Using Friendship Ties and Family Circles for Link Prediction

Elena Zheleva and Lise Getoor
Computer Science
University of Maryland
College Park, MD
{elena,getoor}@cs.umd.edu

Jennifer Golbeck
College of Information Studies
University of Maryland
College Park, MD
jgolbeck@umd.edu

Ugur Kuter
Institute for Advanced Computer Studies
University of Maryland
College Park, MD
ukuter@cs.umd.edu

ABSTRACT

Social networks can capture a variety of relationships among the participants. Two of the most commonly studied are friendship and family, or kinship, ties. Most existing work studies these networks in isolation. Here, we study how these networks can be overlaid. We study the predictive power of overlaying friendship and family ties on a trio of interesting real-world social networks. We show that when there are tightly-knit family groups, which we refer to as family circles, in a social network, we can improve the accuracy of our link prediction models. This is done by making use of the family-circle features based on the likely structural equivalence of participants in these groups. Our experiments confirm this, and we achieve significantly higher prediction accuracy (between 15% and 30% more accurate) as compared to using more traditional features such as descriptive node attributes and structural features. We also show that a combination of all three types of attributes results in the best precision-recall trade-off.

1. INTRODUCTION

There is a growing interest in social media and in data mining methods which can be used to analyze, support and enhance the effectiveness and utility of social media sites. The analysis methods being developed build on traditional methods from the social network analysis community, extend them to deal with the heterogeneity and growing size of the data being generated and use tools from graph mining, statistical relational learning and methods for information extraction from unstructured and semi-structured text.

Traditionally, social network analysis has focused on actors and ties, or relationships, between them such as friendships and kinships. The notion of "structural equivalence," when two actors are similar based on participating in equivalent relationships, is fundamental to finding groups in social networks. Similarly, there has also been much work in community finding, where densely connected groups of actors are clustered together into communities.

Most of the existing work focuses on networks that exhibit a single relationship, such as friendship or affiliation. The two most common types of networks are unimodal networks, where the nodes are actors and the edges represent ties such as friendships, and affiliation networks which can be represented as bipartite graphs in which there are two types of nodes, the actors and organizations, and the edges represent the affiliations between actors and organizations.

In this paper, we investigate the power of combining friendship and affiliation networks. Our approach is an attempt to bridge approaches based on structural equivalence and community detection. We show how predictive models, based on descriptive, structural and community features, perform surprisingly well on challenging link-prediction tasks. We validate our results on a trio of novel social media websites describing pets, and their friendships and family relationships. With our results, we hope to motivate further research in discovering closely-knit groups in social networks and using them to improve link-prediction performance.

Our link-prediction approach can be applied to a variety of domains. The important properties of the data that we use are that there are actors, links between them and closely-knit groups. In some data, groups are given; in other datasets, it may be necessary to first cluster the nodes in a meaningful manner.

Our contributions include the following:

- We propose a general framework for combining social and affiliation networks.
- We show how to instantiate it for overlaying friendship and family networks.
- We show how features of the overlaid networks can be used to accurately predict friendship relationships.
- We validate our results on three social media websites describing pets and their social networks.

In this research, we used data from three social media websites: Dogster, Catster, and Hamsterster [1]. These pet social networking websites, or *petworks*, allow members to post and share information describing their pets, their pets' friends, and their pets' family members. Figure 2 shows part of a representative profile from Dogster. On all these sites,

[1]At `http://www.dogster.com`, `http://www.catster.com`, and `http://www.hamsterster.com`.

The 2nd SNA-KDD Workshop '08 (SNA-KDD'08), August 24, 2008, Las Vegas, Nevada, USA.

