caldarelli, G.: The Structure of Financial Networks. In: Estrada, E., Fox, M., Higham, D.J., Oppo, G.-L. (eds.) Network Science: Complexity in Nature and Technology, pp. 131–163. Springer, London (2010)
14. Scher, M.: Bank-firm Cross-shareholding in Japan: What is it, why does it matter, is it winding down? DESA Discussion Paper No. 15. ST/ESA/1999/DP.15. United Nations (2001)
15. Souma, W., Fujiwara, Y., Aoyama, H.: Heterogeneous Economic Networks. In: Namatame, A., et al. (eds.) Proc. of the Workshop on Economics and Heterogeneous Interacting Agents. Springer, Tokyo (2005)
16. Caldarelli, G., Battiston, S., Garlaschelli, D., Catanzaro, M.: Emergence of Complexity in Financial Networks. In: Ben-Naim, E., Frauenfelder, H., Toroczkai, Z. (eds.) Complex Networks. Lect. Notes Phys., vol. 650, pp. 399–423. Springer, Heidelberg (2004)
17. Piccardi, C., Calatroni, L., Bertoni, F.: Communities in Italian corporate networks. Physica A 389, 5247–5258 (2010)
18. Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F., Vespignani, A., White, D.R.: Economic Networks: The New Challenges. Science 325(5939), 422–442 (2009)
19. Köbler, J., Schöning, U., Torán, J.: Graph isomorphism is low for PP. Comput. Complexity 2, 301–330 (1992)
20. Bunke, H., Foggia, P., Guidobaldi, C., Sansone, C., Vento, M.: A Comparison of Algorithms for Maximum Common Subgraph on Randomly Connected Graphs. In: Caelli, T.M., Amin, A., Duin, R.P.W., Kamel, M.S., de Ridder, D. (eds.) SPR 2002 and SSPR 2002. LNCS, vol. 2396, pp. 123–132. Springer, Heidelberg (2002)
21. Tian, Y., Patel, J.M.: TALE: A Tool for Approximate Large Graph Matching. In: Proc. of the IEEE, 24th ICDE, pp. 963–972 (2008)
22. Spielman, D.A.: Spectral Graph Theory and its Applications. In: Proc. of the FOCS, pp. 29–38 (2007)
23. Borgatti, S.P., Carley, K.M., Krackhardt, D.: On the robustness of centrality measures under conditions of imperfect data. Soc. Netw. 28(2), 124–136 (2006)
24. Newman, M.E.J.: A measure of betweenness centrality based on random walks. Soc. Netw. 27(1), 39–54 (2005)
25. Jacob, R., Koschützki, D., Lehmann, K.A., Peeters, L., Tenfelde-Podehl, D.: Algorithms for Centrality Indices. In: Brandes, U., Erlebach, T. (eds.) Network Analysis. LNCS, vol. 3418, pp. 62–82. Springer, Heidelberg (2005)
26. Turlach, B.A.: Bandwidth selection in kernel density estimation: a rewiew. CORE and Institut de Statistique, 23–493 (1993)
27. Superintendencia de Valores y Seguros, http://www.svs.gob.cl (last accessed: November 10, 2011)
28. Monsalve, M.: A study of the structure and dynamics of the Chilean shareholding network. Dissertation, Universidad de Chile (2009)

Mapping Emerging News Networks: A Case Study of the San Francisco Bay Area

Daniel Ramos, Mehmet Hadi Gunes, Donica Mensing, and David M. Ryfe

Abstract. The news and information system in the United States is undergoing a significant transformation. From a limited number of professional, major metropolitan newspapers, television and radio stations to a networked system of hundreds of small and medium size information sources. Structural changes in news production and distribution are significantly altering the supply and flow of news to citizens. Using network analysis, we seek to map changes in the news ecology of the San Francisco Bay area. In this study, we graph the relationships between 143 locally based news sites to examine connections between news organizations, between journalists and their sources and between users of the news sites.

1 Introduction

The changing pattern of news production and consumption in the U.S. over the past 30 years is well documented. Newspaper circulation has declined by 31% and the percentage of people who watch an evening news program on a major American network has declined by 57% since 1990 [9]. The number of professional staff in newspaper newsrooms has declined by more than 25% in the past 10 years [5]. Meanwhile, nearly 60% of all Americans now access news online in a typical day [9] with the Internet now the third most popular news platform, behind local and national television news [7]. These audience consumption patterns reflect a technological, economic and social transformation that is causing significant changes in the news industry. The centralized, one-way distribution model of mass produced news is changing in response to a new communication structure that is far more decentralized, interactive and integrated.

The defining characteristic of mass media is the ability to broadcast messages from one-to-many points. Television, radio, magazines, newspapers and books

Daniel Ramos · Mehmet Hadi Gunes · Donica Mensing · David M. Ryfe
University of Nevada, Reno
e-mail: daniel.ramos@att.net, mgunes@cse.unr.edu,
{dmensing,dryfe}@unr.edu

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 237–244.
springerlink.com

have all been considered part of mass media, while telephones, the telegraph and the postal service are classified as narrowcast industries characterized by point-to-point message transmission [8]. The structure of the Internet, however, is a networked model of distributed communication with broadcast and narrowcast capacity. The reconfiguring of media relationships online is fundamentally reshaping the structure of social communication, including news and information, altering the role and function of message, audience, producer and production. The purpose of our study is to examine the network structure of news production in a specific region, the San Francisco Bay area, to begin understanding the shape and characteristics of this change. Specifically, we are interested in discovering the emerging shapes of the network for evidence of the formation of small worlds or other network configurations [12]. It is clear that changes of this magnitude are impacting the content of the news as well the public's access, response and participation in the information being transmitted, thus affecting many of the functions generally attributed to mass media.

The San Francisco Bay is recognized as one of centers of online innovation and is an ideal setting to analyze the network effects of a changing communication structure, given the high broadband penetration, sophisticated user base and concentration of online publishing experiments such as *Craigslist*.

2 Related Work

One of the earliest studies that applied network theory to online news sites analyzed the quantity of external hyperlinks accompanying individual stories [10, 11]. The author concluded that the increased use of internal links, rather than external, confirmed the preferential attachment theory of network formation and pointed to the evolution of particular stores, story topics and news organizations as hubs forming central nodes in a network. However, he also noted that should the traffic on non-media sites, such as blogs, increase, web editors may wish to reconnect with that network of users, thus altering what constitutes a central node. It appears that may be in fact what is happening as the number of non-professional, new news sites develop and grow. Tremayne [11] noted that the distribution of external links from news stories did not follow a normal distribution but a power distribution, with a handful of stories generating many of the external links.

A more recent network study analyzed 6,298 foreign news stories in 223 news web sites from 73 countries for their use of external hyperlinks [3]. The author found that news organizations rarely used external links, but when they did it followed patterns predicted by the preferential attachment theorem and the world system theory. The *world system theory* suggests that countries of the world can be categorized into three hegemonies: core, semi-periphery, and periphery based on their political and economic characteristics. The author found that only about 6% of foreign stories had one or more external hyperlinks. It is supposed that journalists are usually trained to provide only minimal access to their sources and that news organizations do not trust other sources to be accurate when distributing information, especially when in a different country than in which it operates. Furthermore, providing links to other sources could possibly report conflicting

information and cast doubt on the report containing the hyperlink. The author also concludes that since the media market is largely profit driven, that providing links to other sites would take the user away from the current site and decrease advertisement profits, supported by the finding that public broadcasters were more likely to use external hyperlinks. Finally, the author reported that the network formed by the study followed power-law degree distribution, with the US and UK attracting 10% of the hyperlinks and other countries attracting no more than 2% each.

Finally, a study analyzed Chicago area news websites and how they linked news and information [2]. Authors collected a list of 368 *seed sites* based on a survey and another web site that compiles news feeds by location. They used a web crawler to examine the links on these seed sites and recorded the links to other sites linked by the seed sites. They only recorded sites that were linked two or more times from the original sites to insure relevancy. They performed the process for three iterations and collected a final list of 277 sites. The sites were categorized into the following: *legacy* (i.e., traditional media brands), *legacy-affiliated* (i.e., publications owned by legacy brands), *micropublisher* (i.e., web sites focused on a topic or location), *organization/institution* (i.e., entities that would have, in the past, needed media organizations), *national brand* (i.e., websites of national scope with local presence), and *service* (i.e., websites that help publishers). They found that organizations are authorities, micropublishers and organizations are hubs. They also found that organizations are prominent intermediaries and organizations and some micropublishers are switchboards. From the previous results, it is clear that the authority of organizations/institutions is signaled by high number of inbound links. We will be testing the proportion of inbound and outbound links as part of this analysis, looking to distinguish the authorities, hubs and switchboards of the network.

