Bayes Net Graphs to Understand Co-authorship Networks?

Anna Goldenberg
Center for Automated Learning and Discovery
Carnegie Mellon University
Pittsburgh, PA 15213, USA
anya@cs.cmu.edu

Andrew W. Moore
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
awm@cs.cmu.edu

ABSTRACT

Improvements in data collection and the birth of online communities made it possible to obtain very large social networks (graphs). Several communities have been involved in modeling and analyzing these graphs. Usage of graphical models, such as Bayesian Networks (BN), to analyze massive data has become increasingly popular, due to their scalability and robustness to noise. In the literature BNs are primarily used for compact representation of joint distributions and to perform inference, i.e. answer queries about the data. In this work we learn Bayes Nets using the previously proposed SBNS algorithm [14]. We look at the learned networks for the purpose of analyzing the graph structure itself. We also point out a few improvements over the SBNS algorithm. The usefulness of Bayes Net structures to understand social networks is an open area. We discuss possible interpretations using a small subgraph of the Medline publications and hope to provoke some discussion and interest in further analysis.

Keywords

Bayesian Networks, Structural Learning, Massive Data, Graph Analysis, Co-authorship networks

1. INTRODUCTION

The statistical literature on modeling Social Networks assumes that there are n entities called *actors* and that there exists information about binary relations between them. Binary relations are represented as an $n \times n$ 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 a 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 real valued and not just binary, representing the strength of the relationship between actors i and j [27]. 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 fully observed covariate data that is taken into account in the model [19].

In our work, we assume that there are observations, particularly *events* relating entities (each *paper* is an event in the co-authorship dataset). However, the true underlying structure of relations between entities is not observed. We are not claiming to find the true underlying graph connecting the entities. By probabilistically modeling dependencies from the events data we aim to learn the relations robust to noise, the dependency structure and in the future predict the entities' further actions (using inference). In other words, based on the known information about simultaneous participation of entities in observed events, we construct a probabilistic model that would describe those events.

Studies on gene expression data [12] and social networks in particular suggest that correlations of entities on a local level are very important and in fact are what make up the global network [12, 7]. The SBNS algorithm [14] used here to learn the structure of Bayes Nets precisely makes use of that idea. 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. So, along with being computationally practical, Bayesian Networks created by our algorithm have a very natural motivation stemming from those important domains.

In this work we turn our attention to the question - how valuable are the graph structures of the Bayes Nets themselves? The resulting learned structures look like directed social networks, but the semantics behind links are different and one needs to be careful interpreting the results. In our experiments section we show two subgraphs built for the same authors of the Medline dataset that exhibit different characteristics. We observe that by learning probabilistic models we are able to draw conclusions that were not possible by simply connecting co-authors together. In fact, we believe that probabilistic graphical models that learn dependencies between entities provide a very rich structure for analysis. This work is just the beginning of the exploration in this area.

This paper is structured as follows. First we introduce notation and concepts essential to understanding the SBNS algorithm. We then provide a shorter more intuitive description of the SBNS indicating improved heuristics where applicable. Further, we give an example of a possible interpretation of the Bayes Nets in terms of the social relations in the co-authorship publications. Finally we discuss related

LinkKDD '05 August 21, 2005, Chicago, IL, USA

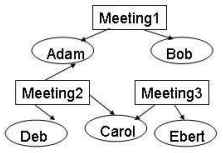
literature and conclude with our thoughts on future work.

2. BACKGROUND

In this section we introduce the terms and concepts that are most relevant to our learning algorithm. It has to be noted that the proposed algorithm readily applies only to binary, otherwise known as *transactional*, data. Thus, first we would like to introduce the general scenario where the algorithm can be applied.

2.1 Data

Assume our training data is a collection of M records of observations of N binary variables $X_1, ...X_N$. Write x_i^j as the value of X_i in the jth record where $1 \leq i \leq N$ and $1 \leq j \leq M$. Intuitively, each record denotes a collection of entities that participated in an "event". We use the words entity and *actor*, as in "social actor", interchangeably throughout the paper. 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 k have co-authored a paper together, then for this event (co-authorship of a given paper) their states are $X_i = 1$ and $X_k = 1$ and the states of all other variables in the database for this event are 0 ($X_t = 0, \forall t \neq i, k$). Examples of co-authorship datasets are the online library of computer science publications Citeseer, the index of online library of medical publications Medline and others, where each record is a list of co-authors of a particular paper.



