Social Network Probability Mechanics

IAN A. MCCULLOH
JOSHUA LOSPINOSO
Department of Mathematical Sciences
United States Military Academy
West Point, NY 10996
UNITED STATES OF AMERICA

KATHLEEN CARLEY
Institute for Software Research
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
UNITED STATES OF AMERICA

Abstract: - A new model for a random graph is proposed that can be constructed from empirical data and has some desirable properties compared to scale-free graphs [1, 2, 3] for certain applications. The newly proposed random graph maintains the same "small-world" properties [3, 4, 5] of the scale-free graph, while allowing mathematical modeling of the relationships that make up the random graph. E-mail communication data was collected on a group of 24 mid-career Army officers in a one-year graduate program [6] to validate necessary assumptions for this new class of random graphs. Statistical distributions on graph level measures are then approximated using Monte Carlo simulation and used to detect change in a graph over time.

Key-Words: - Social networks, random graph, change detection, small-world

1 Introduction

Social network analysis (SNA) examines relationships between social entities (i.e. people, groups, tasks, beliefs, knowledge, etc.). These entities are modeled with nodes or vertices and their connections or relationships are modeled with edges. Not all nodes are connected and some nodes may have multiple connections. This mathematical model is applicable in many content areas such as communications, information flow, and group or organizational affiliation [7, 8]. SNA thus relies heavily on graph theory to make predictions about network structure.

Nodes are defined in terms of a set of n verticies, $V = v_1, v_2, \ldots, v_n$. The nodes are related to each other with a set of edges E, where e_{ij} is a relationship between node v_i and v_j. A social network is often shown as an adjacency matrix, where the rows and columns correspond to the nodes and each cell a_{ij} can take on any numerical value corresponding to the edge e_{ij}. In an unweighted network, cells are Boolean and are represented as 0/1: the presence or absence of an edge or relationship between nodes i and j. Networks where relationships between nodes are always mutual are called undirected networks, and their adjacency matrices will always be symmetric. Directed networks, on the other hand, can model both mutual and directional relationships. A value of 1 in cell a_{ij} represents a directed relation from node i to node j. In application, the diagonal of the adjacency matrix is rarely populated with anything but zeros, since interactions from an entity to itself are not generally interesting.

The potential complexity of interactions within even a small network, while discrete, grows exponentially with the number of entities. For this reason, algorithmic approaches to exploring distributions within constrained networks quickly become computationally challenging. In a directed network, the number of possible relationships among nodes can be found by the following expression, where n represents the number of nodes in the network:

$$n^2 - n$$

The number of possible configurations of a network with a specified number of nodes (n) and edges (ε)

can be thought of as the number of unique combinations of ε nodes within the network:

$$C_{\varepsilon}^{n^2-n} = \frac{(n^2-n)!}{\varepsilon!\cdot(n^2-n-\varepsilon)!}$$

It follows that the total number of possible network configurations with n nodes can be represented by the following:

$$\sum_{\varepsilon=1}^{n^2-n} \frac{(n^2-n)!}{\varepsilon!\cdot(n^2-n-\varepsilon)!}$$

A network of 30 nodes, for example, can be uniquely configured roughly 7.87 x 10^{261} different ways.

To understand the probability of network structures occurring, the degree of the nodes is often investigated [3, 6, 8, 9, 10]. The degree of a node, k_i, is a simple network measure counting the number of edges going into/coming out of a particular node. It is often powerful and accurate at determining who holds the power and influence within a network [8, 9, 11, 12]. If we accept the notion that a random network is one in which nodes have an equal and unchanging probability to have a relationship with all other nodes in the network, random networks have a nice underlying distribution of degree measures. Both the degree of a node and the number of edges in a network both will follow a binomial distribution. As the network gets arbitrarily large, the distribution converges to a Poisson distribution.

There is no shortage of alternative views on what constitutes a random network; nevertheless, empirical work has shown that social networks do not construct themselves in the image of a Binomial random graph [3, 6, 11]. Travers [13] studied social connections in the United States and discovered surprisingly short path lengths, where many strangers were connected by mutual acquaintances. This was termed a small-world network. A network is a small world network if its average path length is much smaller than the number of nodes in the network. This phenomenon in real-world networks is popularly known as "six degrees of separation" [14]. Watts and Strogatz [4] proposed the clustering coefficient as a graph level measure to indicate whether a graph is a small-world network. The clustering coefficient for a directed graph is defined as,

$$C_i = \frac{\left|\{e_{jk}\}\right|}{k_i(k_i-1)} : v_j, v_k \in N_i, e_{jk} \in E,$$

