Statistical and Physical Models for Social Networks and Their Evolution

Anna Goldenberg

November 15, 2004

Recently, the area of Social Networks has gained popularity in computer science partially due to the increased efforts in the domain of security. One of the goals of analysis in this area is to detect and evaluate relationships between individuals that may be part of terrorist networks (Krebs, 2002). It should be noted, however, that statistical analysis of social networks spans over 60 years. Since the 1970s, one of the major directions in the field was to model probabilities of relational ties between interacting units (social actors), though in the beginning only very small groups of actors were considered. Extensive introduction to earlier methods is provided by Wasserman and Faust (1994). Two of the most prominent current directions are Markov Random Fields (MRFs) introduced by Strauss and Frank (1986) and Exponential Random Graphical Models (ERGMs), also known as p^* (Wasserman and Pattison, 1996; Anderson et al, 1999). The ERGM have been recently extended by Snijders et al (2004) in order to achieve robustness in the estimated parameters.

The statistical literature on modeling Social Networks assumes that there are n entities called *actors* and information about binary relations between them. Binary relations are represented as an nxn matrix Y, where Y_{ij} is 1, if actor i is somehow related to j and is 0 otherwise. For example, $Y_{ij} = 1$ if "i considers j to be friend". The entities are usually represented as nodes and the relations as arrows between the nodes. If matrix Y is symmetric, then the relations are represented as undirected arrows. More generally Y_{ij} can be valued and not just binary, representing the strength (or value) of the relationship between actors i and j (Robins et al, 1999). In addition, each entity can have a set of characteristics x_i such as their demographic information. Then the n dimensional vector $X = x_1, \ldots, x_n$ is a fully observed covariate data that is taken into account in the model (e.g. Hoff et al, 2002)

Predominantly the social network literature focuses on modeling P(Y—X), i.e. on probabilistically describing relations among actors as functions of their covariates and also properties of the graph, such as indegree and outdegree of individual nodes. A complete list of graph-specific properties that are being modeled can be found in (Snijders et al, 2004). Thus, the models are geared to probabilistically explain the patterns of observed links and their absence between n given entities.

There are several useful properties of the stochastic models listed in a brief survey work by Smyth (2003). Some of them are:

- the ability to explain important properties between entities that often occur in real life such as reciprocity, if i is related to j then j is more likely to be somehow related to i; and transitivity, if i knows j and j knows k, it is likely that i knows k.

- inference methods for handling systematic errors in the measurement of links (Butts,2003)

- general approaches for parameter estimation and model comparison using Markov Chain Monte Carlo methods (e.g. Snijders, 2002)

- taking into account individual variability (Hoff:2003) and properties (covariates) of actors (Hoff,2002)

- ability to handle groups of nodes with equivalent statistical properties (Wang and Wong, 1987).

There are several problems with existing models such as degeneracy analyzed by Handcock (2003) and scalability mentioned by several sources (Hoff et al, 2002; Smyth, 2003). The new specifications for the Exponential Random Graph Models proposed in Snijders et al (2004) attempt to find a solution for the unstable likelihood by proposing slightly different parametrization of the models than was used before. Experiments show that the parameters estimated using the new approach yield a smoother likelihood surface that is more robust and is less susceptible to the degeneracy problem. The scalability remains to be a major issue. Datasets with hundreds of thousands of entities are not uncommon in the Internet and co-authorship based domains. To our knowledge, there are no statistical models in the social networks literature that would scale to thousands or more actors. Parameter estimation in general for Markov Random Fields is well-known to be intractable for large number of variables due to the computational complexity of the normalization constant which requires summation over all possible graphs with n nodes. The scalability problem has also been attributed to the tendency of the models to be global, i.e. most operate on the full covariance matrices (Hoff et al, 2002). The use of MCMC approaches that tend to have slow convergence rate may also hinder computational speed of the parameter estimation in high dimensions.

One of the more recent directions is latent variable models that may be able to avoid the problems related to the use of Markov Random Graphs. For example, the work of Hoff et al (2002) proposes a model in which it is assumed that each actor i has an unknown position z_i in a latent space. The links between actors in the network are then assumed to be conditionally independent given those positions and the probability of a link is a probabilistic function of those positions and actors' covariates. The latent positions are estimated from data using logistic regression. The general form of the model is:

$$logodds(y_{ij} = 1|z_i, z_j, x_{ij}, \alpha, \beta) = \alpha + \beta^T x_{ij} + d(z_i, z_j) \quad (1)$$

, where $d(z_i, z_j)$ is a distance metric between positions of the actors in latent space. This model is though promising also suffers from the lack of scalability of the parameter estimation.

