An Empirically-Based Model for Network Estimation Under Uncertainty and Policy Analysis

Matthew J. Dombroski, Paul Fischbeck, and Kathleen Carley

Abstract:

Social network analysis has been used to understand groups of individuals and how they operate. Most of the literature in social networks has dealt with overt organizations with an easily discernable network structure. This paper examines the possibilities of using the inherent structures observed in social networks to make predictions of networks using limited and missing information. The model is based on empirical network data exhibiting the structural properties of triad closure and adjacency. Triad closure indicates that if person *i* has a dyad with person *j* and person *j* has a dyad with person *k*, then there is a higher than chance likelihood that person *i* and person *j* have a dyad. The model exploits these properties using an inference model to update adjacent dyads given information on a reference dyad. The model is tested against several networks to understand and discern its behavior. The paper illustrates that if the model is built with careful consideration towards the network being predicted, it will assist in making better decisions regarding uncertain organizational phenomenon. The method is applied in a covert network example, and has been extended to show its usefulness in epidemiological networks and improving performance in organizations operating under stress. The paper opens up new avenues in the development of models designed to make network predictions and use those predictions to make better decisions.

Support: This research has been supported in part by the National Science Foundation IGERT in CASOS, the office of Naval Research ONR 1681-1-1001944 and the center for Computational Analysis of Social and Organizational Systems. The views and results expressed herein are solely the responsibility of the authors and do not represent the official views of the Office of Naval Research or the National Science Foundation.

Unknown Network Structures

Social network analysis has been used to understand organizational dynamics in a variety of application areas (e.g., epidemiology, technological diffusion, and management consulting). A group's behavior, values, and/or performance can be articulated by understanding the relationships that exist within the group. Most applications to date have been on open groups or societies in controlled experiments. Currently there have been very few network applications to covert or "hidden" networks of interest. Social network measures and tools that could efficiently infer "hidden" networks from limited data could allow policy makers to make better informed decisions in a variety of applications.

This paper presents an empirically-based probabilistic model, grounded on observational social networks, to infer network structure using limited and incomplete information. First, relative similarity information is used to build a prior probability assessment of who communicates with whom. As direct information on dyadic likelihood is received, these priors are updated. Adjacent dyads are updated through an empirically based inference model. This continues until the likelihood of every dyad in the probabilistic network is inferred and updated.

The resulting network provides an estimate of the actual network and may be used to guide policy analysis.

Network Properties

Researchers have uncovered inherent structural properties in social networks (Skvoretz, 1990). These properties arise from the structure of the network itself and not from the behavior of the individuals in the network. They include reciprocity, triad-closure, and triad-closure reciprocity. A corollary of the triad properties is an adjacency property. Simply stated, if persons *i* and *j* are talkative with each other, then they are likely to be talkative with others. Formally, if *A* and *B* are adjacent dyads, then

if $n_B > \bar{n} \Rightarrow E(n_A) > \bar{n}$, and

if $n_B < \bar{n} \Rightarrow E(n_A) < \bar{n}$,

where n_B is the number of interactions recorded on dyad *B*, $E(n_A)$ is the expected number of interactions on dyad *A*, and *n*-bar is the mean number of interactions for the whole network. In other words, if *B* has above average activity then the expected value of the distribution of interactions for all of its adjacent dyads will also exceed the mean number of interactions. The degree to which these properties exist varies from network to network (Krackhardt, 1987).

Constructing the Model

The problem domain will determine the relationship of interest (ROI). In most real-world situations, only samples of interactions between individuals can be observed. Depending on the type of interaction, knowing that *i* and *j* interacted will inform our belief about the likelihood of a ROI existing between the individuals. But, what, if any, inference can be made about these individuals' relationships with others in the network?

For illustrative purposes and to facilitate model development, we focus on one social network dataset, Bernard and Killworth's 1979 observed interactions between 58 fraternity brothers at a West Virginia university. Because of the size of this data set, it was not possible to develop a robust inference model based on the triad-closure property. Instead, the model is based on adjacency properties found in the data. Figure 1 shows this relationship between interactions on a reference dyad and the expected number of interactions on an adjacent dyad. As the number of communications for the reference dyad increases, so does the expected number for the adjacent dyads.

