Pruning Social Networks Using Structural Properties and Descriptive Attributes

Lisa Singh
Computer Science Dept.
Georgetown University
Washington, DC, USA
singh@cs.georgetown.edu

Lise Getoor
Computer Science Dept.
University of Maryland
College Park, MD, USA
getoor@cs.umd.edu

Louis Licamele
Computer Science Dept.
University of Maryland
College Park, MD, USA
licamele@cs.umd.edu

Abstract

Scale is often an issue with understanding and making sense of large social networks. Here we investigate methods for pruning social networks by determining the most relevant relationships. We measure importance in terms of predictive accuracy on a set of target attributes of the social network. Our goal is to create a pruned network that models only the most informative affiliations and relationships. We present methods for pruning networks based on both structural properties and descriptive attributes demonstrate it on a network of NASDAQ and NYSE businesses and on a bibliographic network.

1 Introduction

Social networks abound; examples include online community networks, disease transmission networks, corporate executive networks, and criminal/terrorist networks. Scale is often an issue with understanding and making use of large social networks. As the size of the network increases, it is harder to make sense of the network, and it is computationally costly to manipulate and maintain. Here we investigate methods for pruning social networks by determining the most relevant relationships in a social network. We measure importance in terms of predictive accuracy on a set of target attributes of social network groups. Our goal is to create a pruned network that models only the most informative affiliations and relationships. We present methods for pruning affiliation networks based on both structural properties and descriptive attributes.

An affiliation network is described by a set of actors A, a set of events E, and a set of membership relations R. The most common graph representation for affiliation networks is as a bipartite graph with two node types representing actors and events, and a single edge type representing membership relationships between actors and the events in which they participate.

Structural properties are determined by the graph structure of the network. Examples include the density of the graph, the average degree of nodes in the graph, the number of cliques in the graph, etc. Recent research has focused on understanding the structural properties of social networks. For a recent survey, see Newman [6]. Much of the work has been descriptive in nature, but recently there has been more work which uses structural properties for prediction. Within this category, there is work that focus on the spread of influence through the network (e.g., [2, 3]), prediction of future interactions between actors using network topology [5], methods for approximating the connectivity properties of a graph, and classification and clustering [1, 8].

In addition to the nodes and edges themselves, the nodes and edges in the affiliation network can have descriptive attributes or features associated with them. The descriptive attributes provide specific social context to the network. A corporate board social network may contain descriptive attributes representing the job function and age of a board member. A disease transmission social network may contain descriptive attributes representing the location of person's home and date of disease discovery.

2 Prediction in Social Networks

Our goal is to develop principled approaches to compressing and pruning social networks determining which portions of the network can be removed while minimizing information loss. Let $N = (A, E, R)$ be our original network and $N' = (A', E', R')$ be our pruned network.

We will focus on maximizing our predictive accuracy on the event attributes. For ease of exposition, we will assume we are attempting to maximize the predictive accuracy for a single event attribute $E.C_i$, based on attributes of related actors found using the co-membership information and based on attributes of related events found using the event overlap information. The idea is to construct a classifier, using local neighborhood information, to predict $E.C_i$. Now it is easy to see the difficulty with this setup. Each event may have

a different number of related actors and a different number of related events, so how can we construct features to use in our classifier? We solve this problem by computing an aggregate over the set of related actors and over the set of events. Aggregation is a common technique used to construct feature vectors in relational domains [4, 7]. Here we assume some set of aggregates is associated with each attribute.

We compare the classifier F_N constructed from the original social network $N = \{A, E, R\}$ with the classifier $F_{N'}$ constructed from a pruned social network $N' = \{A', E', R'\}$. We compare both accuracy on the training sets and accuracy on test sets. Our goal is to find pruned networks that are both compact and accurate.

3 Pruning Techniques

We consider two categories of techniques for pruning the network. The first involves removing edges from the affiliation network. The second involves removing actors (and incident edges) from the affiliation network. We can use different techniques for pruning a network. The three techniques of interest to us are: 1) pruning based on structural properties, 2) pruning based on descriptive attribute values, and 3) pruning based on random sampling.

Structural Pruning Structural network properties or measurements involve evaluating the location of actors in a social network. Two well known structural measures are *degree* and *betweenness*. The degree of a node is defined as the number of direct connections a node has to other nodes in the network. The nodes with the most connections are considered the most active nodes in the network and are referred to as *hubs*. Betweenness of a node corresponds to the number of cliques to which a node belongs. Nodes with high betweenness are considered to have great influence in the network and are referred to as *brokers*. Therefore, when pruning based on structure, we will be interested in removing actors that are not hubs and/or brokers from the network.

