Proposal for a scalable class of graphical models for Social Networks

Anna Goldenberg

November 24, 2004

Abstract

This proposal is about new statistical machine learning approaches to detect evolving relationships among large numbers of entities. One example is the set of friendship relations among participants in Friendster. Three major recent developments make it an important area of study. First, there has been a dramatic increase in collection of data, providing input to a system that deduces an underlying social-network-like structure. Second, there are immediate uses for this technology in business, security and the social sciences. Third, the recent rapid growth in probabilistic and statistical approaches to tractable machine learning mean that it may be possible to analyze networks with millions of entities. Statistical models in traditional social network literature can manage at most a few hundred nodes.

In this thesis I am planning to develop a new stochastic model for describing relations in a social network and evolution of those relations over time. Initial explorations led to the creation of an algorithm for structural search of Bayesian Networks from sparse data (Goldenberg and Moore, 2004). SBNS is a scalable search procedure that learns Bayes Nets from the binary events data, i.e. the estimation is based solely on information about which entities (variables) participated in the set of given events (records). I am planning to build on this work by analyzing the properties of the obtained links. Also, it is important to introduce secondary characteristics of the entities into the model, such as profession or location for people, to make sure that the strength of relationships between them is affected by their similarity or dissimilarity on a more personal level. This will be accomplished by incorporating the secondary information in the parameter estimation process. Finally, I am planning to extend the model to account for the evolution of the social networks over time.

1 Introduction

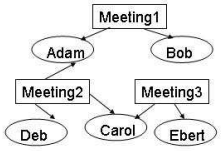
Several upcoming areas can benefit from scalable stochastic models that infer complex network structures from transactional or more detailed data. Extensive growth of the online social forums, cross-world collaborations, preferential customer networks measuring in hundreds of thousands or even millions of participating entities all require robust scalable analysis. These communities are naturally represented using *social networks* - structures where nodes represent people (or groups, organizations) and links (or edges) between them represent their relationships, interaction, influence. Ongoing work on stochastic models in social science (Snijders et al, 2004) and descriptive models of complex networks in physics sheds light on the topological properties and relationships and their dependencies in real life networks. However, most of this research assumes that the links and relations between entities are observed, given. It is then possible to study properties such as degree distribution or number of particular patterns, such as k-stars in the network. In our work, we assume that there are observations, particularly *events* relating entities, however the true underlying structure is not observed. We are not attempting to find the true underlying graph connecting the entities, rather by probabilistically modeling dependencies between entities from the events data we aim to make inferences about relations between them and perhaps their further actions.

We are proposing the use of graphical models to describe dependencies in the unobserved networks. The use of graphical models in modeling event data has been advocated in several domains in the literature already. For example, inferring customer preferential networks has been extensively studied by Breese et al (1998) and later Heckerman et al (2000). Graphical models are also very prominent in modeling relational data especially when detailed information about relations and or social actors is available (Getoor et al, 2002; Heckerman et al, 2004; McGovern, 2003, etc). Another significant area of graphical models application is genetics where they can be used to describe regulatory networks (Friedman, 2004).

1.1 Datasets

Let's start with the simplest datasets: binary, where each record denotes a collection of entities that participated in an "event". Examples are a software purchasing database where each record is a set of items purchased in a single transaction; the online library of computer science publications Citeseer, where each record is a list of co-authors of a particular paper; a click-database where each record is a set of clicks a user made when visiting a site of a particular company. The term *binary* dataset comes from the fact that each record r consists of a set of ones and zeros: $r_{ij} = 1$, if entity i participated in event j, and $r_{ij} = 0$ otherwise. An example of a dataset and corresponding representation are depicted on Figure 1.

These datasets have one important property in common. Each record in these large datasets consists mostly of zeros: they are extremely sparse. Sparse-



	Adam	Bob	Carol	Deb	Ebert	...
Meeting1	1	1	0	0	0	0
Meeting2	1	0	1	1	0	0
Meeting3	0	0	1	0	1	0
.....	...	...	...	...	...	...

Figure 1: An example of representation (on the left) of the data (on the right). Nodes in the network are people. Rectangles are events relating them.

ness has been considered hazardous in statistics as it may give rise to degeneracy in models. In fact, sparseness has many advantages that are very important for computational scalability. While the problems of degeneracy arise when attempting to build a global model, sparseness is helpful to quickly identify significant local models that can later be combined into a global model. It also is instrumental in greatly improving the speed of counting that is essential in obtaining sufficient statistics.

1.2 Modeling Networks

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).

1.2.1 Network Modeling in Social Science

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. Those 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.

1.2.2 Network Modeling in Physics

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), where expected number of edges in the graph is $E(N) = p(\frac{(n-1)n}{2})$, where p is the probability of having any edge, 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 its relation to statistical physics is given by Barabasi et al (2002).

