Panel Discussion

David M. Blei

Princeton University
Princeton, NJ 08544, USA
blei@cs.princeton.edu

In this volume, we have seen several compelling reasons for the statistical analysis of network data.

1. Find statistical regularities in an observed set of relationships between objects. For example, what kinds of patterns are there in the friendships between co-workers?
2. Understand and make predictions about the specific behavior of certain actors in a domain. For example, who is Jane likely to be friends with given the friendships we know about?
3. Make predictions about a *new* actor, having observed other actors and their relationships. For example, when someone new moves to town, what can we predict about his or her relationships to others?
4. Use network data to make predictions about an actor-specific variable. For example, can we predict the functions of a set of proteins given only the protein-protein interaction data?

All of the analysis techniques proposed here are model-based: one defines an underlying joint probability distribution on graphs and considers the observed relationship data under that distribution. Loosely—and this will be a point of discussion among the panelists—the models can be divided into those that are "descriptive" or "discriminative" and those that are "generative."

A descriptive graph model is one where the number of nodes in the observed graph or graphs is held fixed and the joint distribution is defined over the edges of that fixed set. The influential exponential random graph model is a general formulation of a descriptive graph model [1,2]. In this framework, the distribution of the entire graph structure is an exponential family with sufficient statistics that are aggregates of the entire graph, e.g., the number of triangles.

In a generative graph model, there is a clear probabilistic mechanism for expanding the graph to new nodes and new edges. The paper by Goldenberg and Zheng is a full generative graph model: the joint distribution is built around the notion of new actors and new connections between existing actors. There is still a joint distribution over the observed graphs. However, the probability of a new node is well-defined and the probability of a new edge can be computed without recalibrating the distribution.

There is ample room for middle ground between these categories. Several papers define hierarchical models based on the latent space approach [3]. These models are generative in the sense that new edges are conditionally independent of the others and have a well-defined probabilistic generative process. But they

E.M. Airoldi et al. (Eds.): ICML 2006 Ws, LNCS 4503, pp. 186–194, 2007.

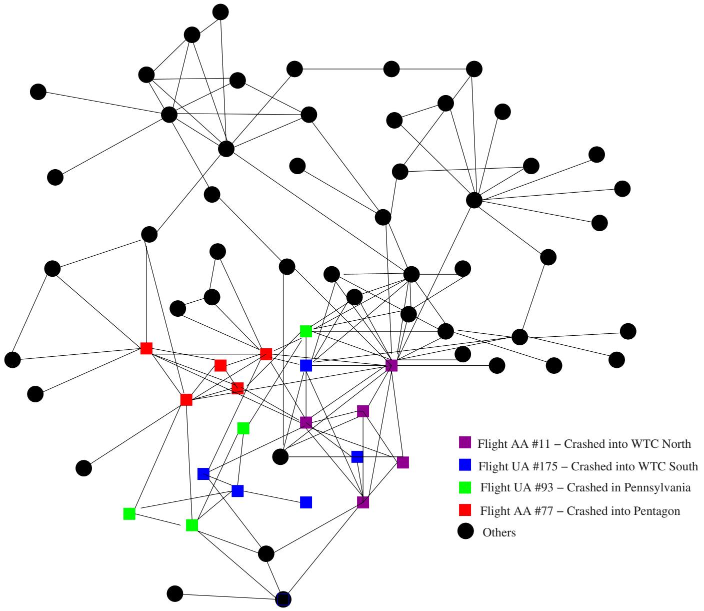
🟪 Flight AA #11 – Crashed into WTC North
🟦 Flight UA #175 – Crashed into WTC South
🟩 Flight UA #93 – Crashed in Pennsylvania
🟥 Flight AA #77 – Crashed into Pentagon
● Others

Fig. 1. 9-11 Hijacker/Terrorist Network. Source: [4].

are somehow not "as generative" as Goldenberg and Zheng's model, where the evolution of the social network is part of the fabric of the generative process.

This distinction was only one of the issues addressed in the workshop that accompanied this volume. As in any kind of data analysis, the tools required depend on the job at hand. We saw work on modeling sequential observations from social networks, modeling multiple data types such as citations and text, fitting graphs organized into hierarchies, and developing new statistics for the exponential random graph model.

Many of these tools were presented for the first time at the workshop. In this panel discussion, we have asked some of our distinguished participants to reflect on the contents and offer a comparative perspective.