3 Methodology

As exploratory research, our analysis was driven not by specific hypotheses but by questions about three types of relationships within the news network:

1. Relationships between news organizations,
2. Relationships between journalists and individuals/institutions/events external to the news organization, and
3. The relationship between the users of the Web sites and the organizations.

We explored these three relationships using social network analysis. In the following sections, we will describe the methodology of capturing each network.

3.1 News Organization to News Organization

The first network we were interested in capturing was the relationship between news organizations. Using search engines and links on news sites to search for relevant sites, we developed a list of websites that (1) produce news, (2) are updated

at least weekly and are (3) physically based in one of the nine San Francisco Bay area counties. The list includes organizations of all sizes from traditional media outlets to small blogs and other non-traditional formats. We compiled a list of 143 web sites we feel represent the news ecosystem of the San Francisco Bay area.

We used web crawling to discover the connections between organizations. For this, we modified the WebSPHINX crawler. All links to outside destinations were recorded to a database, including the number of times linked. A link was considered duplicate if it had the same domain name up to the subdomain level (e.g. `www.example.com/page1.html` is the same as `www.example.com/page2.html`, but `other.example.com` is not). We made this delineation because many of the sites in our list used hosting services like *BlogSpot* or *WordPress* and therefore all have the same second-level domain name.

Next, using the data from the crawl of each seed site, a directed, weighted graph was generated for each site in which each site and its linked sites were nodes. A directed edge is drawn from the seed site to the outside site indicating the one-way nature of a hyperlink. The weight of the edge was calculated by how many times that particular domain was linked to from the seed site. The next phase merged all the graphs for each seed site into a large graph containing all news organizations. From this graph we measured *in-degree* and *out-degree* to examine which news organizations are authorities or hubs, *betweenness* to examine which sites might link otherwise unconnected sites, and *PageRank* to examine which sites are more important in terms of the number links.

Most of the seed sites were crawled successfully but there were some sites that were not crawled either because they were not accessible or they prevented web crawlers from accessing them via the *Robots Exclusion Standard* [6]. In all, 118 of the original 143 sites were crawled in some way, shown in Fig. 1. Some of the larger sites were only partially crawled due to their size. In this case, we tried to focus on content from recent years. We did not include edge weights because many of the sites have standard headers, footers, and navigational sidebars that appear on many pages. Links in these page elements increased the reference count artificially and skewed our metrics.

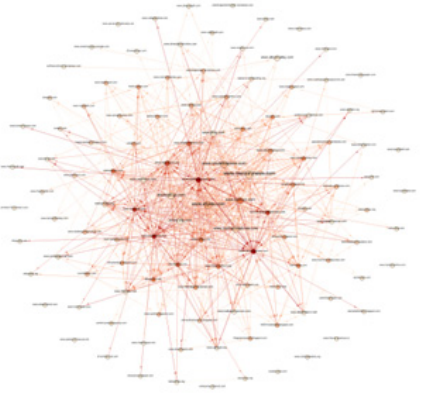


Fig. 1. News organization network

The average incoming and outgoing links for a site in the network shown in Fig. 1 is approximately 12 with the in and out degree averages being very close to 6. There are 43 sites that did not link to another news organization, but most are linked to by another site. Not surprisingly, the sites with the most incoming links (i.e., authorities) were traditional news organizations such as the *San Francisco Chronicle*, *San Jose Mercury News*, and the *Oakland Tribune*. A large

number of links to an organization most likely indicates a greater level of authority and completeness compared to other sites. Similar to [3], the top eight sites for incoming links, traditional news organizations, have a very low number of outgoing links. We theorize this is because traditional news reporters are likely to follow the mass media model of news reporting without adapting to a networked structure.

The sites with the most outgoing links (i.e., hubs) are generally independent or non-traditional news organizations. This fits the expectation that non-traditional or newly created news media sites are more likely to utilize the full potential of the network and provide links to sources and supplemental information about the story being reported.

We used the centrality measures as another way to evaluate the relative importance of a site in the graph. As expected, the sites we previously found to be authorities (i.e., the traditional news organizations) had the highest page ranks. The alternative news site *The Easy Bay Express* has the highest betweenness ranking. *The East Bay Express* receives almost as many incoming links as it has outgoing. Also, it links and is linked to by both traditional and non-traditional sites. Therefore, it could be considered a bridge site between news organizations.

3.2 Journalists to the Community

The second network we were interested in was the linking patterns of journalists. We used modified WebSPHINX to crawl the seed sites mentioned previously. We crawled only the larger sites that had author credits clearly identified for each story. The challenges in creating a graph of this network included the fact that some sites do not have author credits in their articles nor is there any standard format for website bylines. We had to develop site-specific web crawling rules so the crawler knew where to find the author's information. The crawler recorded each author found and each external hyperlink reference in their stories. Duplicates were not counted as in the method mentioned previously.

The network formed for journalists was a bipartite graph with the sets being all the journalists discovered and the other set was all the external sites referenced in their articles. Using this graph, we were able to measure the following metrics: the *degrees of the journalists* (which journalists used hyperlinks most often) and the *degrees of the sites* (which sites are linked to most often by journalists). Sites not considered community (e.g. advertisements) were removed manually.

We crawled three large sites representing three different types of news media: www.baycitizen.org as independent, laurendo.wordpress.com as a blog, and www.sfgate.com as traditional. We only examined links in news stories that were reported for the past few years (mainly 2010). As before, non-community sites were removed.

The Bay Citizen's journalists are in the middle range in terms of how frequently they link to external sites. The range of external links is between 4 and 75 with the top five journalists all having above 30 links. A minimum of four links are automatically included on every news story to facilitate sharing on various social networking sites. Besides these sites, the most referenced sites are *Oakland North*,

YouTube, *Berkeleyside*, *Wikipedia*, and the *San Francisco Chronicle*. Since *The Bay Citizen* is established in 2010, it tends to utilize newer journalistic practices by providing references to outside sources, but clearly these practices are unevenly spread among writers.

Blogging Bayport Alameda is a blog about local news in Alameda written by one author, Lauren Do. Do is very prolific and has linked to thousands of external sites in her blog. As a personal project, Do does not host any advertising on her site nor does she attempt to cover all the issues and events mentioned on her blog.

Despite the *San Francisco Chronicle's* larger size, it had fewer journalists than *The Bay Citizen* and did not link to many outside sources. The range of external links is between 3 and 14 with the top five journalists all having between 8 and 14 links. Like *The Bay Citizen*, the minimum 3 links are on every news story as a method to share it on various social networking sites. No other sites are heavily linked. The lack of external links is parallel to the results of the study by [3]. Interestingly, the journalists who did utilize external links more frequently are freelancers and other non-staff.

It is clear the importance and popularity of social networking has reached most news organizations. Even if a site generally did not link to external sites, it still provided links to share its stories on multiple social networking sites. This practice alone insures some connection to the network, even if generated automatically.

3.3 Commenters to News Organizations

The final network we were interested in is the links between user commenters on news web sites and how they interact with news organizations and other commenters. A recent feature of many of the larger news sites is the ability for the reader to comment directly on a news story. This feature highlights the collaborative nature of the Internet.

We focused on a small subset of seed sites that had story comment features and a larger population of registered users. It also required site-specific web crawling rules so the crawler knew where to find the comments section and the commenters' names. The crawler recorded each user found and each story they commented on. A manual coding process was required to remove spam bots and other non-human commenters. Since sites do not have a common user pool, each site was its own graph.

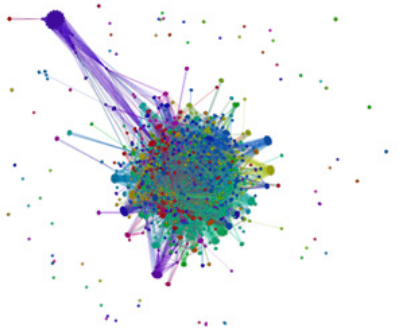


Fig. 2. *Berkeleyside* commenter network

The first networked formed was a bipartite graph. The set of nodes were all the different users found and the second was all the different stories commented on. We were able to measure the *degrees of the users* to determine which users are most prolific in their commenting and the *degree of the stories* to examine which stories garnered the most attention.