1.2.3 Other

A variety of other domains greatly benefit from network analysis. For example in Biology, motif search in biological networks is facilitated by studying various graph properties such as local graph alignment (Berg and Lassig, 2004). Even though this work is probabilistic in nature and reveals topological graph properties from the given graph, it is not generative and does not generalize to answer

other queries about the data. Friedman (2004) uses probabilistic graphical models to gain insights into biological mechanisms governing cellular networks. This work is probably the closest to ours in its nature. Several assumptions about domain knowledge are used in the constructed models. We make no assumptions about the structure of the graph underlying the resulting set of events and learn the dependencies among actors from data directly.

Another work similar in spirit to ours is use of relational dependency networks RDNs to answer classification queries posed to the network of entities (Neville and Jensen, 2003). Though the network in this case appeared to be quite large (over 300,000 entities), it was considered to be given. McGovern et al gives another nice analysis of the high energy physics community based on the citation graph was done by (McGovern et al, 2003). In this work, graph properties for very large graph of citations (static analysis) was done using sampling. Also, several predictive tasks were tested by learning Relational Probability Trees (Neville et al, 2003) based on featurized data. Again, sampling was used to train the model resulting in good prediction accuracy. Above work on analyzing social communities and learning tasks is based on given network structure, whereas our work is geared to learning that structure. These two directions are complementary to each other.

Mapping Knowledge Domains is yet another area that unites physicists, biologist, computer scientists in order to understand the formation of knowledge domains, their structures and properties. A lot of the work in this area assumes that a knowledge database can be represented and visualized as a graph structure. The research is geared towards understanding various properties of the domains, such as topics clustering, number of papers written by authors in single or across multiple domains, length of the path between actors in co-authorship networks, etc. Methods used to describe those properties and the networks are similar in spirit though not always limited to those found in physics literature. A great overview of this work can be found in the special issue of PNAS (April 2004) dedicated to this topic.

1.2.4 Network Evolution

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 relatively short period of time. In terms of modelling a change in a social network, 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.

1.3 Thesis goals

In my recent work on obtaining scalable models for binary event data, I developed an algorithm for structural search of Bayesian Networks, named Screen-based Bayes Net Structural search (SBNS). Some of the useful properties of the algorithm are scalability to over 10^5 actors while maintaining the ability to search pairwise, three-way and higher-order interactions. Experiments showed that SBNS finds better fitting models than the only available scalable alternative: random hill-climbing approach. It was also evident that inclusion of high order interactions increased the accuracy. My goals are to examine and improve on the properties of the SBNS-created-model when applied to social network domains; to extend the model to incorporate secondary characteristics, i.e. additional information about actors and finally to modify the model to allow evolution over time. I believe that SBNS is a solid framework that could easily facilitate further developments to achieve scalable models for social networks.

The rest of this proposal is structured as follows. First, I briefly introduce the existing SBNS model and discuss its benefits and shortcomings and then elaborate on the exact goals and my plan to achieve them.

2 SBNS model

My initial explorations in the area of large scale graphical models from sparse data led to the creation of SBNS (Goldenberg and Moore, 2004). SBNS is an algorithm for tractable structural learning of Bayesian Networks in the presence of sparse event data as described in Section 1.1.

2.1 Bayes Nets

Lets call entities about which the information is collected "actors". Let's associate random binary variables $X_i, ...X_n$ with an actor's participation in any event. The state of X_i is 1 when actor i has participated in a given event and is 0 otherwise. For example, for a citation database, if two people i and j have co-authored a paper together, then for this event (co-authorship of a given paper) their states are $X_i = 1$ and $X_j = 1$ and the states of all other authors in the database for this event are 0 ($X_k = 0, \forall k \neq i, j$).

We would like to learn the underlying dependencies that trigger the events. In other words, based on the known information about simultaneous participation of actors in observed events, we would like to construct a probabilistic generative model that would describe those events.

A probabilistic model describes a joint over all the random variables of interest. Probabilistic graphical models represent joint distributions that can be expressed as a product of terms each depending on a few random variables. The graph provides a dependency structure between variables for each of the terms in the product and is essential in the inference step.

Bayesian Network (BN) is a set $\{\mathcal{G}, \theta\}$ where $\mathcal{G}$ is a Directed Acyclic Graph $\{\mathbf{V}, \mathbf{E}\}$ ($\mathbf{V}$ is a set of nodes and $\mathbf{E}$ is a set of edges) and θ is a set of parameters obtained by maximizing a Bayesian score, which is usually a penalized likelihood. BNs a type of probabilistic graphical model, where the joint distribution is determined by a product of conditional probabilities, i.e.

$$P(X_1 \dots X_n) = \prod_i P(X_i | Pa(X_i)) \tag{2}$$

, where $Pa(X_i) \in \mathbf{X}$ is a set of parents of the variable X_i in the DAG. Graphically, BNs are represented using directed edges from parents $Pa(X_i)$ to children X_i, for each $i = 1 \dots n$. Acyclicity of the DAG guarantees the product in Equation 2 to be a coherent probability distribution. More information on Bayesian Networks can be found in (Cooper and Herskovitz, 1991; Heckerman et al, 1995).