Stephen E. Fienberg
Carnegie Mellon University
Pittsburgh, PA 15213, USA
fienberg@stat.cmu.edu

The papers at this workshop when taken together capture many fascinating aspects of network modeling. Prodded by some earlier discussion, I thought it would be useful to begin by reminding us about the two different kinds of graphical representations of the traditional $n \times p$ data array for n individuals or units by p variables. Graphical models [5] are used to represent relationships among the variables in terms of independence relationships. Graphical representations for networks are used to represent relationships among the units, and it is precisely because we don't have conditional independences in the usual dyadic models that things are somewhat complex. Our goal in this workshop has been to both focus on the latter kind of network models and to show how to link them in different ways to probabilistic/statistical models for the variables, either through representations of covariate information (see the paper here by Handcock) or or via mixed membership models (see the papers here by Airoldi et al. and by McCallum et al.).

As others have already noted, we have seen two different types of network models—generative (or what Shalizi et al. call agent-based in their paper) and descriptive. In the latter, which includes the class of p^* models described here by Wasserman et al., we identify *motifs* such as triads or stars and then build models that use them as primary data summaries, e.g., sufficient statistics. When we focus on the evolution of networks we can often be blending the two types of models although they can still be purely descriptive, c.f. the paper here by Henneke and Xing. An interesting question we have raised is whether the latent space models of [3] are descriptive or generative. At one level they appear to be descriptive but they are quite similar in other ways to the mixed membership stochastic block models in the papers by Airoldi et al. and by McCallum et al., which are generative in nature see also [6] and the discussion that follows it.

One issue on which we have not dwelled but which is implicit in the discussions of the distinction between the models is the nature of the data at hand. When we ask what are the data and where do they come from we are really asking a generative question which frames the nature of the models we should be considering. Consider the reported "network" demonstrating the links among the 9/11 hijackers that the press and administration officials are so fond of describing. Figure 1 shows perhaps the most carefully constructed version of it due to [4] What types of linkages do the edges in the graph represent, i.e., to what variables do they correspond? Was the graph constructed to assure that there are paths linking the hijackers to one another? The network picture shows linkages to others beyond the 9/11 hijackers with Arabic names. Are these individuals to be considered hijacker accomplices or confederates? What about others to who linkages could have been made beyond the horizon of observability from each other? After all, many hijackers on the same flight were more than 2 steps away from each other. Finally, real linkages in a terrorist network are dynamic but Figure 1 represents data collapsed over time.

Let me end by summarizing what I see as three major statistical modeling challenges in the analysis of network data. These relate to both the quality and the ease of inference:

Computability. Can we do computations exactly for large networks, e.g., by full MCMC methods, or do we need to resort to approximations such as those involved in the variational approximation such as in the paper by Airoldi et al.?

Asymptotics. The is no standard asymptotics for networks, e.g., as n goes to infinity, which can be used to assess goodness of fit of models. Thus we may have serious problems with variance estimates for parameters and with confidence or posterior interval estimates. The problem here is the inherent dependence of network data.

Embeddibility. Do our data represent the entire network or are they based on only a subnetwork or subgraph, as in Figure 1? When the data come from a subgraph we need to worry about boundary effects and the the attendant bias they bring to parameter estimates. The one result I know in this area is due to [7] for scale-free models in which they show the extent and nature of the bias. My suspicion is that there are similar issues for most of the models discussed at this workshop and we need to explore the consequences of these.

Andrew McCallum
University of Massachusetts
Amherst, MA 01003
mccallum@cs.umass.edu

Task-Focussed Social Network Analysis

I am a relative newcomer to social network analysis. Although I have been doing some research in SNA for the past few years, most of my research over the past decade has been in natural language processing. With this "outsider's perspective," I'd like to offer a couple of thoughts about possible fruitful future directions for SNA.

First, I encourage work in discriminatively-trained social network models.

Many of the recently-proposed models in my local sub-area of SNA are generative directed graphical models. These include various mixed-membership "topic models" and related models, such as *author-topic* [8], *author-recipient-topic* [9], *role-author-recipient-topic* [10], *group-topic* [11], *infinite-relational* [12], *entity-topic* [13], *relational mixed membership* [14], and *community-user-topic* [15]. Other generative models are mentioned by the other panelists.