We then transformed the first network into a 1-mode network by removing the story nodes from the graph. Using this new network we were able to tell which users are connected together, meaning they commented on the same stories. The edge weights in this new network were the number of stories that both users commented on. Using this network we were able to tell if any users form clusters (i.e., commenting on the same stories) and how well connected these communities might be. To create the initial networks, we crawled two independent sites: berkeleyside.com and socketsite.com. We found that traditional news sites, tended to use JavaScript based commenting systems that prevented our crawler from recording. We limited the results to only news stories occurring in 2010. We also manually removed comments that were clearly spam or an automated posting.

On *Berkeleyside*, the stories with the most commenters are on a wide variety of subjects, but many of the stories are about retail and commerce. *Berkeleyside* has a fairly active community, shown in Fig. 2, with the top 4 users commenting on over 100 stories and the top user commenting on 60% more stories then the next highest user. When converting the network to only contain users, we found that a user was linked to, on average, 19 other users via commenting on the same story. The commenters were be divided into 142 different *communities* using the Louvain method [1]. These communities also include users who are the only commenter on a story and therefore are a community by themselves (as seen by the many single nodes in Fig 2). The top two communities contain 29.75% (17.34% and 12.41%) of users which would suggest there is a large portion of users who frequently comment on the same stories.

Socket Site mainly focuses on real estate, so most of the top commented stories are in regards to that subject. Since Socket Site is a niche topic site, the community is smaller than *Berkeleyside*, but the top 4 posters still have commented on 80-100 stories each. The community for *Socket Site*, shown in Fig. 3, appears to be more closely knit than *Berkeleyside*. We find that a user was linked to, on average, 57 other users. The commenters on this site could be divided into 19 different communities where the top 3 communities contain 53.36% of all users. This tighter knit group might be explained by the specific focus on the site which fosters more. There is also the possibility that real estate professionals who work together also interact on this site.

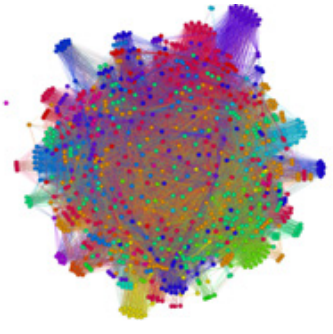


Fig. 3. *Socket Site* commenternetwork

4 Conclusion

This study represents an initial step in understanding the emerging news networks focusing on the San Francesco Bay Area. The network graphs reveal a distinct difference between the linking patterns of traditional news media sites and newer

and/or alternative news sites. These results were mirrored when examining individual reporters on the various types of sites. We found that some small percentage of journalists in each of our examined news sites tended to link far more than the average journalist. Finally, we found that commenters on sites sometimes form communities in which they often comment on the same stories. This seems to happen even more if a site is primarily about one topic.

These results indicate that the structure of the news ecology in the Bay area is indeed changing. New patterns of relationships, production and distribution are evident. Newer news organizations are facilitating practices that set them apart from older news organizations that came of age in the mass media era. Some journalists are utilizing practices different from others, building content conducive to networking and linking rather than stand alone, authoritative reports. News users are organized according to a predictable range of relationships, knowledge of which could facilitate better communication on news sites.

References

1. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of community hierarchies in large networks. Journal of Stat. Mech (2008)
2. Gordon, R., Contractor, N., Johnson, Z.P.: Linking Audiences to News and Information: A Network Analysis of Chicago Websites (September 23, 2010)
3. Himelboim, I.: The International Network Structure of News Media. Journal of Broadcasting & Electronic Media 54(3), 373–390 (2010)
4. Jarvis, J.: New rule: Cover what you do best. Link to the rest in Buzzmachine (2007)
5. Kirchhoff, S.M.: The U.S. Newspaper Industry in Transition. Congressional Research Service 7-5700 (2010)
6. Koster, M.: A standard for robot exclusion (June 30,1994)
7. Purcell, K., Rainie, L., Mitchell, A., Rosenstiel, T., Olmstead, K.: Understanding the participatory news consumer. Pew Internet and American Life Project (March 2010)
8. Rosse, J., Dertouzos, J.: Proceedings of the Symposium on Media Concentration, vol. 1. Bureau of Competition, Federal Trade Commission (December 14, 1978)
9. State of the News Media, An Annual Report on American Journalism. Pew Project for Excellence in Journalism (2010)
10. Tremayne, M.: The Web of Context: Applying Network Theory to the Use of Hyperlinks in Journalism Stories on the Web. Journalism & Mass Communication Quarterly 81(2), 237–253 (2004)
11. Tremayne, M.: Applying Network Theory to the Use of External Links on News Web Sites. Internet Newspapers: Making of a Mainstream Medium. Lawrence Erlbaum Associates (2006)
12. Watts, D.J.: Networks, Dynamics, and the Small-World Phenomenon. The American Journal of Sociology 105(2), 493(435 pages) (1999)

Identifying Critical Road Network Areas with Node Centralities Interference and Robustness

Giovanni Scardoni and Carlo Laudanna

Abstract. We introduce the notions of centrality interference and centrality robustness, as measures of variation of centrality values when the structure of a network is modified by removing or adding individual nodes from/to a network. Centrality analysis allows categorizing nodes according to their topological relevance in a network. Thus, centrality interference analysis allows understanding which parts of a network are mostly influenced by a node and, conversely, centrality robustness allows quantifying the functional dependency of a node from other nodes in the network. We examine the theoretical significance of these measures and apply them to classify nodes in a road network to predict the effects on the traffic jam due to variations in the structure of the network. In these case the interference analysis allows to predict which are the distinct regions of the network affected by the function of different nodes. Such notions, when applied to a variety of different contexts, opens new perspectives in network analysis since they allow predicting the effects of local network modifications on single node as well as global network functionality.

1 Introduction

Study of complex networks currently spans several disciplines, including biology, pharmacology, economy, social science, computer science and physics [1]. One of the major goals of modern network science is the quantitative characterization of network structure and functionality with the purpose of inferring emergent properties of complex systems, abstracted as networks and represented as graphs [2]. Notably, since network structure always affects function [3], the topological analysis approach allows understanding networks functionality through the analysis of their

Giovanni Scardoni
Center for BioMedical Computing (CBMC), University of Verona
e-mail: giovanni.scardoni@gmail.com

Carlo Laudanna
Department of Pathology, University of Verona
e-mail: Carlo.laudanna@univr.it

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 245–255.
springerlink.com

specific structure. For instance, the topological structure of the road network affects critical traffic jam areas, the topology of social networks affects the spread of information and diseases, and the topology of electrical grids affects the robustness and stability of energy distribution. Remarkable results have been reached in this field and, even if far from being complete, several key notions have been introduced. These unifying principles underlie the topology of networks belonging to different fields of science [4],[5],[6],[7],[8],[9]. Currently, network analysis mainly focuses on global network properties and on their global modifications [10]; [11]; [12]; [13] as for example in the case of the vitality index [9] or attack tolerance of networks [14]. Recent fundamental results [15] show how analysis on the topology of the network allows identifying the driving nodes of a network, i.e. the nodes that have to be controlled in order to control the entire network, suggesting that identification of these nodes depends on the network topology and not on the network dynamics. These results may suggest the utility of a deeper analysis of biological networks, with the purpose of analyzing not only global network properties, but especially local properties affecting those nodes that are, more than others, central to the global functionality of the network. In this context, network centralities, such as degree, eccentricity, closeness, betweenness, stress, centroid and radiality [9]; [16]; [17] are topological parameters allowing understanding the importance of single nodes in a network.