Note that directed arrows in the graph represent direct dependency of the outcome of variable X_i on its parents $Pa(X_i)$. Note that the dependencies can only be described in terms of the observed data, for example in a citation database case, a relation $X_i -> X_j$, where $Pa(X_j) = X_i$, means that author X_j is likely to appear as co-author of the paper if X_i is one of the co-authors. The dependence can also represent a negative correlation, i.e. in the above case knowing that X_i is one of the co-authors, would make X_j unlikely to be one of the co-authors.

2.2 Algorithm

The scalability of SBNS is achieved by exhaustively searching over structures only on the local level for a large set of small subsets of variables. The advantage of such a structural learning algorithm is that the optimization never needs to be carried out on the global scale. We exploited the computational efficiency of *Frequent Sets* (Agrawal, 1993) for gathering statistics that are most likely to be useful for structure search given the assumption of sparse data. A Frequent Set with support s is a collection of variables (entities, actors) that have co-occurred (i.e. simultaneously had 1 in the binary dataset) more than s times. In our work we show that in sparse data the evidence for positive correlation if such exists is going to be much stronger than for negative one. Given sparse data and a support s greater than about 3, it is surprisingly easy to compute all Frequent Sets (Agrawal, 1994). There is an abundance of literature on Frequent Sets as their collection is an essential part of the association rules algorithms widely used in commercial data mining (Agrawal, 1993; Han and Kamber, 2000). It is important to note that usage of Frequent Sets facilitates finding interactions of order higher than just dyads and triads, the limitation that still exists in majority of the Social Networks models. Finally, there is a simple heuristic

that iterates over potential edges once, efficiently exploiting locally collected statistics and structures to create the global Bayes Net. It was not necessary to make simplifying assumptions such as restricting the number of possible parents and thus impacting the structure of the network, since the local searches are quite inexpensive. Pseudocode for the SBNS algorithm is available in Table 1, the detailed explanation of SBNS is available in (Goldenberg and Moore, 2004).

2.3 Advantages of SBNS framework

The main contribution of SBNS is the ability to perform structural search on datasets such as Citeseer with over $100,000$ variables (authors). Experiments show that SBNS finds models fitting the data better than the only scalable alternative, *random hill-climbing*, on several large datasets including Citeseer. Random hill-climbing is a search algorithm where at each step one of the three operations {addition, deletion, reversal} of an edge is applied to the given graph. The modification to the graph is accepted with probability p only if it improves the score. The table of results and their analysis are available in (Goldenberg and Moore, 2004). Beyond finding the structure, which carries important information in itself, we also showed that it was possible to execute certain queries very quickly even when the networks are so large. For example, one of the queries that was reported by Goldenberg and Moore (2004) is similar to all-but-one task in collaborative filtering, where the goal is to find the most likely subset of variables of given cardinality to complete the given partial set (Heckerman et al, 2000).

The algorithm can be treated as a more general framework. The main ideas that are useful for scalability of modelling any social network are

1. to be able to identify subsets of actors that

   (a) are small enough to allow finding nearly optimal local models optimizing the same scoring function quickly

   (b) upon local parameter estimation provide a substantial amount of statistics (such as counts) necessary for further (global) estimation

2. the model has to be decomposable so that the full joint probability does not require estimation of the full covariance matrix

Block models of Wang and Wong (1987) satisfy the first condition, though the subsets need not be disjoint. The first condition could include soft clustering instead of the Frequent Sets, provided condition 1(b) is satisfied. The second condition eliminates Markov Random Field (MRF) type models in their present state, where the normalization constant has to be computed over all possible states of all variables (computationally prohibitive step).

algorithm	SBNS
input	**K** - max Frequent Set size
	s - support
output	BN - Bayes Net

Also: Ed - Edgedump - a collection of directed edges represented as (source,dest,count)
DS - DAG storage

1. **for** k = 2 .. K
2. obtain counts for all Frequent Sets of size k
3. **foreach** Frequent Set
4. find best scoring DAG
5. **if** DAG contains a node that has $k-1$ parents
6. store DAG in DS
7. **end foreach**
8. **end for**
9. **foreach** DAG in DS
10. store all edges $\{source, dest, count{+}{+}\}$in Ed
11. order Ed in decreasing order of edge counts
12. **foreach** edge $e \in Ed$
13. **if** e doesn't form a cycle in BN
14. **and** e improves BDeu
15. add e to BN
16. **end foreach**
17. return BN