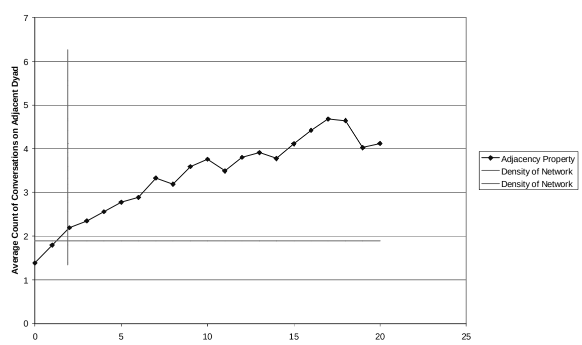

Figure 1—Adjacency Property Illustrated using Bernard and Killworth Fraternity Data

Transforming Data

In order to build a model of dyadic dependency, the network data of "counts of interactions" has to be converted into probabilities that an ROI exists between any pair of individuals. To do so, requires a careful definition of what constitutes an ROI. Two important considerations must be made.

1. The number of interactions needed to define when a ROI exists
2. The marginal increase of each additional interaction towards the probability of a ROI existing

We also need to establish the functional form that relates additional interactions to the probability of ROI. In this paper, we are assuming a concave function (marginal decreasing value). A standard exponential functional form is used:

$$d_{ij} = \frac{(1-e^{-\lambda x_{ij}})}{(1-e^{-\lambda(\max(x_{ij}))})}$$

where d_{ij} is the probability of the relationship of interest, λ is the shape parameter of the function (higher values more concave), x_{ij} is the interaction dyadic data, and $\max(x_{ij})$ is the ROI threshold value. For this model, a ROI threshold of 21 interactions and a λ value of 0.14 were used. The 21 interaction threshold value was chosen so that strong relationships could be modeled and that a large distribution of reference probabilities could be considered. The 0.14 λ value was chosen as a moderate value to attain some concavity to the curve.

Building a Model of Dyadic Dependency

The dependency relationship between dyads can be illustrated by plotting adjacent dyads' probabilities against reference dyads probabilities, for all dyadic pairs. Figure 2 shows the percentile contours for the transformed fraternity data. Note that the percentile contours are generally increasing with probability.

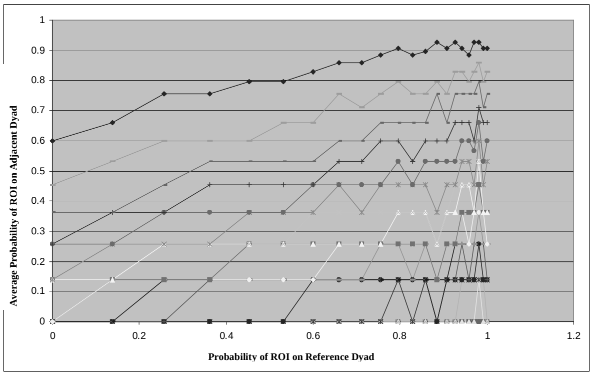

Figure 2---Raw Fraternity Data Showing the Dependency Relationship Between Dyads

By fitting lines through these data and smoothing the parameters, the dependency between dyadic probabilities in the fraternity data can clearly be shown. For this application, the relationships shown in the plots were modeled using a neural network. (Freeman and Skapura, 1991) The neural network consists of 9 nodes and 3 layers.

Implementation

Priors to dyads may be assigned using homophily (McPherson and Smith-Lovin, 1987, McPherson, Popielarz, and Drobnic, 1992, Valente et al., 1997) and a database of social relationships and attributes such as the PCANS methodology (Krackhardt and Carley, 1998). As an observation comes in to inform the model, Bayes Rule performs the direct update and the inference model can then propagate the information to update other relationships in the network. Suppose that I_{ij} is the event that an interaction is observed between nodes *i* and *j*. Suppose also that L_{ij} is the event that the ROI exists between nodes *i* and *j*. $P(I_{ij}|L_{ij})$ and $P(I_{ij}|L_{ij}^C)$ can be assessed for each piece of incoming information. Then the conditionals can be used to update the probability of the reference dyad, $P(L_{ij})$.

$$P(L_{ij} \mid I_{ij}) = \frac{P(I_{ij} L_{ij})}{P(I_{ij})} = \frac{P(I_{ij} \mid L_{ij}) P(L_{ij})}{P(I_{ij} \mid L_{ij}) P(L_{ij}) + P(I_{ij} \mid L_{ij}^{\ C}) P(L_{ij}^{\ C})}$$

Once the probability of the initial dyad is calculated using Bayes Rule, there are several choices for how to propagate the update through the rest of the network. In this paper three alternative models are considered:

1. Bayes Rule is used to update only the reference dyad.
2. Bayes Rule is used to update the reference dyad. A secondary round of updates are applied to the dyads immediately adjacent to the reference dyad, using the inference model.
3. Bayes Rule is used to update the reference dyad. A secondary and tertiary round of updates are applied to the dyads immediately adjacent to the reference dyad and adjacent dyads using the inference model.