Here, we introduce the notion of centrality interference and robustness, as measurements of changes in the local topological structure of the network as a consequence of single nodes removal or addition, in order to quantify the influence of single nodes in different parts of the network. Our approach allows addressing the following question: "if we remove or add one node in the network, how do other nodes modify their functionality because of this removal?". In some cases, such as in social and financial networks, the structure of the network is naturally modified over time; in other cases this can be due to specific network changes: power grid failures, traffic jam or work in progress in a road network, temporary closure of an airport in an airline network and so on. In a biological network one or more nodes (genes, proteins, metabolites) are possibly removed from the network because of gene deletion, pharmacological treatment or protein degradation. Understanding the topological consequences of such changes in the network means to understand how the network functionally rearranges. For instance in the case of a drug treatment, we can potentially predict side effects of the drug by looking at the topological properties of nodes in a drug-treated network, meaning with that a network in which a drug-targeted node (protein) was removed. Similarly, we can understand new critical traffic points in a road or airline network after a modification of its structure. Notably, our perspective concerns node-by-node modifications: a single node modification can be irrelevant to the overall organization of the network (for instance its scale-free structure), but can profoundly modify the properties of one or more nodes in different regions of the network, thus changing, for instance, the network modular structure. Since centralities are single-node properties, the effects of single node alterations can be calculated by analyzing modifications of centralities values due to single node alterations. As the centrality value of a node is strictly dependent on the network structure and on the properties of other nodes in the network, if we

add or remove a node in the network the consequences of this modification on the network structures are reflected on the centrality values of all the other nodes. Such a situation, similarly to the case of interference for computer programs [18] can be analyzed introducing the notions of centralities interference and robustness. We first introduce interference for the betweenness centrality. The interference definition can be applied to other centrality measures. All definitions consider connected networks (i.e. networks where each node is reachable from all the others) and that remain connected also after nodes deletion.

2 Betweenness Interference

We consider a network as a graph $G = (N, E)$ where N is the set of nodes and E is the set of edges. Betweenness of node n is defined as $Btw(G,n) = \sum_{s \neq n \in N} \sum_{t \neq n \in N} \frac{\sigma_{st}(n)}{\sigma_{st}}$ where σ_{st} is the number of shortest paths between s and t and $\sigma_{st}(n)$ is the number of shortest paths between s and t passing through the vertex n. We consider the relative value of betweenness normalizing it as $relBtw(G,n) = \frac{Btw(G,n)}{\sum_{j \in N} Btw(G,j)}$ in order to have the fraction of betweenness of each node with respect to the rest of the network. Consider the example in figure 1a. If we remove node k from the network, node b become the only node connecting a to all the other nodes in the network (fig. 1b), so its betweenness value will increase. This is a case of betweenness interference since removing node k from the network "interferes" with the betweenness value of node b and can be measured as follow. $G_{|i}$ is the network obtained from G removing node i and all its edges from the network. The *betweenness interference* of node i with respect to node n in the network G is $Int_{Btw}(i,n,G) = relBtw(G,n) - relBtw(G_{|i},n)$. The measure shows which fraction of betweenness value a node loses or gains with respect to the rest of the network when the node i is removed from the network. The interference value can be positive or negative. If it is negative, it means that the role of node n in the network is higher when the node i is not present in the network. So we can say that node i has *negative* interference on node n, in the sense that the

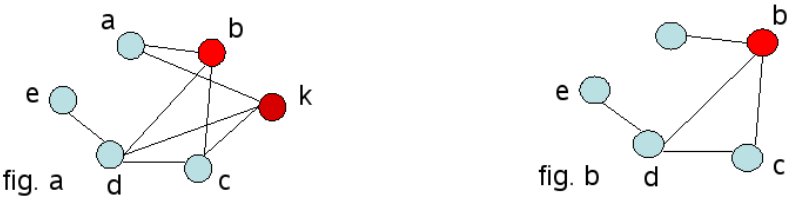


Fig. 1. a. Node k and b are in the shortest paths from node a to the other nodes. **b**. Node k has been removed. Node b is now essential for connecting node a to the rest of the network: it is the only node in the shortest paths connecting a to the other nodes: node b betweenness increases.

presence of node i in the network is "negative" for the node n to play a "central role" in the network. If the interference value is positive, it means that betweenness value of node n is higher if node i has been added to the network. In this case we say that i has *positive* interference on node n, in the sense that the presence of node i is "positive" for node n to play a "central role" in the network. The meaning of negative and positive interference strictly depends on the kind of network they are applied to.

Note. Even if interference is calculated removing a node from the graph, it is a measure of the influence of this node with respect to the rest of the network. Besides it can also be used to model some frequent situations where nodes are added or removed to/from a network. In these cases adding a node means to add a node whose interactions are known. As example, adding a protein to a protein-protein interactions network we exactly know its interactions with other proteins (the new edges to add to the graph).

3 Centralities Interference definitions

The notion of interference can be easily extended to other centrality values and other interference based measures as *modulus interference* ($ModInt_{Btw}(i,n,G) = |Btw(G,n) - Btw(G_{|i},n)|$) and *absolute interference* ($AbsInt_{Btw}(i,n,G) = Btw(G,n) - Btw(G_{|i},n)$) can be used to enrich the analysis. Finally, a successive step for a complete analysis of interference is to quantify the interference of a single node with respect to the entire network. The question is: How important is node i for the functionality of the entire network? A node can interfere with high value with respect to few nodes and can have low interference value with respect to many others. Alternatively one node can interfere with significant values with respect to the most of the nodes in the network. In the second case the node can have importance for the entire network functionality and not only for one or few nodes. In order to quantify the interference with respect to the entire network we introduce the *max* interference value and the *global* interference value. The *betweenness max interference* value of node i is defined as $maxInt_{Btw}(i,G) = \max_{n \in N \setminus \{i\}} \{Int_{Btw}(i,n,G)\}$. If it is high at least one node is consitently affected by node i. The *betweenness global interference* value of node i is $Int_{Btw}(i,G) = \sum_{n \in N \setminus \{i\}} (Int_{Btw}(i,n,G))$. If it is high the nodes interferes with high values with respect to the most of nodes in the network. In order to compare different networks these two values can be normalized by dividing them by $|N| - 1$ where $|N|$ is the number of nodes of the network.

4 Centralities Robustness, Dependence and Competition Value

We approach now the reverse problem of interference: we know that a node has a central role in the network and we would like to know if its functionality can be affected by other nodes and how much. The question is, conversely to interference: "which are the nodes affecting node n?". To answer to this question we introduce the

notion of robustness, competition and dependence value of a node. The *betweenness robustness* of node n is defined as $Rob_Btw(n,G) = \frac{1}{\max_{i \in N_{|n}} \{|Int_{Btw}(i,n,G)|\}}$. Robustness depends on the maximum interference value that can affect the betweenness value of the node. If it is low, the node can be easily "attacked" by removing or adding particular nodes. If it is high, the node is "robust", i.e. there is no node removal or adding that can affect its betweenness value and consequently its functionality. Note that we consider the modulus value of interference. Similarly to interference, positive and negative robustness can be defined but it is more intuitive to consider their reciprocal values, respectively dependence and competition. The *dependence value* is $Dep_{Btw}(n,G) = \max_{i \in N_{|n}} \{Int_{Btw}(i,n,G)\}$ where $Int_{Btw}(i,n,G) \geq 0$. If it is high, this value means that the node is "central" because of the presence of at least another node in the network: if that node is removed then node n loses a consistent part of its central role (its centrality measures decreases). If low the central role of node n is not dependent on other nodes and there is no node removal that can consistenly affects its relevance in the network. Similarly we define the *competition value* as $Comp_{Btw}(n,G) = \max_{i \in N_{|n}} \{|Int_{Btw}(i,n,G)|\}$ where $Int_{Btw}(i,n,G) \leq 0$. High competition value means that the central role of node n can be "improved" removing a particular node from the network (node n betweenness increases). In this sense the two nodes, node n and the removed one are "competitors" in the network. If low, the central position of the node cannot be improved removing a particular node from the network. Because of our specific focus on single node analysis, the betweenness variation due to robustness, competition and dependence can be related to the betweenness value of the node in the starting network (the network with no node deletion) ($relRob_{Btw}(G,n) = \frac{Rob_{Btw}(G,n)}{relBtw(G,n)}$, $relDep_{Btw}(G,n) = \frac{Dep_{Btw}(G,n)}{relBtw(G,n)}$, $relComp_{Btw}(n,G) = \frac{Comp_{Btw}(n,G)}{relBtw(n,G)}$ Similarly to the interference definitions, total robustness dependence and competition value can be also used as global parameters in order to characterize the entire network. All robustness, competition and dependence definitions can be extended to other centrality values. Next example shows the role of node centrality robustness, dependence and competition value. Consider the network in figure 2a. Node3 and node6 have the highest values of betweenness (25.64), node4 and node5 present the third highest value (12). A Robustness analysis