Table 1: Pseudocode for the SBNS algorithm

3 SBNS and Social Networks

I would like to draw a parallel with the Social Network (SN) literature. In the SN paradigm, it is assumed that the links between the actors representing their relationships are given. Having the event database, we could also link the actors based on their simultaneous participation in an event. However, we observe many events in which actors i, j, k, etc may occur together or separately. Thus, these relations are of probabilistic nature and not considered to be given. It is also important to note the difference between the "links" or edges that are found using a structural learning algorithm such as SBNS and the "relationships" between actors. The edges found using SBNS do represent dependency which however cannot be readily interpreted as a relationship in a sense common to SN analysis. SN relationships can be inferred from a learned Bayes Net by asking questions such as "are i and j close collaborators". One of the ways of answering this question would be to find $p(X_i|X_k), \forall k \neq j$ and then find whether $p(X_i|X_j)$ is in the desired percentile, such as in the top 5%.

Another difference between the Social Network paradigm and the probabilistic graphical model (PGM) as described above is the ability of the PGM to succinctly represent the joint distribution, where as the focus of the graphical models commonly used for Social Networks is to fit a model to the existing graph of relationships as closely as possible. In case of PGMs, one of the goals is to represent the joint while minimizing the number of parameters used. This approach allows us to handle models with hundreds of thousands of variables without running out of memory.

One of the benefits of the commonly accepted models describing social networks is the possibility of directly assesing the properties of the graph, such as the number of triangles, k-stars, or degree distribution. It is not possible to answer the questions about properties using Bayes Nets without a number of complex inference steps. However, though it is hard to create easy visualizations of underlying relationships and draw conclusions from SBNS about the underlying network itself, we believe that by using inference it is possible to make judgements about the relationships between actors, one of the ultimate goals of modeling social networks.

4 Example

The purpose of this example is to show similarity and differences in conclusions that can be drawn from a social network represented by an adjacency matrix A and a Bayes Net learned using SBNS from the event data that preserves constraints specified by A.

This illustration is based on a sample dataset *borg4cent* available in the UCINet package[1]. The dataset provides a relationship (adjacency) matrix for

[1]UCINet is a social network analysis package available at `http://www.analytictech.com/ucinet.htm`

19 actors labeled 'a' through 's'. The graphical representation of the dataset is shown on Figure 2

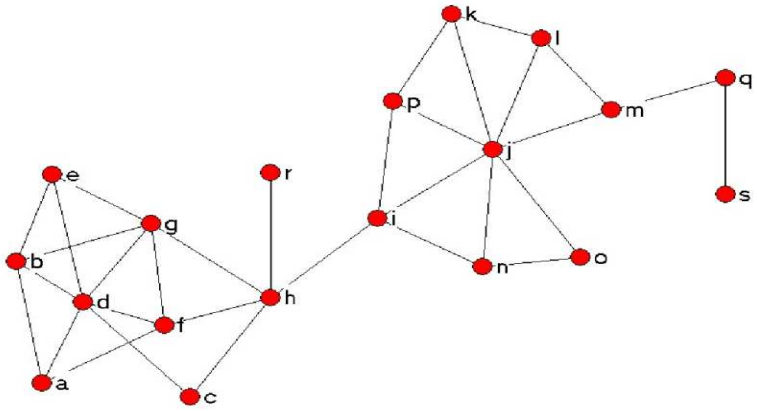


Figure 2: Graphical representation of the borg4cent dataset

For the purposes of this example we suppose that the adjacency matrix A is constructed based on the event dataset E. A single event can be simulated from the adjacency matrix as follows:

1. with probability $p = .5$ pick either
   (a) pair (i, j) from A that has not been picked before or
   (b) pick an actor i uniformly at random and pick an actor j uniformly from the neighbors of i
2. pick actor k from among the common neighbors of the already picked actors or quit with probability .5
3. repeat step (2) until the set of common neighbors is empty

By repeating the above procedure $max(M, |A_{ij} > 0|)$ we obtain an event database E with at least M records. Step 1(a) insures that the sampling algorithm terminates and steps 1 and 2 together guarantee that only the relations described in A are present in the event dataset E. Thus, if we wanted to link

actors explicitly based on their participation in the events in the simulated dataset, we would obtain graph on Figure 2.

In our simulation we generated 500 events as described above. To illustrate the framework we first learned a Bayes Net model using only positive correlations (BN_1). The resulting graph with a subset of the parameters is shown on Figure 3. From the graph on Figure 3 it can be seen that a large number of the relations present in A exhibit strong positive correlation in the data.