	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.

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

2.2 Frequent Sets

Let the N variables be represented by integers $\{1, 2, \dots N\}$. Let the *co-occurrence frequency* of a set of attributes $S \subseteq \{1, 2, \dots M\}$ be the number of records in which all the attributes in S are simultaneously set to 1.

$$\text{freq(S)} = |\{i : \forall j \in S, x_{ij} = 1\}| \quad (1)$$

Given $s \geq 1$ we say S is a *Frequent Set of m attributes* if S contains exactly m attributes and $freq(S) \geq s$. Threshold s is called *support* in the data mining literature. Given sparse data and a support s greater than about 3, it is surprisingly easy to compute all Frequent Sets [2]. There is an abundance of literature on Frequent Sets as their collection is an essential part of the association rules algorithms [1, 2, 15] widely used in commercial data mining.

2.3 Bayes Nets

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 likelihood penalized for complexity. BNs are factored probabilistic graphical models, where the joint distribution is determined by a product of conditional probabilities, i.e.

$$P(X_1 \dots X_n) = \prod_i P(X_i | Pa(X_i)) \quad (2)$$

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 is a coherent probability distribution. More information on Bayesian Networks can be found in [10, 17].

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

3. SBNS ALGORITHM

Here we give a brief description of the SBNS (Screen-based Bayes Net Structure search) and for details we refer the reader to [14].

The SBNS algorithm is a two stage process. During the first stage, which we will call *Local Screening*, SBNS performs a Bayes Net structural search on each of the small subsets of variables defined by Frequent Sets. The resulting local structures comprise the restrictive pool of edges from which the global Bayes Net will be constructed at the second stage.

3.1 Local Screening

The intuitive idea behind the local search stage is that we do a full structural search on very small subsets of variables. One of the ways to identify the (not necessarily disjoint) subsets is to use Frequent Sets.

Screening the Frequent Sets. Suppose we have a collection of Frequent Sets $\{X : |X| = m, m \geq 2\}$. We call *screening* the process of finding the optimal Bayes Net structure for each of the Frequent Sets.

First, we screen the pairs to find pairwise correlations. We add an edge between two variables to the *Edgedump* if and only if a model that has an edge in either direction was found to have a higher score than a complete independence model between the two variables in the pair. We then in turn screen Frequent Sets of size 3, 4, etc.

It is possible that the dependencies in the Frequent Set S of size $m > 2$ are already well-explained by interactions of

order less than m. For example, suppose variables X_i, X_j and X_k co-occur frequently, but their co-occurence is well explained by the local Bayesian Network DAG structure of $X_i \leftarrow X_j \rightarrow X_k$. In that case when searching through pairs, (X_i, X_j), (X_j, X_k), (X_i, X_k), the two-way interactions will already explain all dependencies of S. In fact, only DAGs that contain a node with $m-1$ parents could be missed by not considering an m-size tuple.

We implement a *Screening* test by searching over all possible DAG structures for S and finding whether the best scoring structure has an $m-1$-parent node (we call it an *m-way* interaction). We thus allow S to pass the screening test *if and only if* S is best explained by a DAG structure containing an m-way interaction. If S passes the Screening test, all edges of the highest scoring DAG are added to the *Edgedump* – the set of edges that will eventually be considered for addition to the global Bayes Net.

3.2 Stage 2: Global Bayes Net

Once the Edgedump is created, there are several ways to construct the global Bayes Net. In this work we use the following heuristic: prioritize the edges by the score of the highest scoring m-way interaction in which they participated; create the global Bayes Net by adding the highest correlated variables first. Not all edges in the Edgedump will be added due to the acyclicity property of BNs. Note that this approach is different from the heuristic originally proposed in [14] and seemed to have resulted in higher scoring Bayes Nets in practice. We start with an empty (edgeless) global Bayesian Network and iterate through the ordered contents of the Edgedump, allowing each edge in turn to be added if and only if it improves the current score and avoids cycles. If the algorithm fails to add an edge with the direction stored in the edgedump, it tries to add the reversed edge to avoid cycles.