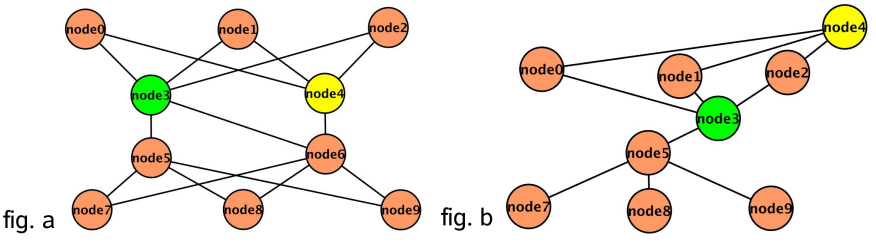


Fig. 2. Node3 and node6 have highest betweenness (25.64). Betweenness value of node4 and node5 is 12.

of node3 and node4 allows to understand if and to what extent their high betweenness values depend on other nodes of the network. Node3 has higher robustness value (0.046) than node4 (0.036). In fact node4 is in the shortest paths connecting node0, node1 and node2 with node7, node8 and node9 (fig. 2a), but if we remove node6, node4 loses this role and becomes a "peripheral node" connecting only node0, node1, node2 between them (fig. 2b). This can not happen to node3 since it is connected to both node6 and node5. Node3 has highest dependence on node5 equals to 0.0999. The relative dependence value is 0.3118 indicating that node3 loses about the 31% of its starting betweenness value if node5 is removed from the network. Indeed, if we delete node5 the betweenness value of node3 becomes the same as node4, since they connect the same nodes through the same paths: those passing through node6. But dependence of node4 on node6 is higher (0.1143, with relative dependence 0.7619 i.e it loses about 76% of its starting betweenness value if node6 is removed from the network): as previously seen, if we remove node6 then node4 becomes a "peripheral" node and node3 becomes the only way to connect the "top" of the network with the "bottom". Also the competition value of both nodes is very informative. The highest value of node3 depends on deletion of node4 and the highest value of node4 depends on node3. In this sense they are really "competitors" in the network. But this also means that, missing one of the two nodes, its role can be replaced by the other one. If we remove node3 then node4 becomes the only connection between the "top" and "bottom" of the network. The same for node3 if we remove node4. But node4 competition values is higher (0.2786 vs 0.2162). This is due to the fact that starting betweenness value of node4 is lower (12) than node3 value. So the increase of betweenness of node4 is higher, the 185% of the starting value.

5 Interference in a Transportation Network: The Case of Italy North-East Highway

We applied interference to the highway network of the north-east of Italy, the region included between Milan, Bologna and Trieste (see fig. 3). The network, containing 136 nodes and 144 edges has been compiled with the distance in minutes between each highway exit as reported by the official Italy highway website [19]. We chose three highway exits as example to evaluate betweenness interference: Melegnano, Como, Mestre. Positive and negative values for each of these exits are reported in figures 4 and 5. Due to lack of space only the first ten values are reported, but they are enough to illustrate the notion of interference in a real world example. Firstly we analyzed Melegnano betweenness interference: Melegnano is a critical node to connect the Milano area with the Bologna one. Closing the highway in this point means to stop the main traffic from Milano to Bologna. As expected the first ten positive interference values are all the towns between Milano Sud and Parma (see fig. 6, the red road). This region is the one that is more affected by Melegnano. If Melegnano is part of the network, these towns are in the shortest way to connect Milano and its area with Bologna and its area. This is perfectly captured by the

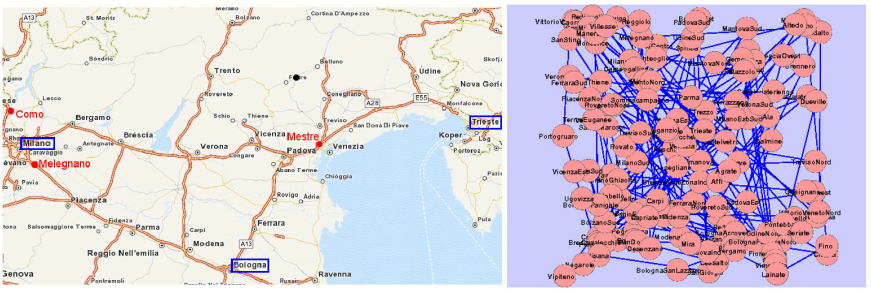


Fig. 3. The north-east of Italy highway network and its representation as a graph

Melegnano

Node name	Betweenness Interference
MilanoSud	0.33
Lodi	0.25
Sesto	0.24
Casalpusterlengo	0.23
PiacenzaNord	0.2
Fidenza	0.18
Fiorenzuola	0.18
Parma	0.18
PiacenzaSud	0.18
ReggioEmilia	0.18

Como

Node name	Betweenness Interference
Fino	0.08
Lainate	0.07
MilanoNord	0.05
Agrate	0.04
Cavenago	0.04
MilanoEst	0.04
Monza	0.04
Trezzo	0.04
Bergamo	0.03
Capriate	0.03

Mestre

Node name	Betweenness Interference
Mira	1.77
PadovaEst	0.06
Mirano	0.05
Grisignano	0.04
Montebello	0.04
Montecchio	0.04
PadovaOv...	0.04
Soave	0.04
VeronaEst	0.04
VicenzaEst	0.04

Fig. 4. First ten positive interference value of Melegnano, Como and Mestre

Melegnano

Node name	Betweenness Interference
BresciaCentro	-0.23
BresciaOvest	-0.2
Ospitaletto	-0.2
Grumello	-0.19
Manerbio	-0.19
Palazzolo	-0.19
Ponteoglio	-0.19
Pontevico	-0.19
Rovato	-0.19
Bergamo	-0.18

Como

Node name	Betweenness Interference
PadovaEst	-0.03
Grisignano	-0.02
Mestre	-0.02
Mira	-0.02
Mirano	-0.02
Montebello	-0.02
Montecchio	-0.02
PadovaOvest	-0.02
VeronaSud	-0.02
VicenzaEst	-0.02

Mestre

Node name	Betweenness Interference
Spinea	-1.72
Preganziol	-1.66
Venezia	-1.06
Ala	0.0
Belluno	0.0
BolognaArcoveggio	0.0
BolognaFiera	0.0
BolognaPanigale	0.0
BolognaSanLazzaro	0.0
BolzanoNord	0.0

Fig. 5. First ten negative interference value of Melegnano, Como and Mestre

positive interference of Melegnano with the highway exits of this region. If Melegnano is removed from the network, for example if it is blocked by a road accident, the road between Milano sud and Parma can not connect Milano and Bologna. To understand the alternative paths, we consider the negative interference of Melegnano. As reported in figure 5, the first ten negative values belong to the region around Brescia Centro. As expected, if Melegnano is blocked, the interference analysis predicts that Brescia Centro is the new critical point to connect Milano and Bologna, through the highway from Brescia Centro to Fiorenzuola (see fig. 6, the blue road). Even if these nodes are far from Melegnano in the network, the interference analysis can easily predict that they are undirectly influenced by Melegnano.

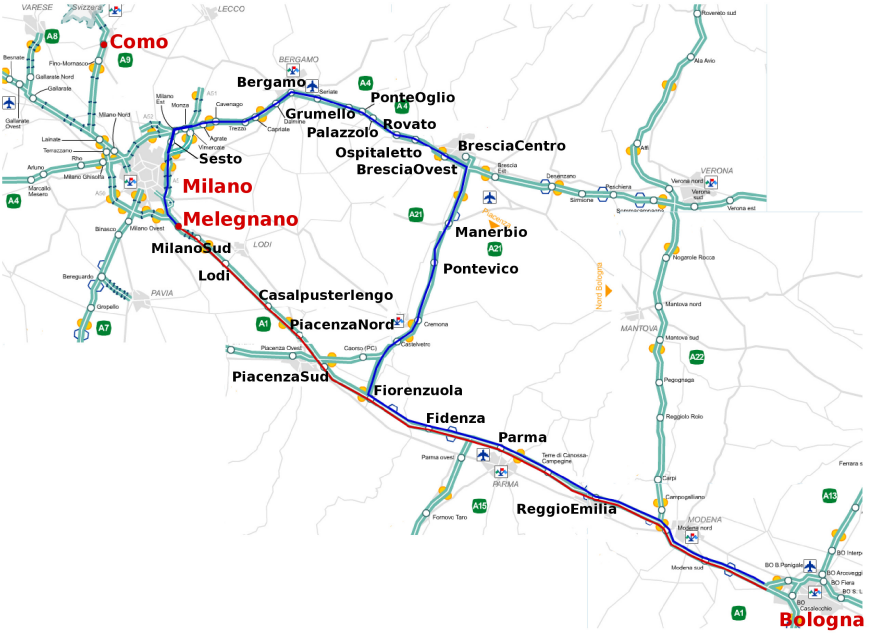


Fig. 6. The shortest road from Milano to Bologna passes through Melegnano (red road). If Melegnano is blocked, the shortest road is the one passing through Brescia (blue road). This behaviour is exactly predicted by the interference analysis.