It is also worth mentioning that a graph theoretic area of physics that studies complex systems is directly applicable to social network modelling. Though modeling of complex systems has developed seemingly in parallel to the statistical modeling of social networks in social science, the findings in this area can assist in understand further the phenomenon of real networks organization and structure. The assumptions are the same: there are n actors (nodes) and there are N links between those nodes representing relationships among actors. The goal is also to understand and model structural properties of the naturally occurring networks. The base model describing random graphs was developed by Erdos and Reny (1959,1969,1961), where expected number of edges in the graph is $E(N) = p(\frac{(n-1)n}{2}$ and the probability of obtaining the observed graph is $P(G_o) = p^N(1-p)^{\frac{n(n-1)}{2}-n}$. However, it was noted that the degree distribution in the random graphs does not follow power law $P(k) \sim k^{-\gamma}$ common to the realistic networks. Thus, "scale-free networks" were introduced

(Barabasi and Albert, 1999; Barabasi et al, 1999) . Newman et al (2001) have developed a generalized random graph model where the degree distribution is given as an input parameter. The research in the field of physics gives more insight into graph growth, given the proposed models, such properties of the emerging graphs as clusterability, graph diameter and the formation of a large component. A great summary of the past and ongoing work and it's relation to statistical physics is given by Barabasi et al (2002).

One of the important properties of real life networks is evolution over time. It can be expected that co-authorship networks can be relatively stable, whereas such dynamic online communities as Friendster may significantly change in a matter of weeks. In terms of modelling a change in a social network, emergence of a new interaction means an addition of a new edge, whereas severing a relationship means deletion of an edge. The principles underlying the mechanisms by which relationships evolve are still not well understood (Liben-Nowell and Kleinberg, 2003). There are several potentially different approaches based on the objective the researchers are optimizing. Works of Jin et al (2001), Barabasi et al (2002) and Davidsen (2002) in physics along with Van de Bunt et al (1999) and Huisman and Snijders (2003) in social sciences, generally evaluate their models for network evolution by comparing structural properties and features of the developed models to those of real networks. Another direction is to model evolution aiming to make inferences, i.e. based on the properties of the network seen so far, to infer who are the most likely future friends or collaborators (Newman, 2001; Liben-Nowell and Kleinberg, 2003). Such models are still in their infancy, having similar problems with scalability and incorporation of secondary factors, such as graduation or relocation that have great impact on real life networks.

References

Albert, R., & Barabasi, A.-L. (2002). Statistical mechanics of social networks. *Reviews of Modern Physics*, *74*.

Anderson, C., Wasserman, S., & Crouch, B. (1999). A p^* primer: logit models for social networks. *Social Networks*, *21*, 37–66.

Barabasi, A., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaboration. *Physica A*, *311*, 590–614.

Barabasi, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509–512.

Barabasi, A.-L., Albert, R., & Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physica A*, *272*, 173–187.

Butts, C. (2003). Network inference, error, and informant (in)accuracy: a Bayesian approach. *Social Networks*.

Davidsen, J., Ebel, J., & Bornholdt, S. (2002). Emergence of a small world from local interactions: Modeling acquaintance networks. *Physical Review Letters*, *88*.

Frank, O., & Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, *81*, 832–842.

Handcock, M. (2003). *Assessing degeneracy in statistical models of social networks*Working Paper 39). University of Washington.

Hoff, P. (2003). Random effects models for network data. *Proceedings of the National Academy of Sciences*.

Hoff, P., Raftery, A., & Handcock, M. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, *97*, 1090–1098.

Huisman, M., & Snijders, T. (2003). Statistical analysis of longitudinal network data with changing composition. *Sociological Methods and Research*, *32*, 253–287.

Jin, E., Girvan, M., & Newman, M. (2001). The structure of growing social networks. *Physical Review Letters E*, *64*.

Krebs, V. (2002). Mapping networks of terrorist cells. *Connections*, *24*, 43–52.

Liben-Nowell, D., & Kleinberg, J. (2003). The link prediction problem for social networks. *Proc. 12th International Conference on Information and Knowledge Management*.

Newman, M. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences USA* (pp. 404–409).

Robins, G., Pattison, P., & Wasserman, S. (1999). Logit models and logistic regressions for social networks iii. valued relations. *Psychometrika*, *64*, 371–394.

Smyth, P. (2003). Statistical modeling of graph and network data. *IJCAI Workshop on Learning Statistical Models from Relational Data*.

Snijders, T. (2002). Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, *3*.

Snijders, T., Pattison, P., Robins, G., & Handcock, M. (2004). New specifications for exponential random graph models. Submitted for publication.

Van De Bunt, G., Duijin, M. V., & Snijders, T. (1999). Friendship networks through time: An actor-oriented dynamic statistical network model. *Computation and Mathematical Organization Theory*, *5*, 167–192.

Wang, Y., & Wong, G. (1987). Stochastic blockmodels for directed graphs. *Journal American Statistical Association*, *82*, 8–19.

Wasserman, S., & Pattison, P. (1996). Logit models and logistic regression for social networks: I. an introduction to markov graphs and p^*. *Psychometrika*, *61*, 401–425.