where N_i the neighborhood for a vertex v_i and is defined as it's immediately connected neighbors,

$$N_i = \{v_j\} : e_{ij} \in E.$$

The degree k_i of a vertex is the number of vertices, $|N_i|$ in it's neighborhood $|N_i|$. Albert and Barabasi [3] review current methods of constructing random graphs throughout the field of Network Science and compare the degree distribution, clustering coefficient, and average path length of multiple real-world networks with various types of random networks. They find that real-world networks have a higher average clustering coefficient and a shorter average path length than randomly generated networks with the same number of nodes and edges. Furthermore, they show that several networks have degree distributions that follow a power-law distribution, which means that very few nodes have a large degree, and many nodes have a small degree.

Barabasi [1] proposed the scale-free graph which creates a condition on the random graph that the degree distribution must follow a power law distribution. These networks were shown to resemble real-world networks [3]. While scale-free networks may appear to be similar to real-world networks in terms of structure, they are not a sufficient framework to truly understand the stochastic nature of networks.

A new framework for random networks is proposed, based upon empirical data collected on real-world networks. This new approach produces networks that have equivalent properties to the scale-free networks outlined by Albert and Barabasi [3], however, it is constructed in such a manner as to describe the close relationships between some nodes and distant relationships between others. This framework holds the promise of a new line of research to explore the stochastic behavior of networks.

2 Problem Formulation

Individuals in a social network are not connected to other individuals with uniform random probability. The probability structure is much more complex. Intuitively, there are some people whom a person will communicate with or be connected more closely than others. In a study of email communication conducted at the U.S. Military Academy [6], one subject emailed his wife more than ten times per day on average, while other people that he worked with received an email from him once or twice per month. For this reason, real-world networks tend to have

clusters or cliques of nodes that are more closely related than others [3, 15, 16, 17]. This can be simulated by varying the probabilities that certain nodes will communicate.

Consider a group consisting of 15 individuals, organized into three subgroups. Individuals within each subgroup work closely together and communicate more frequently than they do with people in the larger group. Each day individuals may communicate with others in the group, but probably not everyone. If we let the probability that an individual will communicate with someone in their subgroup with probability 0.8 and communicate with someone outside their subgroup with probability 0.2, we have a network probability matrix (NPM) shown in Figure 1.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
A		0.8	0.8	0.8	0.8	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
B	0.8		0.8	0.8	0.8	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
C	0.8	0.8		0.8	0.8	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
D	0.8	0.8	0.8		0.8	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
E	0.8	0.8	0.8	0.8		0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
F	0.2	0.2	0.2	0.2	0.2		0.8	0.8	0.8	0.8	0.2	0.2	0.2	0.2	0.2
G	0.2	0.2	0.2	0.2	0.2	0.8		0.8	0.8	0.8	0.2	0.2	0.2	0.2	0.2
H	0.2	0.2	0.2	0.2	0.2	0.8	0.8		0.8	0.8	0.2	0.2	0.2	0.2	0.2
I	0.2	0.2	0.2	0.2	0.2	0.8	0.8	0.8		0.8	0.2	0.2	0.2	0.2	0.2
J	0.2	0.2	0.2	0.2	0.2	0.8	0.8	0.8	0.8		0.2	0.2	0.2	0.2	0.2
K	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2		0.8	0.8	0.8	0.8
L	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.8		0.8	0.8	0.8
M	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.8	0.8		0.8	0.8
N	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.8	0.8	0.8		0.8
O	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.8	0.8	0.8	0.8	

Figure 1. Network Probability Matrix.

Using this NPM, Monte Carlo simulation was used to generate 5000 instances of the network. The average clustering coefficient was 0.463 ± 0.0014 compared with a clustering coefficient of 0.329 ± 0.0024 in a random graph of uniform probability. The graph generated with the NPM has a clustering coefficient that is comparable to a small world graph [4] with the same number of nodes and edges. It can be conjectured that the clustering coefficient will become greater as the within group edge probability increases. Furthermore, as the probability of certain key nodes being connected to others increases, the degree distribution will more closely follow a power law distribution. The newly proposed random network, therefore, achieves equivalent performance as the small world graph in modeling real-world networks, and can be extended to model scale-free graphs, yet preserves the flexibility to model dyadic relationships between nodes.