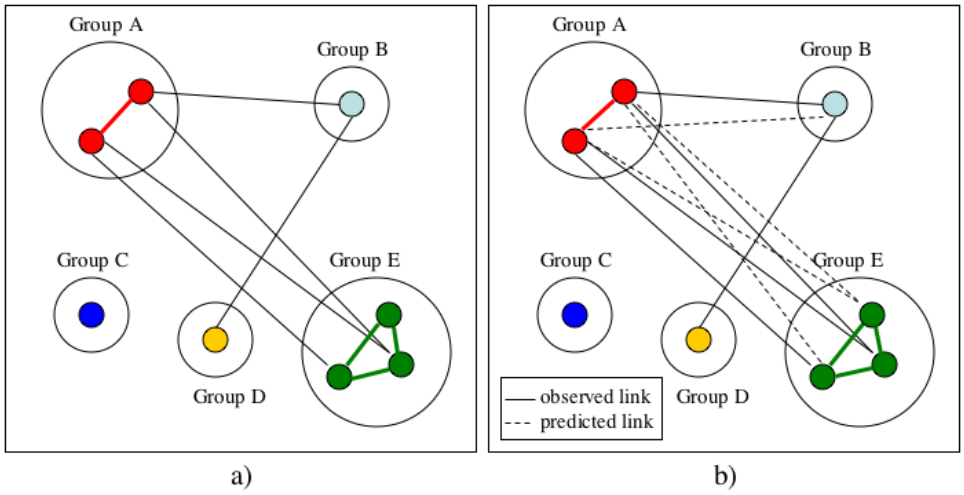

Figure 1: Actors in the same tightly-knit group often exhibit structural equivalence, i.e., they have the same connections to all other nodes. Using the original network (a), and a structural equivalence assumption, one can construct a network with new predicted links (b).

members maintain profiles that include photos, personal information and characteristics. Members also maintain links to friends and family members. As of February 2007, Dogster has approximately 375,000 members. Catster is based on the same platform as Dogster and contains about 150,000 members. Hamsterster has a different platform, but it contains similar information and has about 2,000 members.

2. SOCIAL NETWORK MODEL

We begin with a description of the social network model. Social networks describe actors and their relationships. The actors can have attributes such as *age* and *income*. Relationships can represent dyadic (binary) relationships or they can represent group memberships (cliques or hyperedges); in addition, relationships can be directed, undirected and/or weighted. Here we consider both dyadic and group-membership relationships. Specifically, we consider friendship relationships and family group memberships. In our domain, these are undirected, unweighted relationships.

More formally, the networks we consider consist of:

actors: a set of actors $A = \{a_1, \ldots, a_n\}$,

and their partitioning into non-overlapping groups:

groups: a group of individuals connected through a common affiliation. The affiliations group the actors into sets $G = \{G_1, \ldots, G_m\}$.

We consider the following relationships:

friends: $F\{a_i, a_j\}$ denotes that a_i is friends with a_j, and

family: $M\{a_i, G_k\}$ denotes that a_i is a part of family G_k.

Actors can have attributes; if b is an attribute, then we use $a_i.b$ to denote the b attribute of actor a_i. We denote the set of friends of actor a_i by $a_i.F$, and the set of family members of the same actor as $a_i.M$.

Figure 1(a) shows an example network of eight actors and five groups. Each node represents an actor, and a group is shown as a circle around the actors. The thick lines inside a group mark family relationships, and the thin black lines denote friendship relationships. There are single-member groups, and there are actors without friends.

3. PREDICTING LINKS IN SOCIAL NETWORKS

Here we study the problem of predicting friendship links in multi-relational social networks. Link prediction is useful for a variety of tasks. The most straight-forward use is for making data entry easier – a link-prediction system can propose links, and users can select the friendship links that they would like to include, rather than users having to enter the friendship links manually. Link prediction is also a core component of any system for dynamic network modeling—the dynamic model can predict which actors are likely to gain popularity, and which are likely to become central according to various social network metrics.

Link prediction is challenging for a number of reasons. When it is posed as a pair-wise classification problem, one of the fundamental challenges is dealing with the large outcome space; if there are n actors, there are n^2 possible relations. In addition, because most social networks are sparsely connected, the prior probability of any link a priori is extremely small, thus we have to contend with a large class skew problem. Furthermore, because the number of links is potentially so large, the number of the negative instances will be huge, so constructing a representative training set is challenging.