Although research NLP was dominated by generative models (such as hidden Markov models and probabilistic context free grammars) for decades, the past five years have seen a much stronger emphasis on "discriminative" conditional-probability-trained models, such as logistic regression, maximum entropy models and conditional random fields. Here model parameters are estimated by maximizing the conditional probability of some output variables given some input context. Because the model is not responsible for generating the input context, we need not be concerned about violated independence assumptions among the input variables, and we are free to use rich, expressive sets of overlapping input features. In NLP, the move from generative models to discriminative models typically yields significant gains in accuracy.

Like in natural language, social network data sets are often rich in context, multiple modalities, and other non-independent variables that would benefit from a discriminative approach. We have begun research toward "conditionally-trained topic models" with our work on *multi-conditional learning*, and in particular *multi-conditional mixture models* [16].

Second, in a related point, I encourage emphasis on particular tasks.

Much past work in SNA approaches the problem as a scientist—we observe some natural phenomenon, and attempt to build models that capture them.

These include foundational work in descriptive graph properties, generative models of graphs, etc. It is also interesting (and sometimes more useful) to approach a domain as an engineer—asking "What is the real task we are trying to solve?" "What is the use-case?" "What is the decision problem?"

There are, of course, many important use-cases for social network analysis: deciding who to promote, finding an expert, selecting the right actions to improve (or harm) an organization, identifying likely illicit behavior, selecting the best collaborator, finding new music I'm likely to enjoy, predicting which team will get the job done best.

Scientifically descriptive models may have something to say about these tasks, but discriminative SNA models could focus on tuning their parameters for best accuracy on these particular tasks. As interest in SNA expands, I predict that there will be more research on models designed to address particular tasks.

Cosma Rohilla Shalizi
Carnegie Mellon University
Pittsburgh, PA 15213, USA
cshalizi@cmu.edu

Looking back over the papers presented at this workshop, I am struck by two cross-cutting contrasts, which I want to explore a little here. The first contrast is between models of phenomena and models of mechanisms, which doesn't quite, I think, map on to Prof. Fienberg's contrast between descriptive and generative models. The other contrast is between small networks which we know in rich contextual depth, and big networks where our knowledge is shallow and impoverished. Before elaborating on our divisions, however, I would like to say a few words about what we all seem to have in common. As the token representative of statistical physics on the panel, I will be deliberately provocative and say that what unites us is a devotion to the ideals of statistical mechanics.

Of course, of the participants at the workshop, only Dr. Clauset and myself were even arguably statistical physicists — and really he's a computer scientist and I'm a statistician. But the goal of statistical mechanics is to explain large-scale, macroscopic phenomena as the aggregated result of small-scale, microscopic behavior, as the result of interactions among individuals in contexts which themselves result from small-scale interactions among individuals. Global patterns should derive from local interactions. And this, I think, is something we would all be comfortable endorsing. Certainly when I heard Prof. Krackhardt explain that social networks matter because they show the contextual determinants of behavior, or when I saw Prof. Handcock check his ERGMs by seeing whether they could go from homophily and transitivity (local interactions) to the distribution of geodesic distances (a global pattern), my inner statistical mechanic felt right at home.[1] So, I'd claim that we're united by wanting to understand context and interaction, and how these lead to global patterns.

[1] To be sure, an inner economist, or an inner evolutionary biologist, would *also* have felt at home.

The first contrast, then, concerns *how* we do this. Once we have committed ourselves to *generating* the macroscopic patterns, we still need to decide whether we do this by modeling the *mechanisms* of action and interaction, and hope we get them right, or by modeling the consequences of interactions, and hope the details don't matter. The former leads to mechanistic models, the latter to what physicists call "phenomenological" ones. Take, for example, homophily. The model presented by Goldenberg and Zheng, for instance, is a mechanistic model, with a fairly detailed representation of the process by which people come to form social bonds, one consequence of which is homophily.[2] We also had phenomenological models of homophily, including both the ERGMs of Handcock and Morris, and the dynamic latent space model of Sarkar, Siddiqi and Gordon. Random walks in social space are obviously unrealistic, but may well be a good first approximation to reality; the ERGMs are simply silent about the dynamical processes by which networks form[3]. I *want* to have that a mechanistic understanding of the systems I study, so I find phenomenological models, as such, less than fully satisfying. But I recognize that there are very good reasons to use them, not the least of which is that they are much easier to get right. If Handcock and Morris want to measure the strength of homophily relative to transitivity, their problem is comparatively straightforward: estimate some parameters — with sufficient statistics, no less. If Goldenberg and Zheng want to make the same measurement for their model, the inferential problems are much more complicated, *because* their model includes mechanisms and not just phenomena.