The proposed deterministic approach for creating a global Bayes Net is fast and performs better on average than if the edges were added randomly. However it is a simple heuristic that imposes an ordering on the variables that is not necessarily optimal.

4. NEGATIVE CORRELATION

In the previous section we pointed out that Frequent Sets allow the algorithm to consider only interactions that cause co-occurrence (and thus most likely *positive* correlations). Due to the sparse nature of the data we are not omitting the strongest correlations in general. There is, however, still a danger that if a few variables have relatively high univariate marginal probability, they could cause significant negative correlations that we would miss. Fortunately, such negative pairwise correlations can be detected cheaply by looking at a fraction of the pairs that have never occured together. We reduce the total number of entities significantly by only considering ones that occurred more than support s times in the dataset. This step is statistically justified because fewer occurrences mean lower possible mutual information. We then look at the pairs starting with the highest frequencies first.

There are two possibilities for introducing the negatively correlated pairs. One is to introduce the edges to the Edgedump from which the DAG will be constructed. Another possibility is to augment the DAG created from positive correlations. Each of the approaches has its own biases.

When we decide whether to add an edge between possibly negatively correlated variables X and Y to the Edgedump before the DAG is created, we compare the scores of the model $X-Y$ vs $X \quad Y$ and add an edge if the former scores higher (note: the direction of the edge does not matter if the scoring metric is structurally equivalent). This approach has the disadvantage of not taking into account other dependencies that may already be modeled by the existing edges in the Edgedump. It also might result in considering too many edges. However, the advantage of this approach is that when building a DAG, the pool of dependencies is more complete.

The second approach is to add edges between negatively correlated variables to the constructed DAG. In this case, we add an edge only if it does not cause a cycle and improves the score. Notice that neither of these conditions exist prior to building the DAG and are thus impossible to verify in the alternative approach describe above. The advantage of this approach is that we are likely to consider fewer pairs and thus it may be more appealing for larger networks.

5. EVALUATION CRITERION

There are several standard Bayesian scoring functions that are often used in the literature to evaluate structural learning algorithms. The structures learned were evaluated based on one of the most often used: BDEu, with an equivalent sample size of 1. The general form for the BDeu scoring function is presented in Equation 3. The BDe score was originally suggested by [8]. The u in BDeu just means a uniform prior over structures. The different scoring metrics are described in detail by [17].

$$S(G,D) = \log\left(\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\frac{1}{q_i})}{\Gamma(\frac{1}{q_i}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(\frac{1}{q_i r_i}+N_{ijk})}{\Gamma(\frac{1}{q_i r_i})}\right) \quad (3)$$

where i is the ith variable, q_i - the number of states of the parents of x_i and r_i - the two states (true/false) of x_i, in our case of binary variables. Thus N_{ijk} is the number of records in our data where $X_i = k$ and $Pa(X_i)$ are in the j^{th} state.

5.1 Datasets

We have applied the Bayes Net structural learning algorithm to several co-authorship datasets (sizes are in Table 1).

Table 1: Datasets and their sizes

Datasets	Entities	Records
Institute	456	1488
NIPS	2037	1740
Medline	19499	6217
Citeseer	104801	180395

1. The *Institute Data* is a set of records of collaborations between professors and students collected from publicly available web pages listed on the Carnegie Mellon University Robotic Institute's web site.

2. The *NIPS Data Set* contains co-authorship information of the Neural Information Processing Systems con-

ference (NIPS) contained in proceedings 1-12, the pre-electronic submission era [1].

3. The *Medline Data* is a sample of the co-authorship information of the publically available medical publication database Medline.

4. The *Citeseer Data* is a set of co-publication records from the Citeseer online library and index of computer science publications. Since the entities are represented by first initial and last name, a single name might correspond to several people.

One of the key reasons why the algorithm we propose is computationally feasible is the natural tendency of large social networks to be very sparse. In other words, most of the authors tend to co-author papers with only a handful of the others considered, while very few authors co-author with a large number of others. This effect of social nets has been extensively discussed in the social networks literature [3]. Since the co-authorship data can be interpreted as a bi-partite graph, where one type of nodes are authors and the other is publications, it is interesting to note that the number of people per publication also exhibits a Power Law property: there are few publications that have many authors and majority of publications have just a few authors. In Figure 2 we provide the frequency plots for each of the datasets for both papers-per-author and authors-per-paper frequency distributions. From the plots it is apparent that co-authorship data is indeed distributed similarly to a Power Law, though some datasets tend to be particularly sparse (Medline) and some datasets tend to have heavier tails (Institute).