The edge probabilities can be derived from empirical data in several ways. Given network data collected over multiple time periods on a group of subjects, the edge probabilities can be estimated by the proportion of edge occurrences, e_{ij}, for each cell in the adjacency matrix, a_{ij}. In the case of communication networks, statistical distributions can be fit to the time between messages for each potential edge in the network. For a specified period of time, t, the edge probability p for each set of entities i and j can be found. Let x_{ij} be the time between messages in a communication network. The probability density function for any x can then be defined as $f_{ij}(x \mid \theta_{ij})$, where θ_{ij} is the set of parameters for the distribution. Then, the probability, p, of an edge occurring within some time period t is the probability that $x < t$, which can be expressed as,

$$p = \int_0^t f_{ij}(x \mid \theta_{ij})\, dx$$

In practice, the function $f_{ij}(x \mid \theta_{ij})$ must be estimated using techniques such as maximum likelihood estimation from empirical data collected on the group being studied. It may be desirable to construct a network based on a restriction such as, "two emails within a time period demonstrate a relationship, but one does not." In this case, it is necessary to compose a function of random variables. If $h_{ij}(2 \mid t, \theta_{ij})$ represents the probability density function of time between two sets of two emails and $f_{ij}(x \mid \theta_{ij})$ represents the probability density function of time between one set of two emails, then

the following is true under certain assumptions:

$$h_{ij}(2 \mid \theta_{ij}) = \left(\int_0^t f_{ij}(x \mid \theta_{ij}) dx \right)^2$$

It is possible to generalize this idea; if $h_{ij}(x \mid \theta_{ij}, t)$ is the probability that x or more communications occur within time t, then the following is true:

$$h_{ij}(x \mid \theta_{ij}, t) = \left(\int_0^t f_{ij}(y \mid \theta_{ij}) dy \right)^x$$

This newly proposed framework for viewing the probability space of a social network preserves the same flexibility for modeling dyadic relationships, however, it provides researchers with a means to understand the probability space of the network and thus devise more robust and appropriate statistical tests for social network analysis.

3 Example Problem Solution

Researchers at the U.S. Military Academy monitored the e-mail traffic of 24 mid-grade Army officers for 24 weeks as they were in a one year graduate program at Columbia University [6]. The group had been organized with a formal leadership structure among the 24 officers, they all lived on the West Point Military Installation, and they had regular social events for the officers and their families. The degree distribution followed a power law distribution like the social networks analyzed by Barabasi and Albert [1, 2, 3], and Newman [11, 15]. The time between emails for each possible pair of nodes was calculated. There were only 65 directed pairs of nodes that had greater than 30 messages over the course of 24 weeks. Statistical distributions were fit to the time between email for the 65 pairs of nodes. All of them followed a lognormal distribution. Figure 2 shows the empirical distribution of one directed pair and four distributions fit to the data: exponential, lognormal, pareto, and zipf.

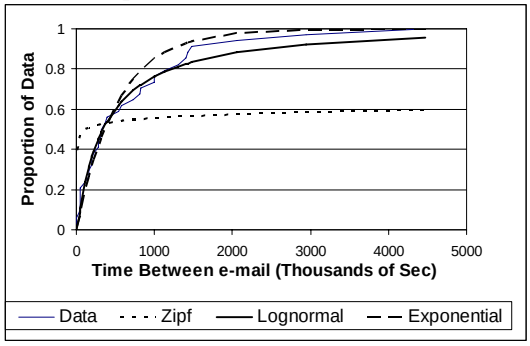

Figure 2. Distributions Fit to Time Between E-mails in Army Officer Study.

One could conjecture that the parameters of the lognormal distributions may be dependent upon various social factors, such as formal position in the network, friendship, common interest, etc. Unlike traditional social network analysis, using the NPM, an analyst can use the edge probabilities as dependent variables to study the causes of relationships, communication frequency, and ultimately network structure.

4 Conclusion

A new approach to modeling a random network has been proposed that resembles real-world networks, preserves dyadic relationships, and can be estimated from empirical data. While the approach is surprisingly simple, it opens the door for many new analysis opportunities in social network analysis. The cell entries in the NPM can be treated as dependent variables, while various properties describing the dyadic relationships between nodal pairs can be used as independent variables. This will reduce variance in the model and increase the coefficient of determination, thereby explaining the complex behavior of a social network much better than existing methods.

Other research building from this new approach to modeling a random network can include building empirical distributions of social network measures. This newly proposed framework allows analysts to randomly generate instances of social networks under investigation. Parameters of distributions for social network measures can then be estimated using Monte Carlo simulation.