As a second example we computed Como interference. In the north-east highway Como is only a peripheral node (see fig. 6). As expected its interference value is high only for its neighbour, and susbstantially smaller than interference of Melegnano (Como interference max value = 0,08, Melegnano interference max value 0,32). This shows that interference analysis really reflects the importance of undirect interaction between nodes. As third example we computed interference for Mestre (see fig. 7). Mestre is well known as an important connection between Trieste (the extreme east of Italy) and important nodes as Milano and Bologna. Its interference analysis results in high negative interference with respect to Spinea, Preganziol, Venezia. This is totally in agreement with the real situation: the road passing through Spinea, Preganziol and Venezia called “passante of Mestre” was recently built in order to solve traffic jam problem of the Mestre Area, always congested because of traffic from Milano and Bologna to the Venice port and to Trieste. To confirm this analysis we can modify the distance in minute between Mira and Mestre. In high traffic condition, as for example during summer weekends when a lot of people moves to the Venice area for holidays, the real distance between Mira and Mestre is more than 20 minutes. In this case the shortest path connecting Trieste with Milano and Bologna is the one passing through Spinea. We modified the distance of the network according with these value. An interference analysis of Spinea in the updated

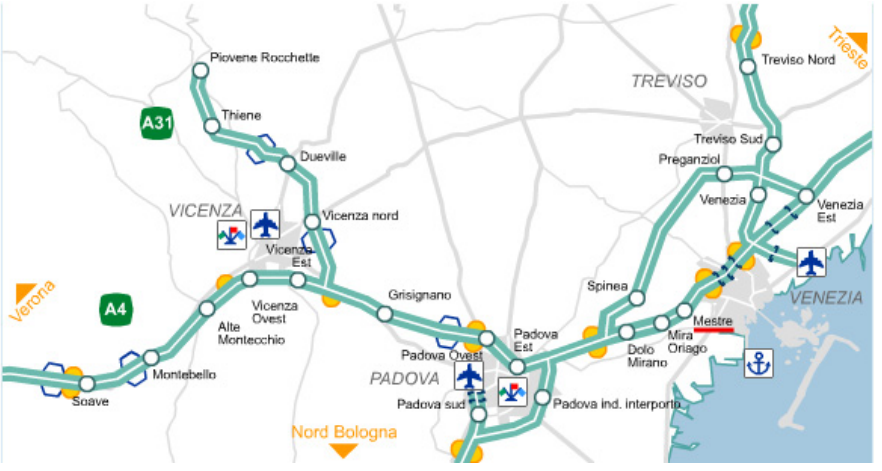


Fig. 7. Mestre is a critical node to connect Trieste with the rest of Italy. Note the recently built alternative path passing through Spinea, called "passante of Mestre"

network, shows high negative interference (= -1.7) with respect to Mira and Mestre. As expected, according to the interference analysis, the role of Spinea is exactly to reroute the traffic of Mestre: if Spinea is not part of the network, its negative interference with Mestre and Mira predicts that Mestre and Mira are more congested than they were before the "passante of Mestre" building.

6 Further Considerations and Conclusions

As showed above, the interference analysis allowed identifying critical areas in roads network. This doesn't result in a real dynamic prediction of traffic jam but, only through the analysis of the network structure, we have been able to identify those parts of the network that more than others can be affected by particular modification of single nodes (traffic jam, closure of an exit, work in progress). As explained in the introduction, the interference and robustness analysis can be applied also to several other kinds of network (biological networks, social networks, electrical grids, transportation networks and so on) and to other centralities measures. So for any case study, the methodology and the interpretation of the analysis strictly depends on the kind of network and the kind of centrality that is used. As a further implication of our approach, we can consider centrality interference and robustness as natural generators of network modularity. Indeed, a new clusterization algorithm can be derived if we group nodes depending on their interference value. Given a node, we may compute its interference activity of the network and, then, we may group in the same cluster all nodes having high interference values. This interference-based modular decomposition of a network allows grouping of nodes according to their response to the deletion (or addition) of specific nodes in the

network. Importantly, this approach may lead to a less purely mathematical, but more contextual-oriented method of network modularization. Notably, it is well known that scale free networks are not easily affected by randomly removing single nodes [10][11][12]. So a possible scenario of application of interference analysis implies removal of groups of nodes. Definition of interference can be easily adapted to such a situation, where removing a subset of nodes is considered.

In conclusion, the introduction of centrality interference and robustness allows understanding how a network locally rearranges itself when nodes are removed or added from/to a network, a common situation in several applications of networks analysis. An interference analysis allows also to identify which parts of the networks are influenced by single nodes or by modification on the functionality of such nodes; with robustness and related notions (dependence and competition values) we may infer how much the central role of a node can be affected by other nodes in the network.

References

1. Caldarelli, G.: Scale-Free Networks: Complex Webs in Nature and Technology (Oxford Finance). Oxford University Press, USA (June 2007)
2. Bhalla, U.S., Iyengar, R.: Emergent properties of networks of biological signaling pathways. Science 283 (January 1999)
3. Strogatz, S.H.: Exploring complex networks. Nature 410(6825), 268–276 (2001)
4. Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. Science 286(5439), 509–512 (1999)
5. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabási, A.L.: The large-scale organization of metabolic networks. Nature 407(6804), 651–654 (2000)
6. Newman, M.E.J.: Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103(23), 8577–8582 (2006)
7. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: Simple building blocks of complex networks. Science 298(5594), 824–827 (2002)
8. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature 393(6684), 440–442 (1998)
9. Koschützki, D., Lehmann, K.A., Peeters, L., Richter, S., Podehl, D.T., Zlotowski, O.: Centrality indices. In: Brandes, U., Erlebach, T. (eds.) Network Analysis: Methodological Foundations, pp. 16–61. Springer (2005)
10. Barabási, A.-L., Oltvai, Z.N.: Network biology: understanding the cell's functional organization. Nature Reviews Genetics 5(2), 101–113 (2004)
11. Jeong, H., Mason, S.P., Barabási, A.L., Oltvai, Z.N.: Lethality and centrality in protein networks. Nature 411(6833), 41–42 (2001)
12. Albert, R., Jeong, H., Barabási, A.-L.: Error and attack tolerance of complex networks. Nature 406(6794), 378–382 (2000)
13. McCulloh, I., Carley, K.: Detecting change in longitudinal social networks. Journal of Social Structure 12 (2011)
14. Crucitti, P., Latora, V., Marchiori, M., Rapisarda, A.: Error and attack tolerance of complex networks. Physica A: Statistical Mechanics and its Applications 340(1-3), 388 (2004); News and Expectations in Thermostatistics
15. Liu, Y.-Y., Slotine, J.-J., Barabási, A.-L.: Controllability of complex networks. Nature 473(7346), 167–173 (2011)

16. Freeman, L.C.: Centrality in social networks: conceptual clarification. Social Networks 1, 215–239 (1978)
17. Scardoni, G., Petterlini, M., Laudanna, C.: Analyzing biological network parameters with CentiScaPe. Bioinformatics 25(21), 2857–2859 (2009)
18. Goguen, J.A., Meseguer, J.: Security policies and security models. In: 1982 Symposium on Security and Privacy, pp. 11–20. IEEE Computer Society Press (1982)
19. The official ”autostrade per l’Italia” website (2011), http://www.autostrade.it/

Software Collaboration Networks

Christopher Zachor and Mehmet Hadi Gunes

Abstract. The need to work together with others on large projects has been emphasized with industrialization. As the software industry grew, it should be no surprise that communities to serve this purpose appeared. With the creation of websites such as `Sourceforge.net`, `Github.com`, and `Freshmeat.net` developers from around the world are able to collaborate on open source projects. This paper will attempt to extend previous studies of software collaboration networks through the use of network analysis. Examining Sourceforge.net, we analyze this community of developers who have contributed greatly to open source software despite not being paid to do so.