Decision Analysis-Covert Networks

Consider a subpopulation of conspirators is being examined to determine its network structure and who to target in that population with an unknown network structure. Borgatti (2002) identifies a key player metric for network analysis that we use to develop inoculation strategies. Suppose that resources are available to remove 7 out of the 20 individuals. We can examine the efficacy of the recommended strategy by comparing the recommended strategy (developed from a network prediction after 100 updates) to that of the real network. Minimums, averages, and maximums for the 20 simulation runs under models 1, 2, and 3 are plotted in Figure 3.

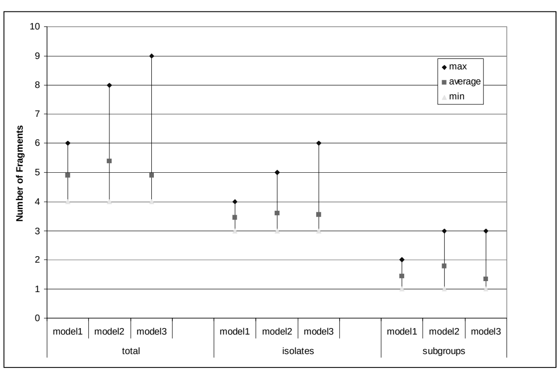

Figure 3---A comparison of models for covert networks; for all possible isolations, model 2 has a higher average value for the number of fragments created

Table 1-- Evaluation of models for covert network scenario showing the expected values of fragments, given different models, and EVPI

	Fragments		
	total	isolates	subgroups
perfect information	12	11	1
E(Model1)	4.9	3.45	1.45
E(Model2)	5.4	3.6	1.8
E(Model3)	4.9	3.55	1.35

Figure 5 and Table 1 indicate that on average, model 2 is best at breaking up the network, but model 3 attains a maximum number of fragments that is higher than model 2 for isolates and total number of fragments. Comparing these values to what they would be with perfect information we see that model 2 performs best of all the models, but none of the models perform as well as perfect information. None of the models for any runs attain the perfect information values. The results in Table 4 indicate that on average, model 2 will result in 0.5 more total fragments than either model 1 or 3. But there is uncertainty, as choosing model 3 might yield a better result than model 2.

References

Borgatti, S.P., 2002, "The Key Player Problem," Proceedings from National Academy of Sciences Workshop on Terrorism, Washington DC.

Freeman, J.A. and D.M. Skapura, 1991, Neural Networks: Algorithms, Applications, and Programming Techniques, Addison-Wesley Publishing Company, Mass.

Killworth, P.D. and H.R. Bernard, 1979, "Informant Accuracy in Social Network Data III: A Comparison of Triadic Structure in Behavioral and Cognitive Data," *Social Networks*, **2**, 10-46.

Krackhardt, D., 1987, "Cognitive Social Structures," *Social Networks*, **9**, 104-134.

Krackhardt, D. and K. Carley, 1998, "A PCANS Model of Structure in Organization," In: *Proceedings of the 1998 International Symposium on Command and Control Research and Technology*, 113-119.

McPherson, J.M., L., Smith-Lovin, 1987, "Homophily in Voluntary Organizations: Status Distance and the Composition of Face-to-Face Groups," *American Sociological Review*, **52**, 370-379.

McPherson, J.M., P.A. Popielarz, and S. Drobnic, 1992, "Social Networks and Organizational Dynamics," *American Sociological Review*, **57**, 153-170.

Newman, M.E.J., 2002, "The Spread of Epidemic Disease on Networks," Sante Fe Institute Working Paper, available online: http://www.santafe.edu/sfi/publications/wpabstract/20020402.

Skvoretz, J., 1990, "Biased Net Theory: Approximations, Simulations, and Observations," *Social Networks*, **12**, 217-238.

Valente, T.W., S.C. Watkins, M.N. Jato, A.V. Straten, L.M. Tsitsol, 1997, "Social Network Associations with Contraceptive Use Among Cameroonian Women in Voluntary Associations," *Social Science Medicine*, **45**(5), 677-587.