Consideration of the probability space of entity level communications is imperative for many studies of social networks. Many considerations for designing social experiments rely on conventions within the field. When constructing interaction matrices, for example, experimenters must choose many parameters which may change the conclusion of the study. The experimenters of the U.S. Military Academy e-mail study, for example, had to choose how many emails between two entities demonstrate a relationship to create an unweighted, directional network. To study the dynamics of the network, the experimenters further needed to determine regular intervals to sample, which allowed for a temporal analysis. By instead fitting distributions to the empirical data, experimenters could use statistical techniques to manipulate random variables and sidestep the selection of the potentially influential aforementioned parameters.

Acknowlegements:
This research is part of the Dynamics Networks project in CASOS (Center for Computational Analysis of Social and Organizational Systems, http://www.casos.cs.cmu.edu) at Carnegie Mellon University. This work was supported in part by:
• The Office of Naval Research (ONR), United States Navy Grant No. 9620.1.1140071 on Dynamic Network Analysis (DNA),
•The Army Research Labs for Assessing C2 structures, Collaborative Technology Alliance,
•Alion Science and Technology,
•ARL Telecordia: Communications and Networks Technology Collaborative Alliances,
•The Defense Advanced Research Projects Agency (DARPA),
•The Air Force Office of Sponsored Research (MURI: Cultural Modeling of the Adversary Organization, 600322) ONR N00014-06-1-0104, N00014-06-1-0921, and
•Additional support on measures was provided by the DOD and the NSF IGERT 9972762 in CASOS.
• 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 National Science Foundation or the U.S. government.

References:

[1] Barabasi, A, Scale-Free Networks, *Scientific American,* 2003, 288:60-69.

[2] Barabasi, A, *Linked: How Everything is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*, New York: Plume, 2003.

[3] Albert, R. and Barabasi, A., Statistical Mechanics of Complex Networks, *Reviews of Modern Physics,* 2002, 74: 47-97.

[4] Watts, D.J. and Strogatz, S.H., Collective dynamics of 'small-world' networks, *Nature,* 1998, 393(6684): 440-2.

[5] Milgram, S., The small world problem, *Psychology Today,* 1967, 2:60–67.

[6] McCulloh, I., Garcia, G., Tardieu, K., MacGibon, J., Dye, H., Moores, K., Graham, J. M., & Horn, D. B., *IkeNet: Social network analysis of e-mail traffic in the Eisenhower Leadership Development Program. (Technical Report, No. 1205)*, Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, , 2007

[7] Tichy, N.M., Tushman, M.L., Fombrun, C., Social Network Analysis for Organizations, *The Academy of Management Review,* 1979, 4(4): 507-519.

[8] Wasserman, S., and Faust, K., *Social Network Analysis: Methods and Applications*. New York: Cambridge University Press, 1994.

[9] Scott, J., *Social Network Analysis: A Handbook,* 2nd Ed., Newberry Park, CA: Sage, 2000.

[10] Borgotti, S.P., Carley, K.M., and Krackhardt, D., On the Robustness of Centrality Measures Under Conditions of Imperfect Data, *Social Networks,* 2006, 28:124-136.

[11] Newman, M., The Mathematics of Complex Networks, *Unpublished,* 2007.

[12] Casciaro, T., Carley, K.M., and Krackhardt, D. Positive Affectivity and Accuracy in Social Network Perception, *Motivation and Emotion*, 1999, 23:285-306

[13] Travers, J. and Stanley M., An Experimental Study of the Small World Problem, *Sociometry*, 1969, 32(4): 425-443.

[14] Guare, J. *Six Degrees of Separation*: A Play New York: Vintage Books, 1990.

[15] Newman, M., The Structure and Function of Complex Networks, *SIAM Review,*2003, 45(2): 167-256.

[16] Carley, K.M., A Comparison of Artificial and Human Organizations, *Journal of Economic Behavior and Organization*, 1996, 31:175-191.

[17] Topper, C. and Carley, K.M., A Structural Perspective on the Emergence of network Organizations, *Journal of Mathematical Sociology*, 1999, 24(1):67-96.

[18] Erdős, P., and Rényi, A., On the Evolution of Random Graphs, *Mathematical Institute of the Hungarian Academy of Science*, 1960, 5:17-61.

[19] Faloutsos, M., Faloutsos, P. and Faloutsos, C., On power-law relationships of the internet topology, *Computer Communication Review,* 1999, 29:251.

[20] Reminga, J. and Carley, K.M. *ORA: Organizational Risk Analyzer v.1.7.8.* Pittsburgh, PA: Institute for Software Research, 2007.

[21] Weiss, L., Testing One Simple Hypothesis Against Another, *Annals of Mathematical Statistics,* 1953, 24:273-281.