Descriptive Attribute-based Pruning Another pruning technique of interest involves pruning based on descriptive attributes. We prune edges by selecting on attribute values. We look at both the case where we keep *only* edges or nodes with a particular attribute value and also the case where we keep all edges *except* edges or nodes with value.

Random Sampling As a baseline, we compare pruning based on random sampling. This involves maintaining only a random sample of the actor population for analysis.

It is important to quantify the compression achieved by pruning. We use a relatively generic measure, the description length of the graph, $DL(N) = \log(|A|) + \log(|E|) + |R|(\log(|A|)\log(|E|))$.

4 Experimental Results

We analyzed two affiliation networks. The first data set, the Executive Corporation Network (ECN), contains information about executives of companies that are traded on the NASDAQ and the NYSE. The executives serve on the Board of Directors for one or more of the companies in the data set. This data was collected from the Reuter's market data website (yahoo.mulexinvestor.com) in January 2004. There are 66,134 executives and 5384 companies (3284 NASDAQ and 2100 NYSE). The relational schema describing the ECN is:

- A = Executive(exec_id, exec_name, age, education_level)
- E = Company(co_id, co_name, stock_exchange, sector, stock_price)
- R = BoardMembership(exec_id, co_id, officer_position, join_date)

The average board size is 14, the average number of boards an officer is on is 1.14, the number of officers serving on multiple boards is 6544, and the number of boards these officer are on is 2.4. We predict two attributes, *stock_exchange* and *sector*. A sector is a coarse grouping of industries, e.g., health care. We prune on descriptive attributes such as *officer_position*, e.g., CEO, President, etc.

The second data set, the Author Publication Network (APN), contains information about publications and their authors from the ACM SIGMOD anthology. There are 13,070 authors and 16,287 publications. The schema APN is:

- A = Author(author_id, author_name, affiliation, number_of_publications)
- E = Publication(pub_id, pub_type, pub_date, number_of_references, number_of_citations)
- R = PaperAuthorship(author_id, pub_id)

The average number of authors per publication is 2.4 and the average number of publications per author is 2.9. For APN, we predicted the two event attributes *pub_type* and *number_of_references* (to publication).

Our goal is to find small networks that can accurately predict event attributes. We compare the following affiliation networks:

- no pruning (**full**)
- descriptive attribute pruning (**descriptive**)
- pruning based on hubs and brokers (**structural**)
- random sampling (**random**)

We built event-attribute classifiers from the networks as described in Section 2. For categorical aggregate attributes, we calculated the mode of the neighboring event values, and for numeric aggregate attributes, we calculated the minimum, maximum and average of the neighboring event values. The classifiers were created using WEKA. We tested a