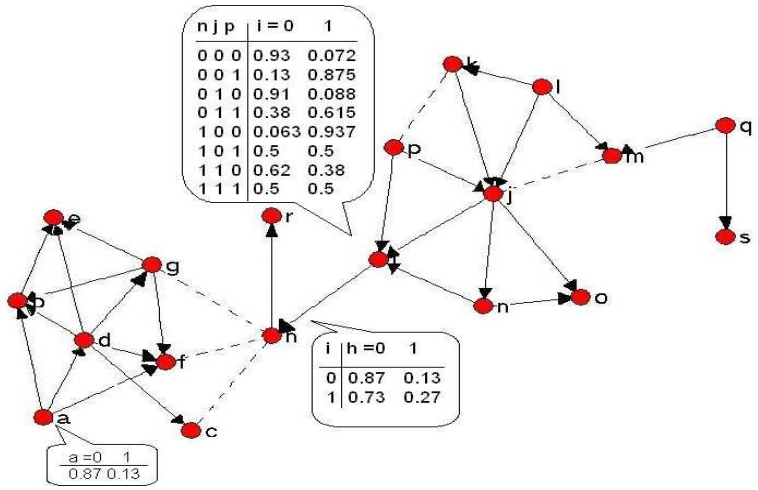


Figure 3: Bayes Net model learned using only positive correlations. Dashed lines represent links that are present in the original adjacency matrix A but are not part of the learned graphical structure

There are several benefits to the generative model that we learn using $SBNS$. Using BN_1 we can immediately answer questions such as what is the probability of actor h participating in an event given that actor i is participating or not participating in that event: $p(h = 1|i = 0) = .13$ and $p(h = 1|i = 1) = .27$. We can also answer other questions that shed light on the relations in the network. For example, we can answer a question "who are the top most likely actors to participate in an event if we know that i is participating". In the case of our data, the answer to that query is n, h, j, p, o, where the order is determined by likelihood, highest first. Note that even though there is no explicit connection between i and o, o turns out to be one of the closer "friends" or "collaborators" in terms of event participation. This information would be very hard to infer just by looking at the adjacency matrix A.

We can also infer negative correlation from data to help improve the accuracy of our inference. In our framework this is done by examining pairs of actors that have never co-occurred but have relatively high mutual information. As can be seen on Figure 4 this leads to a much higher connected BN (BN_2).

Though BN_2 scores significantly higher than BN_1: -4.98 vs -5.62 (average

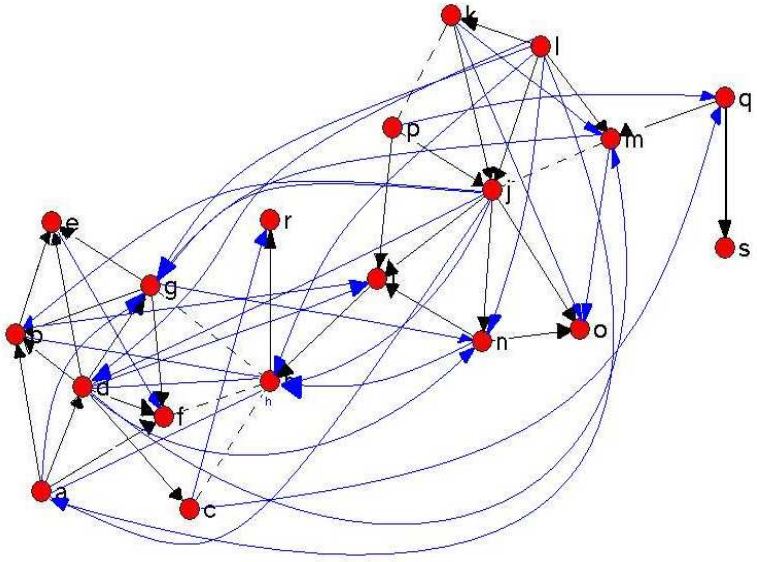


Figure 4: Bayesian Network (BN_2) learned by allowing both positive and negative correlations. The blue edges were learned by using negative correlation information

score on the logscale), it may suffer from overfitting. To check for overfitting, we split the data into two parts. We now re-train the model using two thirds of the data and test on the remaining one third for each of the two models. The results in Table 2 show that the likelihood of test data is slightly worse than training likelihood for both types of models:

	avg train likelihood	avg test likelihood
using positive correlations only	-5.256	-5.42
addition of negative correlations	-5.12417	-5.34815

Table 2: Comparison of BN_1 and BN_2 in terms of overfitting

Table 2 shows that neither of the models grossly overfit though the tendency is to fit training data better due to the noise in the data. Thus, model BN_1 can be preferred to BN_2 as a smaller model that is easier to interpret, whereas BN_2 will give higher accuracy performance when we do inference.