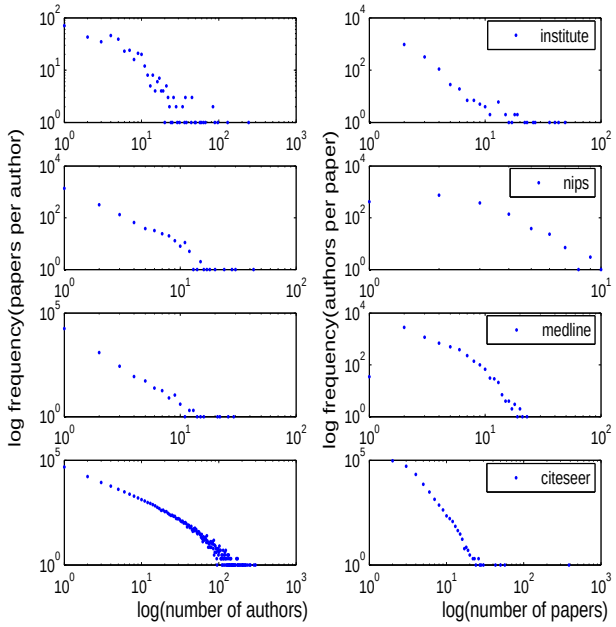


Figure 2: Marginal Frequency distribution plots

[1]This dataset was made available by Sam Roweis and can be downloaded from `http://www.cs.toronto.edu/~roweis/data.html`

5.2 Network interpretation

Usually in social science [19, 30] it is assumed that the connections are given, for example if two people have co-authored a publication they are connected by an edge in the graph. In this case the connectivity in the graph can be easily interpreted in terms of original data and the research in these fields focuses mostly on modeling the generative mechanisms and understanding the global properties of the graph. Even though the networks that we learn using the SBNS algorithm are represented as graphs connecting the same nodes, they have different semantics and one should be careful when interpreting them. We claim however that the graph structure of the learned Bayes Nets can be used effectively to gain a different view of the relations between entities (actors, people, authors).

What does the presence/absence of a link in the graph structure of the learned Bayes Net mean? First of all, the presence of a directed edge $X \rightarrow Y$ means that if the author X is known to be one of the co-authors of a paper, we can infer something about the presence of Y. By further inspection of the corresponding conditional probability table (CPT), we can say whether Y is more or less likely to be an author if X is already an author. This is a standard Bayes Net analysis. It is interesting to note, that many edges in a Bayes Net correspond to the edges in the social network, i.e. some of the edges in the social network represent significant statistical dependence between the authors. Also, due to the fact that SBNS models negative correlations as well, we can gain additional information into the set of relations that normally cannot be inferred from the social networks. For example, two doctors (from the Medline database) never co-author a paper together, but co-author quite often by themselves or with others. Knowing a few of the "negative relations" might help the network analysts to discover polarity in opinions of the corresponding doctors.

5.3 Example

To illustrate how Bayes Nets help to improve understanding about relations among doctors, we give an example of analyzing connections of a random author from the Medline publication dataset. The part of the network shown is obtained by learning the Bayes Net only on the publications that had the key word "overactive bladder", the support was set to 1 and the maximum tuple size was 3. The number of authors were consequently $16,380$ and the number of corresponding publications is $7,575$. SBNS took 1 second to learn the network. Figure 3 represents relations of the 3 levels of predecessors and successors of Alan J Wein in the learned Bayes Net.

From the part of the corresponding probability table shown in Table 2 it is evident that the presence of *Christopher R Chapple* is negatively correlated with the target *Alan J Wein* and that the presence of *Eric S Rovner* by himself is not as strong evidence for the presence of *Alan J Wein* as the presence of both *Eric S Rovner and Flavio E Trigo-Rocha*.