1 Introduction

Open source collaboration systems provide an environment for software collaboration. Such systems are popular. For instance, the Sourceforge.net community has over 250,000 open source projects at the time of this writing. But what are the characteristics of this large and diverse community? When given to tools to collaborate on software, do they work together or separately?

By analyzing this community using social network analysis, we can better understand the developers of open source software. A major challenge for this project is determining what measures will provide interesting and relevant results. We will attempt to focus on measures and metrics that will provide a better understanding of the collaborations (or lack thereof) within the `SourceForge` community.

Previous studies were focused more on the growth of the open source movement rather than the collaborations of the developers [1-3]. While these studies were useful in the sense that they achieved the goal of understanding the open source software movement, they have not explored the developers from social network analysis perspective or their analysis was not tailored to the collaboration aspects of the network.

Christopher Zachor · Mehmet Hadi Gunes
University of Nevada, Reno
e-mail: `zachorc@gmail.com, mgunes@cse.unr.edu`

R. Menezes, A. Evsukoff, & M.C. González (Eds.): Complex Networks, SCI 424, pp. 257–264.
springerlink.com

In this paper, we generate various networks from the SourceForge community and compute network measures. We want to analyze the groups in the SourceForge community to determine if it is a diverse collaboration network. Similarly, we would like to determine if there are closely connected communities (i.e., cliquish sub-groups) who work exclusively with each other, or do developers work with people they don't normally collaborate with.

While previous studies focused on the growth of the SourceForge community and how it relates to the open source movement, we analyze how these communities function and how they work together to produce software. Interpreting the measures extracted from a recent snapshot of the network can provide a wealth of information about how unpaid (with exception to some donations received from the community) developer's work together to produce software.

2 Related Work

Gao et.al. looked at the growth of the SourceForge community over the course of approximately two years [1]. During this time frame, the number of projects grew from about 70,000 to about 90,000. A network was constructed from the relation between the projects and the developers. Three analyses were done on the network, i.e., structure analysis, centrality analysis, and path analysis. They then analyzed the project network, the developer network, and the collaboration network. Moreover, Gao et al. developed an agent based modeling to examine the evolution of the SourceForge communities [2]. While they presented that the SourceForge community was growing, they were not focused on the diversity of collaborations within the network. This will be the primary focus of our paper.

Xu, et.al created multiple networks of SourceForge sub-communities to understand the network and how links are formed within the community structure [3]. After measuring the degree distribution and showing that it follows the power law, they indicated the networks are scale free. They found the small world phenomenon, not only within the project leader network and core developer network, but they also found it in the co-developer and active user networks. However, similar to the previous study, they did not focus on the diverse collaborations.

While writing research papers and developing software is not same, the community structures of people that perform these tasks are, at the very least, similar. On one hand, you have researchers collaborating to produce a research paper. On the other hand, you have developers collaborating to produce software. In this sense software collaboration networks are similar to co-authorship networks.

One of the earlier studies on co-authorship networks was performed by Newman in [4]. He used four different databases from four different disciplines. In the study, Newman was able to point out the difference between fields when it comes to collaboration with other authors. Mathematics was low with about 1 to 1.5 authors per publication. Meanwhile, papers on high-energy physics had an average of about 8.9 authors. Author also pointed out some databases followed the power law degree distribution while some did not. Author also indicated potential flaws in the study. For instance, an author who supposedly published over 1600 papers in a five year window due to several researchers with the same name that was not identified during data collection and processing phases.

3 Methodology

A simple Perl script will download project lists from `SourceForge.net` and match HTML tags using regular expressions. By acquiring the project titles, a script can be written to visit each page through its formatted URL (e.g. `http://sorceforge.net/projects/projectTitle`). The project home page can then be parsed, using regular expressions again, to acquire every developer involved. When all projects and their corresponding developers have been collected, the data will be checked for uniqueness.

The first network formed from this data will be a developer collaboration network. Each node will consist of a developer and every edge will imply project collaboration between the two. Developers working together multiple times will not be taken in to account because we are only interested in the diversity of collaborations, a single edge will be sufficient.

The second topology we will examine is the project-developer network. Nodes are created using both projects and developers. A link between nodes will imply a developer has worked on that project. As nodes will be from two distinct groups, the network will be bipartite.

The first measures we will look at for each graph is *degree* related. The degree of a node is the number of edges connected to it. We will also examine the *degree distribution*. By looking at the degree distribution, we can determine if the network follows a power law. Xu, et al. indicated that the `SourceForge` community does, in fact, follow the power law [3]. However, the community has grown by more than double since their study. While this growth should not change the fact that the degree distribution follows a power law, it is still important to verify it.

The *assortativity* coefficient, indicated by equation 1, is also an important measure [5]. It can give us a measure of the likelihood that nodes of a similar degree will work together. Thus, we will have higher degree nodes working with higher degree nodes and lower degree nodes will work with lower degree nodes. In the case of the `SourceForge` community, we can look at the higher degree nodes as the more social developers in the network. Likewise, we can view the lower degree nodes as the developers who would prefer to work alone. A result closer to 1 will indicate nodes of a similar degree will likely have links to other nodes of a similar degree. On the other hand, a result closer to -1 would indicate that higher degree nodes are working with multiple nodes of a lesser degree.

$$r = \frac{\sum \left(A_{ij} - k_i k_j / 2\mathrm{m}\right) k_i k_j}{\sum \left(k_i \partial_{ij} - k_i k_j / 2\mathrm{m}\right) k_i k_j} \tag{1}$$

Another measure we will examine is the *clustering coefficient* of the developer network. This measure can be useful because social networks tend to have a higher clustering coefficient than technological or biological networks [5]. We can look at the network clustering coefficient with the help of the equation 2.

$$C = \frac{(number\ of\ triangles)x\ 3}{(number\ of\ connected\ triples)} \tag{2}$$

The *small world* phenomenon is important when discussing collaboration networks. Each time you work with a new individual, either you learn something from that individual or they learn something from you. In a small world network, new ideas or methods can spread very quickly. Xu et al. were able to observe the small world phenomenon in the SourceForge community [3]. However, the network has grown drastically over the past years making it necessary to revisit. What we are looking for is a relatively high clustering coefficient and a low average shortest path measure. We say relatively because we will construct random networks of equal nodes and edges. By comparing the differences, we can verify small world properties within the community. The network should have a low average shortest path, but the clustering coefficient of the SourceForge community should be significantly higher than the random network.

4 Results

The data was collected over the course of one week in the middle of November 2010. During this time all projects listed on SourceForge were gathered with the current number of downloads. Then, each project page was visited to gather all developers involved in the creation of the application. A total of approximately **250,000** project titles were collected. However, roughly 20,000 of these project pages were inaccessible. This was due to the page not existing. Many attempts were made to recover the information from these projects but the pages simply do not exist. This leads us to believe the projects have been taken down by the authors and SourceForge leaves a record of their existence in the project database. Of the **230,000** remaining projects, only **115,000** had at least one download or more. The rest of the projects either had no downloads or were marked as inactive by the developer. Projects marked as inactive do not display downloads. While SourceForge claims to have **2.7** million developers, we found only **~200,000** unique developers who have worked on at least one project. The larger number given by SourceForge includes community members whether they have worked on a project or not.

4.1 Degree Distribution

For the Project-Developer Network, there were a total of **430,281** nodes with **276,853** edges. The total number of developers was **204,077**. The total number of projects was **226,204**. The average number of projects worked on by the developers in this network was **~1.35** and the average number of developers per project came out to **~1.22**. These numbers are not surprising because there are a large number of developers who have worked solo on a single project. We will examine this further when we discuss the power law properties of the developer network.

The project that goes by the title "jungerl" was found to have the most developers on a single project with a total of **43** developers. However, on further inspection of the project, they appear to have produced very little. Meanwhile, the developer that has produced the most software goes by the user name "roro01" with a total of 50 open source applications. Most of them appear to be PDF conversion tools. Finally, the developer who have collaborated the most with others is "bfulgham". This developer has worked with a total of **73** other developers on the network.