In our approach to link prediction in multi-relational social networks, we explore the use of both attribute and structural features, and, in particular, we study how group membership (in our case, family membership) can significantly aid in accurate link (here, friendship) prediction.

4. A FEATURE TAXONOMY

We identified three classes of features in these networks that describe characteristics of potential links in the social network: *descriptive attributes*, *structural attributes* and *group attributes*. The descriptive attributes are attributes inherent to the nodes such as 'dog breed,' and they do not consider the structure of the network. The structural attributes include characteristics of the networks based on the friendship relationships such as node degree. The group attributes are based on structural properties of the network when both types of relationships, friendship and family, are

Figure 2: A sample profile on Dogster which includes family and friends.

considered. The groups in this case are the cliques of family members which partition the graph. The latter features are given by the family affiliations of the actors in the network. Each feature within a class can be assigned to an actor or to a pair of actors, a potential edge. The following sections describe our taxonomy of the features in more detail.

4.1 Descriptive attributes

The descriptive attributes are attributes of nodes in the social network that do not consider the link structure of the network. They provide semantic insight into the inherent properties of each node, or compare the values of the same inherent attributes for a pair of nodes. For the pets in Dogster and Catster, we used the following attributes:

1. *Actor features.* These are inherent actor attributes.

   Breed. This is the pet breed such as *golden retriever* or *beagle*. A pet can have more than one breed.

   Breed category. Each breed belongs to a broader category set. In Dogster, the major breed categories are *working*, *herding*, *terrier*, *toy*, *sporting*, *non-sporting*, *hound*, and *other*. The breed category of a dog with multiple breeds is *mixed*.

   Single Breed. This boolean feature describes whether a pet has multiple breed characteristics.

   Purebred. This boolean feature specifies whether a dog owner considers its pet to be purebred or not.

2. *Actor-pair features.* All of the above features describe characteristics of a single pet in the network. The actor-pair features compare the values of the same node attribute for a pair of nodes.

   Same breed. This boolean feature is true if two pets have at least one common breed.

Descriptive features vary across domains. The above features were applicable to Catster and Dogster. Hamsterster had a richer set of features which included a binary feature which described whether the pet is a hamster or a gerbil, and an actor-pair feature which described whether the pets in the pair are both alive.

4.2 Structural features

All of the above features describe characteristics of a node in the network. The next set of features that we introduce describe features of network structure. The first is a structural feature for a single node, a_i, while the remaining describe structural attributes of pairs of nodes, a_i and a_j.

1. *Actor features.* These features describe the link structure around a node.

   Number of friends. The degree, or number of friends, of an actor a_i: $|a_i.F|$.

2. *Actor-pair features.* These features describe how interconnected two nodes are. They measure the sets of friends that two actors have $a_i.F$ and $a_j.F$.

   Number of common friends. The number of friends that the pair of nodes have in common in the network: $|a_i.F \cap a_j.F|$.

   Jaccard coefficient of the friend sets. The Jaccard coefficient over the friend sets of two actors describes the ratio of the number of their common friends to their total number of friends:

   $$Jaccard(a_i, a_j) = \frac{a_i.F \cap a_j.F}{a_i.F \cup a_j.F}.$$

   The *Jaccard coefficient* is a standard metric for measuring the similarity of two sets. Unlike the feature *number of common friends*, it considers the size of the friendship circle of each actor.

   Density of common friends. For the set of common friends, the density is the number of friendship links between the common friends over the number of all possible friendship links in the set. The density of common friends of two nodes describes the strength in the community of common friends. Density is also known as *clustering coefficient*.

4.3 Group features

The third category of features use group membership; in the petworks case, the groups are families. These are the features that overlay friendship and affiliation networks.

1. *Actor features.* They describe the groups to which an actor belongs.