5 Proposed work

5.1 Accuracy improvement

Even though the algorithm is scalable to a very large number of actors, it may benefit from several improvements. Particularly, generating global networks from the local models (step 10–16 in the algorithm in Table 1 can be improved. The edges in local structures learned to be added to the global Bayes Nets are ordered according to their respective number of occurrences. In case of a tie, the ordering is essentially random. It was shown by Friedman and Koller (2003) that ordering is an important factor when building a Bayes Net. One of the solutions to the ordering problem in our case could be a slightly modified version of the variable ordering search algorithm developed in collaboration with Max Chickering (Goldenberg and Chickering, 2004). However, ordering algorithms are computationally expensive. I propose to sample several orders uniformly at random and learn a Bayes Net corresponding to each of them. The number of sampled orders depends strictly on the given time constraint which in the worst case should be equal to picking a single random order. To improve this result, we can bias our sampling in favor of the orders that seem to have resulted in better Bayes Nets so far. We can afford to sample orders due to the small computational cost of learning a Bayes Net given an order. The computational bottleneck remains to be counting of the frequencies.

5.2 Types of relationships and undirected models

As was mentioned above, a substantial body of research in Social Networks explicitly models the properties of ties between actors, such as transitivity and reciprocity (Wasserman and Faust, 1994). However, in real life not all relationships are reciprocal or transitive, thus the properties can be described better using stochastic rather than deterministic models (Snijders, 2004). More recently Hoff et al (2002) have proposed a latent space model where, by estimating parameters (latent positions) empirically from data, such properties will be accounted for automatically.

The SBNS structural search algorithm was not developed with the properties of social relationships in mind. However, some of the well known features of the social net models still hold. For example, in case of reciprocity, in a dyad xy, if x and y are equally likely then $p(y|x) = p(x|y)$ (according to Bayes Rule). In case of transitivity, if x depends on y and y depends on z, then x and y are almost always unconditionally dependent, even though conditionally they could be independent (according to factorizations in the Bayes Net). However, we have to be careful about interpreting the graph. If we tried interpreting the edges as relations between actors directly, in cases like co-authorship, where if x and y have co-authored a paper, it might be desired to have an undirected edge between them. In a Bayes Net semantics is different. The direction of an edge can mean, for example "y is more likely given x", whereas the converse might not be true in the data. Still, it is not clear how much we suffer from using

acyclic graphs to describe distributions underlying social network phenomena.

There are several ways to combat acyclicity in a model. One trivial extension would be to extend SBNS algorithm to learn Dependency Networks as described by Heckerman et al (2000). Dependency networks retain the property that each node in the network is independent of the other nodes given its parents, which allows for computationally efficient estimation of parameters. The extension to learn Dependency Nets using SBNS results in even simpler learning algorithm, where we replace steps $10 - 16$ in the algorithm description in Table 1 by gluing together local structures from the dag storage DS. However, inference is not trivial, since not every Dependency Network encodes parsimonious joint distribution.

It was proved by Heckerman et al (2000) that consistent dependency networks have equivalent representational power as Markov Random Graphs (MRFs). I am interested in extending SBNS to the undirected graphs. The notion of obtaining the global model by identifying a set of locally best models would translate into finding statistically significant cliques. This is easily achievable by finding the best structure for each of the undirected frequent sets by fitting loglinear models using Iterative Proportional Scaling (Bishop et al, 1977). The next step would be to glue the cliques together, however this might result in large cycles, which will then cause a problem when estimating the partition function (normalization constant) and doing inference. Extensive research in the MRFs shows that the computational bottleneck is the estimation of the normalization constant. Several estimation schemes have been proposed (Ghahramani, 2003), but so far no scalable solutions were found.

5.3 Including actors' properties

The work by Hoff et al (2002) has shown that the fit of the model improves if secondary characteristics (e.g. both actors are the same gender) are taken into account. One of the possibilities is to use secondary information when ordering the edges in the SBNS algorithm (step 11 in Table 1). In this case, each actor can be represented as a vector in the space of attributes and in addition to the number of occurrences in different local models, each edge could be weighted by the similarity of the actors participating in it. For example, we could measure similarity by the number of attributes that assume the same value for the two actors. In this case, an edge that does not appear frequently could still outweigh frequent edge if actors of an infrequent dyad are very similar.

The above approach does not take into account actors' properties when evaluating local structures. However, it is easy to imagine a situation where actors' participation in an event is influenced by their properties. For example, if a given event is a weekly meeting of the Artificial Intelligence lab members then it is more likely that students and professors associated with the lab will participate in that meeting, rather than students or professors associated with a Computer Architecture lab. In other words, knowing actors' affiliations can make their presence in an event more or less likely. Thus, we would ideally like to incorporate secondary information about actors into the model selection

process.