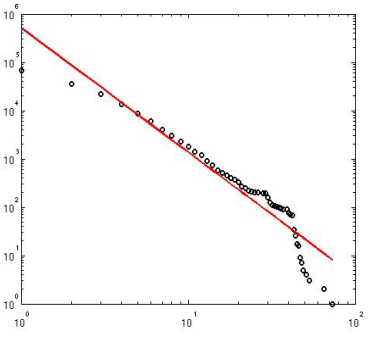


Fig. 1. Degree distribution

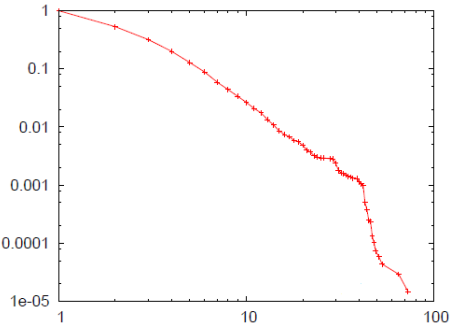


Fig. 2. CCDF of degree distribution

In the Developer Network, we found an average degree of approximately **0.86**. This means that each developer almost worked with an average of at least one person. Despite having nearly **130,000** projects with a single developer, the community as a whole is working together. In the Project Network, the average degree is over one at **~1.27**. This means one of two things. Either solo developers are working on multiple projects, or many developers are working on a diverse range of projects.

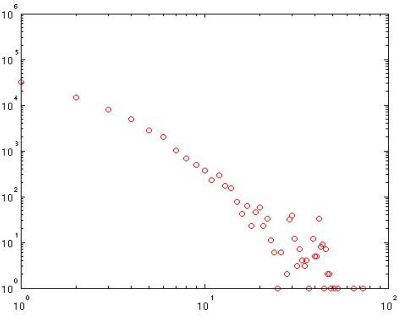


Fig. 3. Non-binned degree distribution with "fat tail"

One interesting fact about the `SourceForge` Developer Network is that it follows the power law. Despite the growth since the last study done on the network, it still follows the power law properties. This means that we have a few developers who have collaborated with many others (in this case, the highest is **73**) and we have many developers who have collaborated with only one or zero other developers. In figure 1, we have the degree distribution where the data has been charted on a log log plot. Figure 3 presents the complementary cumulative distribution function of the degree distribution. Figure 3 shows the data that has not been binned, but we can see the fat tail that is distinctive of a scale free network.

We should expect that only a few developers will work with so many people and we should expect that there will be so many in the community who would rather work alone. Perhaps this tells us something about programmers in general. Some of us would just rather work on a project alone. We might ask others for help occasionally, but the result will be our own.

4.2 Assortativity and Rich Club

Moreover, the network has an assortativity coefficient of 0.85 indicating that high degree nodes are connected to other high degree nodes more than the low degree nodes. This is also reflected in almost exponentially increasing average neighbor degree distribution in Figure 4. This indicates we have a number of developers who are more likely to work with other social developers within the community. In other words, the "social butterflies" of the network do not collaborate with the "hermits" in the community. This makes sense because social developers might actively seek group projects and find other social developers seeking for the same. The developer is also introduced to new developers through these projects. This gives them a greater opportunity to work with a more diverse collection of developers as opposed to a solo developer. Thus, their clique keeps growing while the solo developer collaboration choices remain the same. Additionally, the rich club connectivity graph in Figure 5 indicates there is a rich club in the network. That is, considering the connections a small fraction of users are at the core of the network with majority of links among themselves.

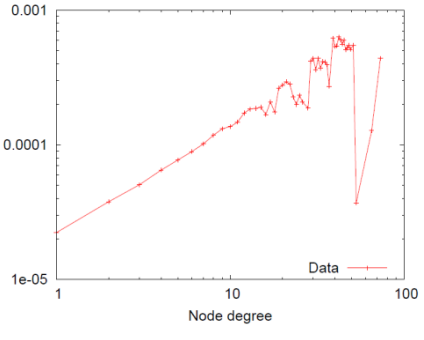


Fig. 4. Neighbor Degree Distr

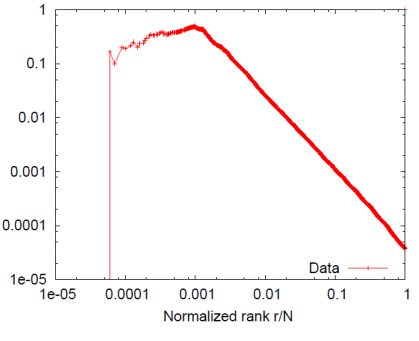


Fig. 5. Rich Club Connectivity

4.3 Clustering Coefficient

The clustering coefficient for the developer network will give us a measure of the transitivity between developers. For example, if we look at three specific developers (A, B, and C), we have a perfect transitivity of one if and only if developer A

has worked with developer B, developer B has worked with developer C, and developer A has also worked with developer C. This forms a closed triad among the three developers. By taking the average of the entire network, we can get an idea of how cliquish the developer community is.

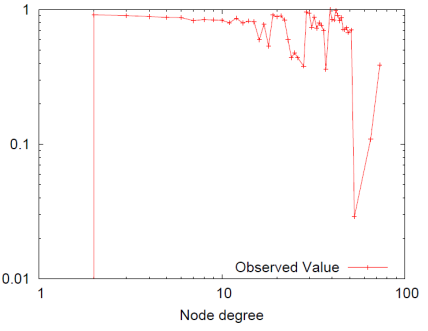


Fig. 6. Clustering Coefficient Distribution

The network clustering coefficient for the developer network is **0.84396**. Figure 6 presents the clustering coefficient distribution with respect to node degrees. While this is a high number, it makes sense because most two path pairs are generally closed because the three developers were working on the same project together. Thus, all three developers have worked together to close the triad. Despite being high, the clustering coefficient is only slightly higher than previous studies [2]. So, even though the network has grown by more than double in the past three years, the transitivity has remained roughly the same. We should expect that as the network grows, developers will branch out and form more cliques to fill the gap between possible triads and closed triads.

4.4 Small World Properties

A random network of the same number of nodes and the same average degree was constructed using the Erdos-Renyi method. The clustering coefficient and average shortest paths were then measured. The result was an average clustering coefficient of **0.0000114** was found. This is significantly lower than the developer network's clustering coefficient of **~0.84**. The average shortest path of the random network was approximately 6.6 among reachable pairs. Meanwhile, the developer network had an average shortest path of nearly **13.5** among reachable pairs.

For a network to exhibit small world properties, it must have a significantly higher average clustering coefficient and a similar average shortest path of a random network with an equal number of nodes and average degree. Because the average shortest distance is significantly higher in the developer network, we must conclude that the developer network is not a small world network. The reason for the higher average distance could be due to the existence of cliquish communities within the network. These groups of developers generally work together, but one or two developers within the group may branch out and work with other groups. This would mean longer paths to reach other developers in other groups. While previous studies found the `SourceForge` community to exhibit small world properties [3], they included non-developers in their network.

5 Conclusion

We have looked at many different aspects of the SourceForge community with a focus on the developer's preferences for working in teams. Despite a large growth in the community, the properties of this social network have remained the same. Even with the addition of many new developers the clustering coefficient has remained high. While most of these programmers are not the most team oriented people on the planet, they are not hermits sitting in a dark room writing code by themselves. However, we did find that the developers have formed groups in a sense. Our assortativity measure would indicate that we have groups of developers who prefer to work with other developers who like to work in teams. They are also not branching out to work with solo developers in the community. Thus, we can conclude there is a divide within the network. This divide is between solo developers, who would prefer to write code alone, and team-oriented developers, who prefer to work with others. Of course, this raises the question of whether the solo developers actually prefer to work alone or perhaps they are not the type of person to go out and find teammates to work with.

Future work for this paper will include the collection of data from Github.com, a similar software collaboration network. This will include a comparison and analysis of the two communities to determine if Github really is a social coding network. It will also include a measure of modularity based on various attributes to better understand what type of developer the communities prefer to work with.

References

[1] Gao, Y., Madey, G.: Network Analysis of the SourceForge.net Community. Spring Sim. 1, 187–200 (2007)
[2] Gao, Y., Madey, G.: Towards understanding: A study of the Source-Forge.net community using modeling and simulation. Spring Sim. 1, 145–150
[3] Xu, J., Christley, S., Madey, G.: Network Analysis to the Study of Open Source Software Application of Social. In: The Economics of Open Source Software Development, vol. 205, Elsevier B.V. (2006)
[4] Newman, M.E.J.: Scientific collaboration networks. Network construction and fundamental results. Physical Review E 64 (2001)
[5] Newman, M.E.J.: Networks: An Introduction. Oxford University Press, New York (2010)

Author Index