   Family Size. This is the simplest attribute and describes the size of an actor's family: $|a_i.M|$.

2. *Actor-pair features.* There are two types of features for modeling these inter-family relations:

   Number of friends in the family. The first feature describes the number of friends a_i has in the family of a_j: $|a_i.F \cap a_j.M|$. This feature allows one to reason about the relationship between an actor and a group of other actors, where the latter is semantically defined over the network through the family relations.

Portion of friends in the family. The second feature on inter-family relations describes the ratio between the number of friends that a_i has in a_j's family (the same as the above feature) and the size of a_j's family.

The idea behind the group features is based on the notion of *structural equivalence* of nodes in a group. Two nodes are structurally equivalent if they have the same links to all other actors. If we can detect tightly-knit groups in a social network and we assume that the nodes in each group are likely to behave similarly, then new links can be predicted such that the nodes in the group become structurally equivalent. In our petworks, such groups are the family cliques. Figure 1 shows an example. If one of the actors from Group A is friends with an actor from Group B, as shown on the original network (a), then it is highly likely that there is a link between the other actor from Group A and the actor from Group B, shown as a dashed line in (b).

5. ALTERNATIVE NETWORK OVERLAYS

The traditional approach to studying networks is to treat all relationships as equal. We propose overlaying networks with different link types in a way that distinguishes between link types, and uses information about affiliation groups. In other words, our link-prediction approach uses information about the actors A, the groups G, the friendship relationships F, and the family relationships M. We call this overlay *different-link and affiliation overlay*. Therefore, a logical question one may ask is what is the benefit of treating links as different, and whether affiliation groups really make a difference in link prediction. Our claim is that affiliations are important and that they can have a predictive value. To illustrate the benefit of our approach as compared to the traditional one, we compare the *different-link and affiliation overlay* to two alternative overlays of the network.

In the first overlay, which we call *same-link and no affiliation overlay*, the family and friendship links are treated the same, and affiliation groups *are not* given. We test the null hypothesis that the alternative overlay can offer the same or better link-prediction accuracy as the overlay that we have been discussing so far. More formally, in this alternative overlay, the graph consists of these components: actors A, and a set of edges to which we refer as *implied friendships* $F_{implied} = F \cup M$. We can compute the descriptive and structural features in this overlay for link prediction.

If the null hypothesis is rejected then we still need to check whether the predictive value of our initially proposed network overlay comes from treating the links as different or from the fact that we are given the affiliation groups. To test that, we also look at a second alternative overlay, the *same-link and affiliation overlay*, in which the family and friendship links are treated the same, and affiliation groups *are* given. In this overlay, the graph consists of these components: actors A, groups G, and implied friendships $F_{implied}$. We can compute all classes of features in this overlay.

6. EXPERIMENTAL EVALUATION

6.1 Data description

We have obtained a random sample of $10,000$ pets each from Dogster and Catster, and all 2059 pets registered with Hamsterster. Each instance in the data contained the features for a pair of pets where some of the features were individual pet features. For each pair of pets in the data, we computed the features from the three classes described in Section 4. An instance for a pair of pets a_i and a_j includes both the individual actor features and the actor-pair features. It has the form $<a_i$ *features,* a_j *features,* (a_i, a_j)*-pair features, class*$>$ where *class* is the binary class which denotes whether a link exists between the actors.

Table 1: Comparison of F1 values in the three datasets, with the feature types from our taxonomy.

Feature Type	Dogster	Catster	Hamsterster
Descriptive (D)	37.6%	0.0%	0.0%
Structural (S)	76.1%	83.1%	59.9%
Group (G)	90.8%	95.2%	89.2%
D and S	78.6%	83.0%	60.3%
D, S and G	94.8%	97.9%	90.5%

For Dogster, the sample of 10,000 dogs had around 17,000 links among themselves, and we sample from the non-existing links at a 1:10 ratio. For Catster, the 10,000 cats had 43,000 links, and for the whole Hamsterster dataset, the number of links was around 22,000. We sampled from the non-existing links in these datasets at the same 1:10 ratio.