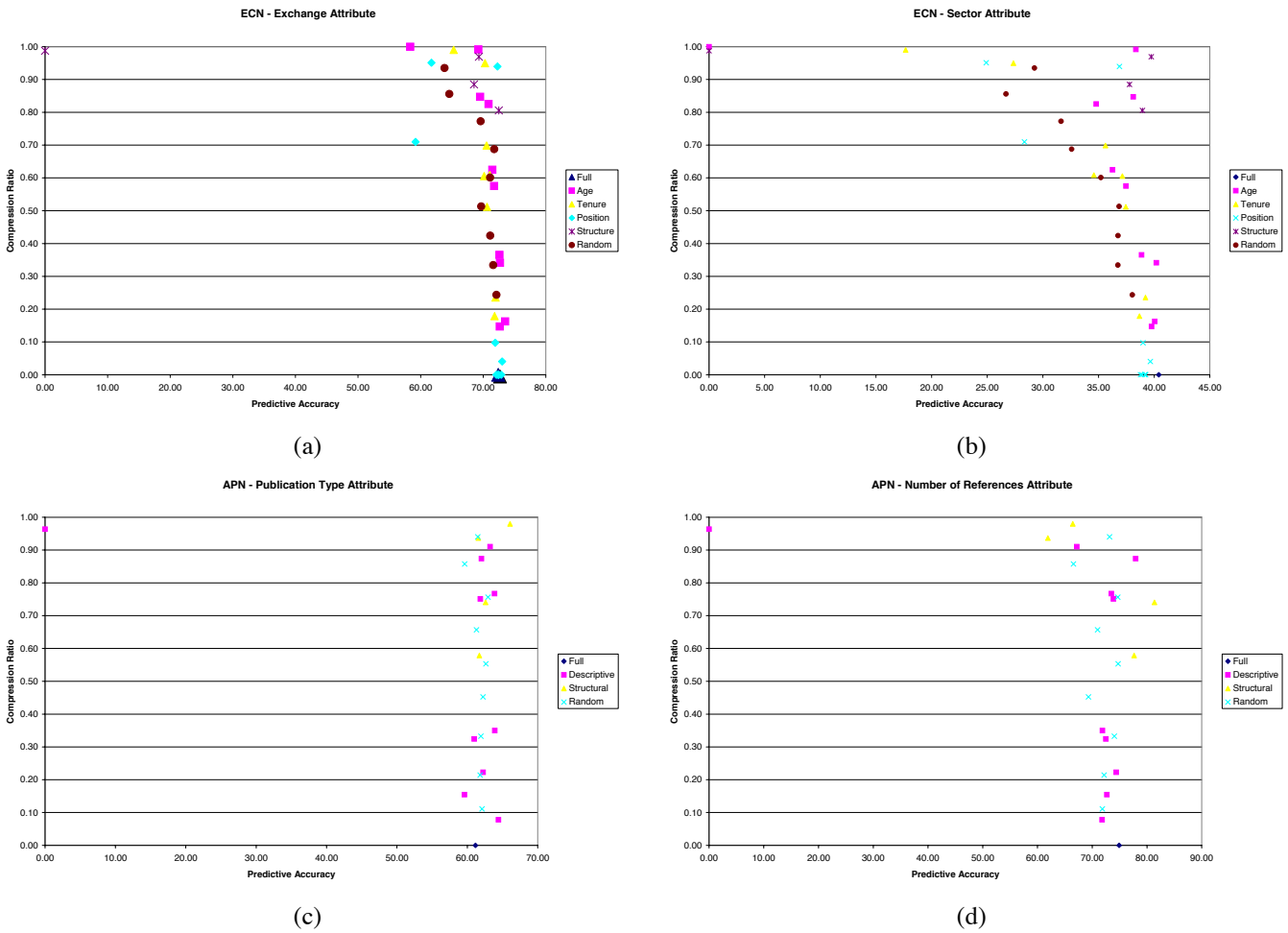


Figure 1. Comparisons of compression vs. accuracy for a variety of network pruning strategies for a) ECN exchange b) ECN sector c) APN pulbication type and d) APN number of references.

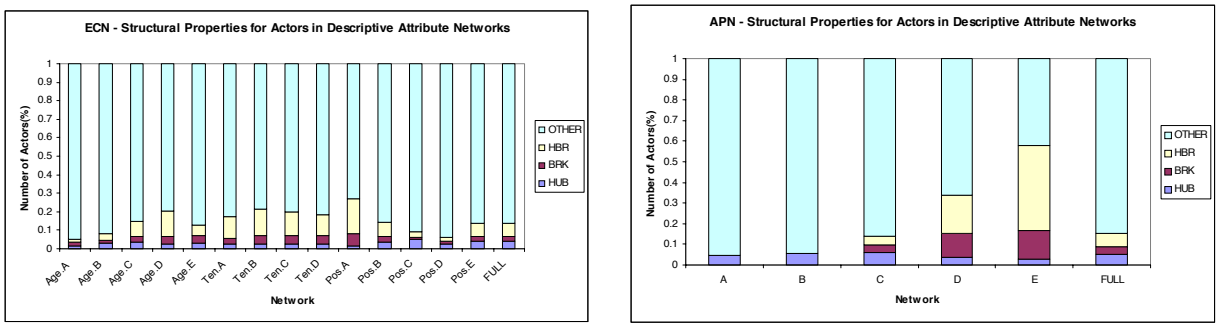
range of classification algorithms including decision trees, naive Bayes, and support vector machines (SVMs). The results were relatively consistent across classifiers; due to space constraints, here we present results only for SVMs using five-fold cross validation.

When constructing our feature vector, we constructed aggregates for the following ECN actor and event attributes: stock exchange, industry, sector, number of officers on a board, number of advanced degrees on a board and officer age of a board. We evaluated three descriptive prunings. The first two descriptive prunings, *position* and *tenure*, involve removing edges from our affiliation graph for executives based on the attributes *BoardMembership.officer_position* and *BoardMembership.join_date*. For example, one pruning of *BoardMembership.officer_position* keeps only edges of CEOs and removes all other membership edges from the network. The third descriptive pruning involves removing actors based on age.