We also provide a social network graph where each link means co-authorship also starting with *Alan J Wein* as the main actor. We limit ourselves in this case to just people that have co-authored with Alan J Wein directly since the network grows very fast. Each link has a weight which represents how many publications the pair have appeared on as co-authors. The graph presented on Figure 5.3 appears much more interconnected with a few fully connected

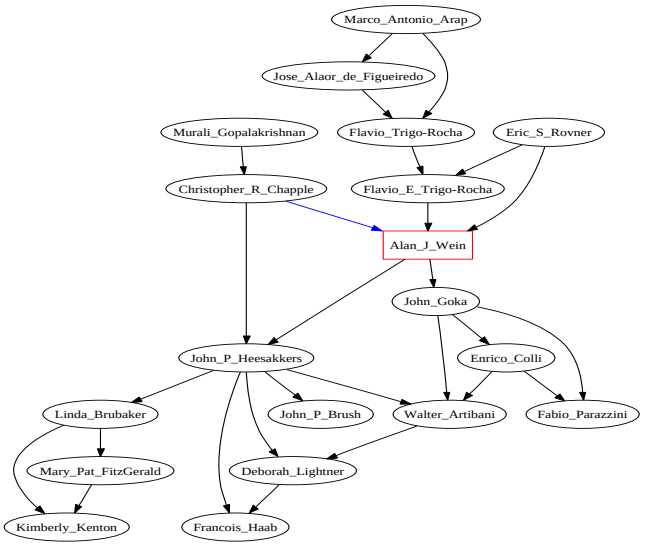


Figure 3: A part of a Bayesian Network learned from Medline publications with the keyword "overactive bladder"

			Alan J Wein	
Christopher R Chapple	Eric S Rovner	Flavio E Trigo-Rocha	0	1
0	0	0	0.997	0.003
0	1	0	0.46	0.54
0	1	1	0.33	0.67
1	0	0	0.75	0.25

Table 2: Part of a Conditional Probability Table (CPT) for *Alan J Wein* from the Bayes Net learned using SBNS

cliques. There are also several people that were not appearing in our Bayes Net. Note that the links with weights higher than 1 appear in the Bayes Net. Most links in the presented Social Network however have a weight of 1, meaning that there is not enough evidence to claim a strong dependency between co-authors. Thus, given the same data, even without increasing the support parameter, our Bayes Net learning algorithm is able to bring more clarity into the picture of relations.

5.4 Dangers in interpretation of the Bayes Net

There are certain things one should keep in mind when interpreting the Bayes Net graphs. Here we list three issues that one must be aware of, but the list might not be complete.

1. If the two nodes are not linked, it doesn't mean they are independent. It means that they are conditionally independent given their parents. Thus one must not ignore the structure of the graph when reasoning about any two nodes.

2. Proximity and number of hops in the network may not necessarily translate into the strength of a relationship as might be done in social networks. For example, in the case of the two small subgraphs presented here, from the Social Network on Figure 5.3 we see that *Deborah Lightner* has co-authored with *Alan J Wein* once and *John P Heessakkers* also co-authored with *Alan J Wein* once. In our Bayes Net on Figure 3 however *Deborah* depends on *Alan* through *John* and another parent. This does not translate into *Deborah is less likely to co-author with Alan than with John*, however it does tell us that if we know that *John* was one of the authors, knowing about *Alan* will not affect our belief in *Deborah's* presence as a co-author.

3. Our networks do not necessarily imply the causality which is usually associated with Bayes Nets. Causality needs to be tested by perturbing the evidence and seeing whether the outcome changes. We do not perform any such tests and thus in general we cannot say that the presence of X causes Y to be present, we can state however that the presence of X makes Ys presence more likely and vice versa, if that is what our conditional probability tables tell us.