I would like to incorporate secondary information regarding actors' characteristics as priors on the observed frequencies of co-occurrences. Following the same notation as in Section 2 and in (Heckerman, 1995), let x_i^k represent the k^{th} possible state of variable X_i and pa_i^j, the j^{th} possible state of parents $\mathbf{Pa_i}$ of the variable X_i. For each variable there are q_i states of its parents. Then, for each state of $\mathbf{Pa_i}$ and given local parameterization θ_i under the model m, $p(x_i^k|pa_i^j, \theta_i)$ is distributed binomially: $p(x_i^k|pa_i^j, \theta_i) = \theta_{ijk}$, where $\sum_{k=1}^{2} \theta_{ijk} = 1$ and $\forall ijk: \quad \theta_{ijk} > 0$. To efficiently calculate the likelihood it is common to assume that θ_{ijk}s are independent and put a Dirichlet prior on the parameters θ_{ij}:

$$p(\theta_{ij}) = \frac{1}{Z(\alpha)} \prod_{k=1}^{r_i} \theta_{ijk}^{\alpha_{ijk}-1} \tag{3}$$

, where normalization constant is $Z(\alpha) = \frac{\prod_{\mathbf{k=1}}^{\mathbf{r_i}} \mathbf{\Gamma}(\alpha_{\mathbf{ijk}})}{\mathbf{\Gamma}(\sum_{\mathbf{k=1}}^{\mathbf{r_i}} \alpha_{\mathbf{ijk}})}$ and parameters α can be interpreted as "prior observation counts". Note that now the structures are not uniformly probable, so the likelihood, i.e. the model selection criteria, has a more general form (Heckerman and Chickering, 1996):

$$p(D|m) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \cdot \prod_{k=1}^{2} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \tag{4}$$

, where N_{ijk} are the observed frequencies of the x_i^k given $\mathbf{pa}_i^j$

There are many ways to calculate prior frequencies from the secondary properties. The simplest way is to empirically estimate α_{ijk} from data by counting. Let $\mathbf{A}$ be a set of all attributes, which is the same for every variable, and $\mathbf{V_a}$ a set of values associated with a particular attribute a. Then the value of the attributes associated with the variable X_i would be $\mathbf{v_{x_i}}$ and the value of the particular attribute of the variable X_i is $v_{a_{x_i}}$. Also, the set of values of a record r_l is indicated as $\mathbf{v_{r_l}} = \{v_{a_{x_i}} : \forall a \in \mathbf{A}, \forall \mathbf{x_i} \in \mathbf{r_l}\}$. To estimate α_{ijk} we need to count the number of records, where the set of values of the attributes of actors include x_i^k and $\mathbf{pa}_i^j$, i.e. $\alpha_{ijk} = \sum_l r_l : \mathbf{v_{r_l}} \supset \{\mathbf{v_{x_i}} \cup \mathbf{v_{pa_i}}\}$. If the number of such records in the dataset is 0, α_{ijk} is set to be 1, since parameters of Dirichlet are restricted to be positive. In other words, we expect to see as many events with the participation of the given actors as there are occurrences of actors with similar combinations of attribute values. This estimation technique of the Dirichlet parameters will increase the likelihood of dependency between actors with parameter values common in the dataset.

5.4 Inference

To answer questions about how strongly individuals are related and to gain better insight into the properties of the relationships modeled by Bayes Nets, we have to do large scale inference quickly. Pavlov and Smyth (2001) have shown that for a query on a subset of attributes it is more efficient to create a local

Bayes Net using only the attributes in the query. The local models in this case outperform inference on global Bayes Nets both in efficiency and in accuracy. However, some queries of interest are more global. For example, we might like to know who are the people that have the most effect on a particular actor. In this case, each actor and/or combination of actors could be potential answers to this query. In a graph with $100,000$ or more actors, the most efficient inference techniques applied to the global Bayes Net are computationally infeasible. However, there are assumptions that can be made to speed up the computation. Obviously, no matter how large the graph is, only a handful of actors could be an answer to the query especially if we set the threshold for "significant effect" apriori. Most of the actors in the graph have marginally negligible influence on any given actor. I plan to explore this and other insights and build on some of the ideas proposed by Pavlov et al (2003) in order to create a general framework for answering queries that are of interest to social scientists using a global Bayes Net.