To group attribute values, we binned numeric attributes and we abstracted categorical attributes. For both our networks, the binnings resulted in four to five bins for each attribute. For APN, we used the attribute *Author.number_of_publications* for descriptive pruning.

As mentioned earlier, descriptive attribute pruning has one of two interpretations for an attribute B with attribute value c: 1) maintain *only* actors with $B = c$ (**only**) and 2) maintain all actors *except* where $B = c$ (**except**). We evaluated pruning on every descriptive attribute value for each descriptive pruning category.

For structural pruning, we tested the following four cases: maintaining only actors who are hubs, maintaining only actors who are brokers, maintaining only actors who are both hubs and brokers, and maintaining only actors who are hubs or brokers. Finally, for random pruning, we com-



(a) (b)

Figure 2. The structural characteristics of actors in different prunings for a) ECN and b) APN.

pared results on random samples for 9 different sample sizes (between 10% and 90% of the actors in the network).

Figure 1 shows compression versus predictive accuracy for different attributes in each of the networks. The right upper corner represents the 'best' networks in terms of compression and predictive accuracy. Figure 1(a) shows results for the ECN exchange. The classifier built using the full network achieves accuracy of 72.4%. The best accuracy and compressions are for the following networks: pruning on position, we achieve an accuracy of 72.3% with a compression of 94% (in this case, we kept only the chairs); pruning on tenure, we achieve an accuracy of 70.29% with a compression of 95%, and pruning on age, we achieve an accuracy of 69.2% with a compression of 99% (in this case, we kept only the older executives). These accuracies are all significantly better than the baseline prediction accuracy of 61% achieved by simply choosing the most common exchange. For predicting the ECN sector, shown in Figure 1(b), the full network achieves accuracy of 40.4% and the the best networks are the ones that prune on age (we achieve accuracy of 40.2% with compression of 34%, in this case we kept the younger executives rather than the older ones) and structure (we achieve accuracy of 39.7% and compression of 97% by keeping only the brokers). Figure 1(c) and (d) show similar results for the pruned APN networks, with many of the pruned networks achieving significantly higher accuracies than classifiers built from the full network. For both APN attributes, the network pruned on structure that achieved the best accuracy-compression tradeoff was the one that kept only the actors that were both hubs and brokers. In all cases, pruning on descriptive attributes and structure properties significantly outperformed random pruning.

To better understand how the two relate, in Figure 2 we show the percentage of structural actor types (hubs, brokers (BRK), hubs and brokers (HBR), and other) preserved under various descriptive pruning strategies. These graphs show that for two different datasets, the networks created using descriptive pruning contain a different mix of actors than those created using structural pruning. This supports our claim that structural pruning and descriptive pruning are two distinct methods for compressing networks.

We believe that exploring descriptive and structural compression techniques together, beyond allowing compact and accurate compression of networks, is also important for identifying actors that are the most useful for network understanding. In this paper we showed how to use structural properties and descriptive attributes to prune social networks. world data sets. While the networks resulting from structural pruning and descriptive pruning are quite distinct, both are viable approaches for reducing the size of a social network while still maintaining predictive accuracy on a set of target event attributes.

Acknowledgments: Work supported under NSF #0423845.

References

[1] R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu. Mining newsgroups using networks arising from social behavior. In *International World Wide Web Conference*, 2003.

[2] P. Domingos and M. Richadson. Mining the network value of customers. In *ACM Intl. Conf. on Knowledge Discovery and Data Mining*, 2001.

[3] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *ACM Intl. Conf. on Knowledge Discovery and Data Mining*, 2003.

[4] A. J. Knobbe, M. de Haas, and A. Siebes. Propositionalisation and aggregates. In *Eur. Conf. on Principles of Data Mining and Knowledge Discovery*. Springer-Verlag, 2001.

[5] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Intl. Conf. on Information and Knowledge Management*, 2003.

[6] M. Newman. The structure and function of complex networks. *IAM Review*, 45(2):167–256, 2003.

[7] C. Perlich and F. Provost. Aggregation-based feature invention and relational concept classes. In *Intl. Conf. on Knowledge Discovery and Data Mining*, 2003.

[8] M. F. Schwartz and D. C. Wood. Discovering shared interests using graph analysis. *Communications of the ACM*, 36(8), 1993.