6.2 Experimental setup

We used three classifiers: Naïve Bayes, logistic regression and decision trees and performed binary classification on the test instances to predict friendship links. The implementations of these classifiers were from Weka (v3.4.12). We measured accuracy by computing precision, recall, and their harmonic mean, F1 score, using 10-fold cross-validation.

6.3 Link-prediction results

We report only on the results from decision-tree classification because it consistently had the highest accuracy among the three classifiers. Table 1 summarizes our results. Adding group features to the descriptive and structural features increased accuracy by 15% to 30%.

6.3.1 Descriptive attributes can be useful in combination with structural attributes

In these experiments, we have investigated the predictive power of the simplest features, i.e., the descriptive attributes versus the impact of the structural attributes. Table 1 shows the accuracy results from the decision-tree classifier. When we use only descriptive attributes, the link-prediction accuracy varies across datasets. In Dogster, there is some advantage to using descriptive attributes, yet the accuracy (F1 score) is relatively low 37.6%. In Catster and Hamsterster, the classifier was not able to separate the positive and negative instances based only on the descriptive features, and it had 0% accuracy. This confirms that, in general, link prediction is a challenging prediction task.

When we used the structural features (such as number of friends that two pets share), the link-prediction accuracy increased to 76.1% in in Dogster. This suggests that the structural features are much more predictive than simple descriptive attributes. This effect was even more pronounced for Catster and Hamsterster.

In Dogster, combining the node attributes and the structural features leads to futher improvement. Using descrip-

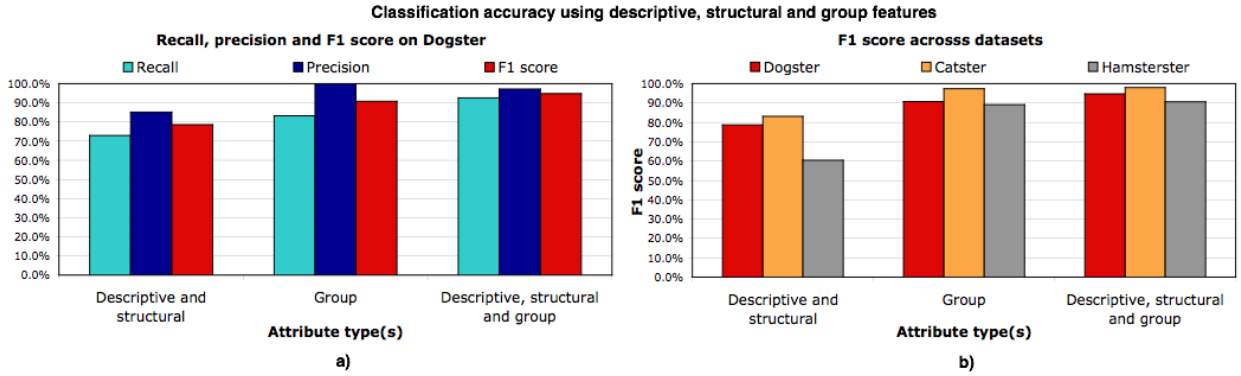

Figure 3: Link-prediction accuracy using all feature classes: descriptive, structural and group features. a) Recall, precision, and F1 score for Dogster; b) F1 score across datasets. Group features are highly predictive, yet adding the other features provided benefit too.