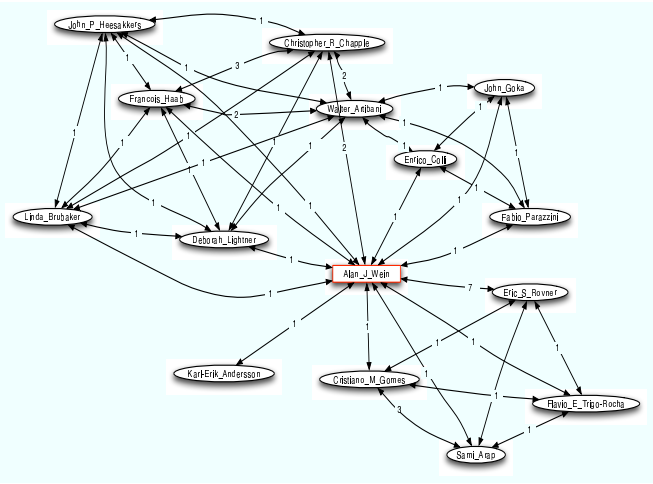


Figure 4: A part of a social network learned from Medline publications with the keyword "overactive bladder" where each link represents co-authorships and weights represent the number of co-authored publications

5.5 Global graph properties of the Bayes Nets

In terms of the global properties of the graph, we also show the graph of degree distributions for the global social network for the overactive bladder and the indegree and outdegree of the learned Bayes Nets in Figure 5.5. The Bayes Net structure seems to follow a Power Law as well. The top indegree nodes do not correspond to the top outdegree nodes. From the graph we can see that there are a few nodes with higher outdegree than the number of publications per person in the data (the social network degree distribution corresponds to precisely that). This is caused by a few of the negative correlations added, i.e. the doctors who are popular (having a high number of publications with other authors) tend to have extra edges corresponding to doctors with high number of publications whom they have never co-authored with.

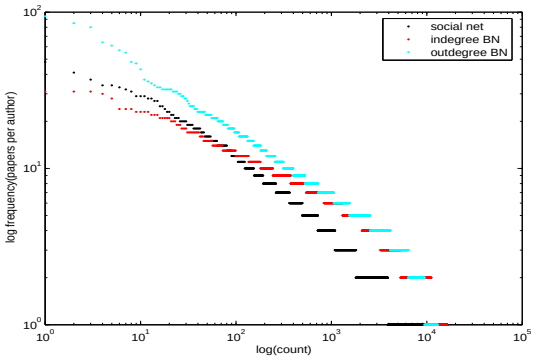


Figure 5: Degree distributions for the social network and the indegree and outdegree for the learned Bayes Net for the publications from the Medline data with the keyword "overactive bladder"

5.6 Maximum Frequent Set Size

In our experiments we tried different maximum Frequent Set sizes: ($mfss = 2 \ldots 5$). The lower bound $mfss = 2$ means that we consider only pairs of entities and thus the structure learned is based solely on two-way marginal counts. Our experiments on large datasets such as Citeseer have shown that there is an obvious loss in accuracy when high order interactions are not taken into account. Beyond a maximum Frequent Set size of 4 the number of Frequent Sets does not increase substantially in these datasets and hence the behavior of $SBNS$ changes little.

We have to note here, that there is a natural upper bound on the maximum tuple size due to the sparsity of the datasets. For example, there are $94,016$ publications in the Citeseer database that have 2 authors and only $3,022$ that have exactly 6 authors. The potential number of publications that have 6 authors, given the total number of authors in the database is 1.8×10^{27}, so the empirical number is only $(1.6 \times 10^{-22})\%$ of the total. Hence, we cannot expect a great improvement in the score of the Bayes Net when increasing the maximum tuple size, since there is not enough support for larger tuples.

5.7 Support

Lowering support greatly increases the number of Frequent Sets to be considered during screening. However, it also introduces quite a few interactions between variables that have low marginal counts. Model fitting in contingency tables in general is sensitive to very low marginal counts even if they are not zero [6]. Here we use BDeu, which is less sensitive to low counts. Despite this, it seems to be a good idea to keep support relatively large in the case of very large datasets. We have tested several support sizes on smaller datasets and found that on very sparse datasets we can use support $s = 1$ without significant overfitting. However for large datasets such as citeseer we used support $s = 3$ to reduce computational cost without affecting the BDeu score too much. We also have to note that if $s = 1$ support is used, we cannot use the approach of adding negative correlations before constructing the DAG, this approach becomes too costly. The addition of negative correlation after the construction of the DAG has showen to improve the score, while keeping the computation costs low.