5.5 Evolution of Social Networks

So far the proposed changes to the model have been static. However, in real data, especially if it is collected over a long period of time, the relationships between actors change. For example, in the publication index database, many of the publications were co-authored by students and their professors. The relationships and publication patterns tend to change as the students graduate. The model is likely to change significantly in the period of ten years. I would like to extend my framework to account for the temporal changes in the network. Since Frequent Sets are tuples that occur at least s times, it is conceivable to assume that the changes can be captured locally. For example, when a student graduates, the frequency of his/her collaboration with the professor changes. If the change is significant then it will be reflected in all the tuples that contain the student and the professor. By monitoring and comparing the global frequencies of each tuple at time $t+1$ to frequencies at time t, it is possible to pinpoint the local models that will need to be re-estimated. Due to the decomposable property of the global model, it is possible to localize the re-estimation of the net and thus incorporate the information about the changes over time in the computationally efficient manner.

6 Timeline

The following is the plan to proceed with my research. The tasks are not necessarily in the order they will be addressed, since some of them are interrelated. I am budgeting about 25% of each task for evaluation and testing on large real world datasets.

1. improve SBNS search by modifying the heuristic used to create global Bayes Nets from local ones (Section 5.1) - 2 months

2. explore in more depth properties of the models found by SBNS and their relation to Social Networks (Section 5.2) - 2 months

3. study algorithms for inference tasks (Section 5.4) - 4 months

4. implement proposed incorporation of the secondary characteristics (Section 5.3) - 2 months

5. apply variational methods to aid with porting the SBNS idea to the undirected graphs (Section 5.2) - 2 months

6. obtain a fitness criteria for undirected graph model selection (Section 5.2) - 3 months

7. evolution of the graph structure over time, possible significant modification of the model, re-using the same ideas for computational feasibility (Section 5.5) - 5 months

I am hoping to make some advances on each of the above tasks, but in reality one or maybe two tasks may be dropped if they run into serious limitations. I'm planning to finish proposed research in the next 18-20 months and graduate in August of 2006.

References

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM SIGMOD 12* (pp. 207–216).

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *VLDB 20* (pp. 487–499).

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

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.

Berg, J., & Lassig, M. (2004). Local graph alignment and motif search in biological networks. *Proceedings of the National Academy of Science*, *101*, 14689–14694.

Bishop, Y., Fienberg, S., & Holland, P. (1977). *Discrete multivariate analysis: Theory and practice*. MIT Press.

Breese, J., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. *UAI 14*.

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

Cooper, G., & Herskovits, E. (1991). A Bayesian method for constructing Bayesian belief network from databases. *UAI 7* (pp. 86–94).

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

Erdos, P., & Reny, A. (1959). On random graphs. *Publicationes Mathematicae*, *6*, 290–297.

Frank, O., & Strauss, D. (1984). Markov graphs. *Journal of the American Statistical Association*, *81*, 832–842.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*.

Friedman, N., & Koller, D. (2003). Being Bayesian about network structure: A Bayesian approach to structure discovery in Bayesian networks. *Machine Learning*, *50*, 95–125.

Getoor, L., Friedman, N., Koller, D., & Taskar, B. (2002). Learning probabilistic models of link structure. *Journal of Machine Learning Research*.

Ghahramani, Z. (2003). Bayesian learning in undirected graphical models. Talk given at the Machine Learning lunch seminar at CMU.

Goldenberg, A., & Chickering, M. (2004). Learning bayesian networks by sampling variable orders. Submitted to AISTATS 2005.

Goldenberg, A., & Moore, A. (2004). Tractable learning of large bayes net structures from sparse data. 21^{st} *International Conference on Machine Learning*.

Han, J., & Kamber, M. (2000). *Data mining: Concepts and techniques*. Morgan Kaufmann Publishers.

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

Heckerman, D., Chickering, D., Meek, C., Rounthwaite, R., & Kadie, C. (2000). Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, *1*, 49–75.

Heckerman, D., Geiger, D., & Chickering, D. (1995). Learning Bayesian Networks: The combination of knowledge and statistical data. *JMLR*, *20*, 197–243.

Heckerman, D., Meek, C., & Koller, D. (2004). Probabilistic entity-relationship models, prms, and plate models. *Proceedings of the 21st International Conference on Machine Learning*.

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.

Jensen, J. N. D. (2003). Collective classification with relational dependency networks. *9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Proceedings of the 2nd Multi-Relational Data Mining Workshop*.

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*.

McGovern, A., Friedland, L., Hay, M., Gallagher, B., Fast, A., Neville, J., & Jensen, D. (2003). Exploiting relational structure to understand publication patterns in high-energy physics. *SIGKDD Explorations* (pp. 165–173).

Neville, J., Jensen, D., Friedland, L., & Hay, M. (2003). Learning relational probability trees. *In Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

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

Pavlov, D., Mannila, H., & Smyth, P. (2003). Beyond independence: probabilistic models for query approximation on binary transaction data. *IEEE Transactions on Knowledge and Data Engineering*.

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.