The contrast between mechanistic and phenomenological models, then, seems to run through almost all the contributions here. But there is no reason we cannot have both sorts of models, or why we should think they contradict each other. In fact, I think this contrast is potentially productive of new research, since there should be ways of systematically deriving phenomenological models from mechanistic ones, and conversely of using well-estimated phenomenological models to constrain guesses about mechanisms.

I turn now to the second contrast, which is not between models of networks, but the networks themselves, or at least our representations of them. In small networks, like the karate club, or even the Colorado Springs sex-and-drugs network, we have, if not necessarily "thick descriptions" in the ethnographic sense, at any rate *deep* ones. We know a reasonable amount about each of the nodes, and sometimes (as in the karate club) can tell a story about each of the edges. We have, in other words, a lot of context, which is what we want. But, precisely because there is some much detail, it can be difficult, at a qualitative level, to distinguish an analytical narrative from a mere just-so story. If we then turn to quantitative models (which, as mathematical scientists, we're inclined to do

[2] In fact, they have what people in complex systems would call an "agent-based model". So far as I know, they are the only people to combine such a model with proper inference.

[3] The interesting paper by Hanneke and Xing adds dynamical detail to ERGMs, but makes no mechanistic commitments.

anyway), the small size of the network severely limits our ability to discriminate among models; the maximum attainable power is low.

Of course, we are no longer limited to small networks like the karate club; some of the graphs we saw presented at the workshop, like the PGP keyring network, had several million nodes, making them about five orders of magnitude larger than the karate club. This is exciting in its own right, because hitherto we have had almost no information about the *fine-grained* social organization of *large* populations. And certainly we no longer have much difficulty statistically distinguishing the predictions of different models! Dealing with this volume of strongly-dependent data does raise interesting technical problems; for instance, it's not obvious that models developed on small- and medium- sized networks can scale up, either computationally or descriptively, to such large networks. But beyond those technical problems, there is what seems like an intrinsic difficulty, which is that our knowledge of these large networks is shallow. It is simply not possible to have richly detailed information on all of the nodes, never mind all of the edges. When we look at *any* network with a few hundred thousand nodes, we are always going to be ignoring a huge amount of context about the nodes and their interactions. This isn't just a problem for social networks, but would also apply to, say, gene-regulatory networks.

So, opening up the divide between small networks and large, we find it contains a dilemma. Either we can possess the rich contextual detail we are interested in, or we can have enough data to severely test our models. Perhaps some clever methodology can cut a path through this dilemma; I myself don't see how.

Mark S. Handcock
Department of Statistics
University of Washington
Seattle, WA 98195-4322 USA
handcock@stat.washington.edu

The development of exponential family random graph models (ERGM) for networks has been limited by three interrelated factors: the complexity of realistic models, dearth of informative simulation studies, and a poor understanding of the properties of inferential methods.

The ERGM framework has connections to a broad array of literatures in many fields, and I emphasize its links to spatial statistics, statistical exponential families, log-linear models, and statistical physics.

Historically, exploration of the properties of these models has been limited by three factors. First, the complexity of realistic models has limited the insight that can be obtained using analytical methods. Second, statistical methods for stochastic simulation from general random graph models have only recently been become available [17,18,19]. Because of this, the properties of general models have not been explored in depth though simulation studies. Third, the properties of statistical methods for estimating model parameters based on observed networks have been poorly understood. The models and parameter values relevant to real networks is therefore largely unknown. Significant progress is now being made in each of these areas. However, despite their elegance and pedigree, the ERGM framework have yet to prove their value in addressing real scientific questions of interest. They have the tendency to produce degenerate behavior as a result of their maximum entropy properties [20]. This hinders simple model specification. The papers presented at the workshop illustrated many alternative approaches that may prove more fruitful.

The discussion of "generative" verses "descriptive" models was dialectic in nature. The exponential random graph models can be clearly be interpreted as descriptive. However, if we take the term generative to mean the ability to simulate network structures with given structural properties, they are also generative. If by generative is meant dynamic changing edges and structures then the paper of Steve Hanneke and Eric Xing illustrated how this can be achieved within the ERGM framework. If a probabilistic mechanism for adding additional nodes temporally is an regarded as an essential characteristic of a generative model then the published work on ERGM models does not meet this criterion. Note, however, that this is well within the capabilities of exponential family models. The model of Goldenberg and Zheng has a more directly generative mechanism and may be preferred for this reason.