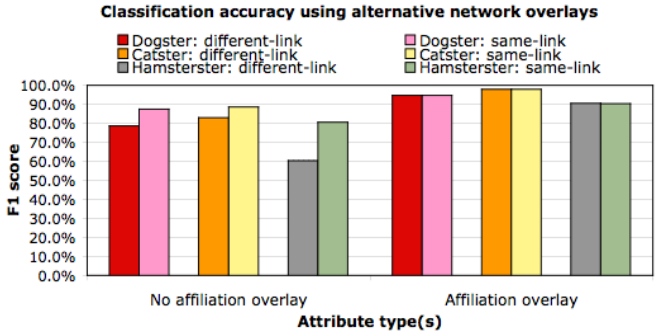

Figure 4: Prediction accuracy when links are treated equal, with and without group affiliations. Group features are the main contributor to the high link-prediction accuracy.

tive attributes together with structural attributes leads to a better F1 score (78.6%) as compared to using either category alone (37.6% and 76.1%, respectively) in Dogster. For Catster and Hamsterster, the difference was less than 0.4%.

6.3.2 *Family group features are highly predictive*

As the previous experiments showed, structural attributes are stronger predictors than the descriptive attributes alone. Next, we investigate the predictive power of the group features in our taxonomy. In Dogster, Catster and Hamsterster, the group features involve the families and friends of the pets. Figure 3 shows our comparisons. Our results suggested that group features are strong predictors for friendship links ($F1 = 90.8\%$ for Dogster). We also ran experiments where we used not only family cliques, but also the structural and descriptive features. In these experiments, the results show that the accuracy (F1) improves by 4% in Dogster, 0.6% in Catster and 1.3% Hamsterster.

6.3.3 *Computing more expensive structural attributes is not highly beneficial*

Some structural features in our taxonomy were more computationally expensive to construct than others. For example, the feature that described the number of pet friends is easy to compute, whereas the feature that described the density of common friends for each pair of pets is the hardest. Using a database, computing density of common friends for all pairs of pets requires several joins of large tables. In order to investigate the trade-off between computing expensive features and their predictive impact on our results, we have performed the following experiments.

We have designed experiments in which we add more expensive structural features one by one, and assess the link-prediction accuracy at each step. We used the following combinations of features: (1) using *number of friends* only, (2) using *number of friends* and *number of common friends*, (3) using *number of friends*, *number of common friends* and *jaccard coefficient*, and finally (4) using *number of friends, number of common friends, jaccard coefficient* and *density of common friends*. We are reporting on the results of these four sets of structural features together with the descriptive attributes since we showed in the previous subsection that using descriptive attributes can sometimes be beneficial. We also report on the setting in which group features were used.

Surprisingly, it turned out that computing the more expensive features added very little benefit. For example, in the Dogster case, adding the *number of common friends* of two nodes improved accuracy (F1 score) by 2% over the individual *number of friends*. Computing the most expensive feature *density of common friends* pays off slightly (improves F1 score by 0.4%) only when there are no group attributes. Computing the more expensive *jaccard coefficient* did not pay off over using the simpler feature *number of common friends*. In the Catster and Hamsterster cases, the improvement was less that 0.5%. Our results also support the claim made in the preferential attachment model [2] that the number of friends of a node (node degree) plays role in the process of new nodes linking to it. They contradicted the link-prediction results in co-authorship networks [10] where *jaccard coefficient* and the *number of common friends* consistently out-performed the metric based on number of friends. This may be inherent to the types of networks discussed.

6.3.4 *Alternative network overlays in petworks*

In the next set of experiments, we used the alternative network overlays to test whether there was an advantage to keeping the different types of links and the affiliation groups. The overlay that we proposed in the paper was *different-link and affiliation overlay*, and the alternative overlays that we compared it to were *same-link and no affiliation overlay* and *same-link and affiliation overlay* (see Section 5). We compute only the descriptive and structural features in the over-

lay with no affiliation information, and compute all classes of features in the overlays where affiliation information was given.

The results on Figure 4 show that when family affiliation was given, it did not matter whether the links were treated equally: the classifiers produces almost the same results. However, in the case when the affiliations were not given, it was better to compute the structural features using both types of relationships and treat them equally. When family links were treated equally with friendship links, the accuracy of the predictions made by the structural attributes improved by 6% to 20%. This may be due to the fact that the overlap between friends and family links in the data was very small, and using both types of links when computing the structural features was beneficial. Using the affiliation information and computing all features on the data led to the best accuracy, and the accuracy was the same both in the different-link and same-link cases. These experiments also confirmed the previous results: group features were the main contributor to the high link-prediction accuracy.