5.8 Other Datasets

We have tried our algorithm on a variety of other datasets, for example IMDB (the Internet Movie DataBase) and IOBDB (the Off-Broadway shows DataBase). These datasets exhibit different properties than the publication data since it is more typical of plays and movies to have many actors. Thus, the distribution of the entities per event is different. The SBNS algorithm learns Bayes Nets that fit the data better (score higher) than the networks found by random hillclimbing. SBNS however is more time consuming since on average the data is somewhat less sparse. We are planning to do the graph analysis of these domains in the future.

6. RELATED WORK

Using Frequent Sets when learning Bayes Nets on the local scale was also explored in [25]. The goal of this work was to answer probabilistic queries on a subset of variables, thus there was no need to combine local information to obtain the joint distribution once the query size was estimated. The authors have explored Frequent Sets for quick computations of the CPTs and have noted that it is enough to look at all pairs to compute the triples without having to scan the dataset directly. The performance of Bayes Nets learned from a selection of variables was reported to be worse though close in accuracy to the inferences drawn from a Bayes Net learned on a full dataset. In [20] it has been proposed to integrate Frequent Sets as a local methodology when modelling joint distributions. This work has shown that mixture models obtained from Frequent Sets using maximum entropy are more accurate, thus supporting our claim that frequent sets contain important local information when modelling joint distributions.

One approach to speed up structural search in Bayes Nets for massive datasets has been to restrict the possible parents. The full Sparse Candidate Algorithm is presented in [13]. In its original form it is a method to speed up hillclimbing at the cost of lower performance, though in practice the performance loss was shown to be insignificant for some of the small datasets. This work is yet another motivation for us, since structural search on the local scale inadvertently restricts the number of parents. However, since on the global scale the number of parents in our Bayesian Network is not limited we perceive it as an improvement on the original Sparse Candidate algorithm.

The idea of augmenting Bayes Nets with high mutual information edges is based on the fact that such dependencies could not be accounted for in frequent sets. The fast computation used in this work is based on [22].

6.1 Statistical Network Modeling

The social network literature focuses predominantly 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 the graph-specific properties that are being modeled can be found in [30]. Thus, the models are geared to probabilistically explains the patterns of observed links and their absence between N given entities.

Several useful properties of stochastic models are listed in a brief survey work [28]. 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 [9]
- General approaches for parameter estimation and model comparison using Markov Chain Monte Carlo methods (e.g. [29])
- Taking into account individual variability [18] and properties (covariates) of actors [19]
- An ability to handle groups of nodes with equivalent statistical properties [31].

There are several problems with existing models such as degeneracy, analyzed by [16], and scalability, mentioned by several sources [19, 28]. The new specifications for the Exponential Random Graph Models proposed in [30] attempt to find a solution for unstable likelihoods by proposing a slightly different parametrization of the models than used previously. 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. 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 for Markov Random Fields is well-known to be intractable in general 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 [19]. 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 [19] 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. The probability of a link is a probabilistic function of the positions and actors' covariates. The latent positions are estimated from data using logistic regression. The general form of the model is:

$$logodds(y_{ij} = 1|z_i, z_j, x_{ij}, \alpha, \beta) = \alpha + \beta^T x_{ij} + d(z_i, z_j) \quad (4)$$

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

6.1.1 Network Modeling in Physics

The graph theoretic area of physics that studies complex systems is directly applicable to social network modeling. Though modeling of complex systems has developed seemingly in parallel to statistical modeling of social networks in social science, the findings in this area help to understand further the phenomenon of real networks organization and structure. The assumptions are the same: there are N actors (nodes) and there are M links between those nodes representing relationships among actors. The goal is also to understand and model structural properties of naturally occurring networks. The base model describing random graphs was developed by [11], where the 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 random graphs does not follow power law $P(k) \sim k^{-\gamma}$ common in realistic networks. Thus "scale-free networks" were introduced [4, 5]. [24] have developed a generalized random graph model where the degree distribution is given as an input parameter. Research in the field of physics gives more insight into graph growth, clusterability, graph diameter and the formation of a large component. A good summary of past and ongoing work and its relation to statistical physics is given in [3].