The latent space framework invented by [3], and expanded by others at the workshop was originally descriptive in nature. However, variants of it can have a generative flavor (e.g., hierarchically adding a Gaussian mixture model for the positions).

As noted in the discussion, there are many challenges facing statistical network modeling. I believe the more traditional ones: inference from sampled data rather than a census, the development of statistical testing procedures, and their associated computational issues, will be overcome. The fundamental challenge is adapting the choice of models to the scientific objectives. Network phenomena are complex and the models must choose the specific features to be represented well while being ambivalent about the others.

Let me end by noting the success of this workshop in bringing together statistical network modeling researchers from distinct disciplines and scientific frameworks. The disciplines have much to communicate to each other especially where their scientific goals overlap. In the few cases were such researchers are brought together to speak, there has been little cross-disciplinary listening going on. This workshop was able to overcome that barrier so that researchers with backgrounds in SNA, physics, computer science or statistics were listened to. This success owes much to the principal organizers.

References

1. Frank, O., Strauss, D.: Markov graphs. Journal of the American Statistical Association **81**(395) (1986) 832–842
2. Strauss, D., Ikeda, M.: Pseudolikelihood estimation for social networks. Journal of the American Statistical Association **85**(409) (1990)
3. Hoff, P., Raftery, A., Handcock, M.: Latent space approaches to social network analysis. Journal of the American Statistical Association **97**(460) (2002) 1090–1098

4. Krebs, V.E.: Mapping networks of terrorist cells. Connections **24**(3) (2002) 43–52
5. Jordan, M.: Graphical models. Statistical Science **19**(1) (2004) 140–155
6. Handcock, M.S., Raftery, A.E., Tantrum, J.M.: Model-based clustering for social networks (with discussion). Journal of the Royal Statistical Society, Series A **170** (2007) in press
7. Stumpf, M.P.H., Wiuf, C., May, R.M.: Subnets of scale-free networks are not scale-free: Sampling properties of networks. Proceedings of the National Academy of Sciences **102**(12) (2005) 4221–4224
8. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. (2004)
9. McCallum, A., Corrada-Emmanuel, A., Wang, X.: A probabilistic model for topic and role discovery in social networks and message text. In: International Conference on Intelligence Analysis. (2005)
10. McCallum, A., Corrada-Emanuel, A., Wang, X.: Topic and role discovery in social networks. In: Proceedings of IJCAI 2005. (2005)
11. Wang, X., Mohanty, N., McCallum, A.: The author-topic model for authors and documents. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. (2004)
12. Kemp, C., Tenenbaum, J.B., Griffiths, T.L., Yamada, T., Ueda, N.: Learning systems of concepts with an infinite relational model. In: AAAI. (2006)
13. Newman, D., Chemudugunta, C., Smyth, P., Steyvers, M.: Statistical entity-topic models. In: Proceedings of the 12th ACM SIGKDD Conference (KDD-06). (2006)
14. Airoldi, E., Blei, D., Xing, E., Fienberg, S.: Stochastic block models of mixed membership. Journal of Bayesian Analysis **(submitted)** (2006)
15. Zhou, D., Manavoglu, E., Li, J., Giles, C., Zha, H.: Probabilistic models for discovering e-communities. In: 15th International World Wide Web Conference (WWW2006). (2006)
16. McCallum, A., Pal, C., Wang, G.D.X.: Multi-conditional learning: Generative/discriminative training for clustering and classification. In: AAAI. (2006)
17. Snijders, T.A.B.: Markov chain monte carlo estimation of exponential random graph models. Journal of Social Structure **3**(2) (2002)
18. Handcock, M.S.: Degeneracy and inference for social network models. In: Paper presented at the Sunbelt XXII International Social Network Conference in New Orleans, LA. (2002)
19. Hunter, D.R., Handcock, M.S.: Inference in curved exponential family models for networks. Journal of Computational and Graphical Statistics **15** (2006) 1–19
20. Handcock, M.S.: Assessing degeneracy in statistical models of social networks. Working paper, Center for Statistics and the Social Sciences, University of Washington (2003)