7. CONCLUSION

Recent work has made it computationally possible to learn Bayesian Networks from very large datasets. One of the areas where such models could be of use is social science. In particular, in this work we focus on the connection between Bayes Nets and social networks and illustrate potential interpretations of the graphical structure learned. Our simple example shows that Bayes Nets, while providing a compact representation, are a potential source for much deeper understanding of the data, such as learning about negative interactions among actors. This work is just the beginning of exploratory analysis using Bayesian Networks to model the structure of social networks themselves. We are currently collaborating with our colleagues at Pfizer to gain deeper understanding into the usefulness of this representation.

8. ACKNOWLEDGEMENTS

We would like to thank Ira Haimowitz and Pinaki Karr from Pfizer for helping with data and model interpretations. We would also like to thank Jens Nielsen and Ricardo Silva for insightful discussions.

9. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD 12*, pages 207–216, 26–28 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *VLDB 20*, pages 487–499, 12–15 1994.

[3] R. Albert and A.-L. Barabasi. Statistical mechanics of social networks. *Reviews of Modern Physics*, 74, 2002.

[4] A.-L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[5] A.-L. Barabasi, R. Albert, and H. Jeong. Mean-field theory for scale-free random networks. *Physica A*, 272(1-2):173–187, 1999.

[6] Y. Bishop, S. Fienberg, and P. Holland. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, 1977.

[7] R. Breiger. Emergent themes in social network analysis: Results, challenges, opportunities. In *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, 2003.

[8] W. Buntine. Theory refinement on Bayesian networks. In *UAI 7*, pages 52–60, 1991.

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

[10] G. Cooper and E. Herskovits. A Bayesian method for constructing Bayesian belief network from databases. In *UAI 7*, pages 86–94, 1991.

[11] P. Erdos and A. Reny. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.

[12] N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 2004.

[13] N. Friedman, I. Nachman, and D. Pe'er. Learning bayes network structure from massive datasets: The "sparse candidate" algorithm. In *UAI 15*, page 206:215, 1999.

[14] A. Goldenberg and A. Moore. Tractable learning of large bayes net structures from sparse data. In 21^{st} *International Conference on Machine Learning*, 2004.

[15] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, August 2000.

[16] M. Handcock. Assessing degeneracy in statistical models of social networks. Working Paper 39, University of Washington, 2003.

[17] D. Heckerman, D. Geiger, and D. Chickering. Learning Bayesian Netowrks: The combination of konwledge and statistical data. *JMLR*, 20:197–243, 1995.

[18] P. Hoff. Random effects models for network data. *Proceedings of the National Academy of Sciences*, 2003.

[19] P. Hoff, A. Raftery, and M. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97:1090–1098, 2002.

[20] J. Hollmen, J. Seppanen, and H. Mannila. Mixture models and frequent sets: combining global and local methods for 0-1 data. In *SIAM ICDM*, May 2003.

[21] G. Hulten and P. Domingos. Mining complex models from arbitrarily large databases in constant time. In *ACM SIGKDD 8*, pages 525–531, 2002.

[22] M. Meila. An accelerated Chow and Liu algorithm: fitting tree distributions to high dimensional sparse data. Technical Report AIM-1652, MIT, 1999.

[23] J. Moreno and H. Jennings. Statistics of social configuration. *Sociometry*, (1):342–374, 1938.

[24] M. Newman. The structure of scientific collaboration networks. In *Proceedings of the National Academy of Sciences USA*, volume 98, pages 404–409, 2001.

[25] D. Pavlov, H. Mannila, and P. Smyth. Beyond independence: probabilistic models for query approximation on binary transaction data. In *IEEE Transactions on Knowledge and Data Engineering*, September 2003.

[26] D. Pelleg and A. Moore. Using tarjan's red rule for fast dependency tree construction. In *NIPS 15*, 2002.

[27] G. Robins, P. Pattison, and S. Wasserman. Logit models and logistic regressions for social networks iii. valued relations. *Psychometrika*, 64(3):371–394, 1999.

[28] P. Smyth. Statistical modeling of graph and network data. In *IJCAI Workshop on Learning Statistical Models from Relational Data*, 2003.

[29] T. Snijders. Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2), 2002.

[30] T. Snijders, P. Pattison, G. Robins, and M. Handcock. New specifications for exponential random graph models. Submitted for publication, 2004.

[31] Y. Wang and G. Wong. Stochastic blockmodels for directed graphs. *Journal American Statistical Association*, 82:8–19, 1987.