7. RELATED WORK

In general, link-prediction algorithms process a set of features in order to learn and predict whether it is likely that two nodes in the data are linked. Sometimes, these features are hand-constructed by analyzing the problem domain, the attributes of the actors, and the relational structure around those actors [1, 5, 10, 14]. Other times, they are automatically generated, i.e., the prediction algorithm first learns the best features to use and then predicts new links [13].

The link-prediction techniques that are based on feature-construction are closest to our work [1, 5, 6, 10, 14]. As most of the relational domains can be represented as a network model, the constructed features not only include the attributes of the actors, but also the characteristics of the structure. Most of this work examines co-authorship and citation networks [5, 10, 13, 14] whereas we validate our method using online social networks. Some of the approaches use machine learning techniques for classification [5, 13, 15], and others rely on ranking the feature values [1, 10, 14]. The novelty in our work is that we overlay two types of networks and explore different combinations of descriptive, structural and group features. Other related work looks at the problems of link ranking [9], link completion [4], link anomaly discovery [6, 14] and group detection [3, 8].

8. DISCUSSION

When studying other large social networks, family information is not always relevant or available. However, groups and affiliations are often available, or communities can be discovered. The networks used here had binary relationships - friend or family - but a similar effect can be achieved in networks where relationships are weighted. For example, co-authorship networks are widely studied as social networks [2, 5, 10, 11, 12, 13, 14], and edges can be weighted by the number of articles a pair of authors have authored together. In email communication networks - the Enron email corpus [6, 7], for example - the number of messages between two senders can be used as a weight. To mimic the strong family-type relationship we used in this article, a threshold weight can be set. Any edge with a weight over that threshold can be treated as a "strong" relationship (like our family relationship). Clusters of nodes connected with strong ties would represent the equivalent of a family unit.

9. CONCLUSIONS AND FUTURE WORK

Link prediction is a notoriously difficult problem. In this research, we found that overlaying friendship and affiliation networks was very effective. For the petworks, we found that family relationships were very useful in predicting friendship links. Our experiments show that we can achieve significantly higher prediction accuracy (between 15% and 30% more accurate) as compared to using more traditional features such as descriptive node attributes and structural features. Family groups helped not only because they represent a clique of actors, but because the family relationship itself was indicative of structural equivalence. As future work, we plan to investigate the use of edge weights and thresholds to define strongly connected clusters, and see if it works as well in link prediction as the family groups did here.

10. REFERENCES

[1] L. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[2] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *PHYSICA A*, 311:3, 2002.

[3] J. Friedland and D. Jensen. Finding tribes: Identifying close-knit individuals from employment patterns. In *KDD*, 2007.

[4] A. Goldenberg, J. Kubica, P. Komarek, A. Moore, and J. Schneider. A comparison of statistical and machine learning algorithms on the task of link completion. In *LinkKDD Workshop*, 2003.

[5] M. Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *SDM Workshop on Link Analysis, Counter-terrorism and Security*, 2006.

[6] Z. Huang and D. Zeng. A link prediction approach to anomalous email detection. In *SMC*, 2006.

[7] B. Klimt and Y. Yang. The enron corpus: A new dataset for email classification research. In *ECML*, volume 3201 of *LNCS*, pages 217–226. Springer, 2004.

[8] J. Kubica, A. Moore, J. Schneider, and Y. Yang. Stochastic link and group detection. In *AAAI*, 2002.

[9] J. M. Kubica, A. Moore, D. Cohn, and J. Schneider. cgraph: A fast graph-based method for link analysis and queries. In *IJCAI TextLink Workshop*, 2003.

[10] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM*, 2003.

[11] M. Newman. Who is the best connected scientist? a study of scientific coauthorship networks. Working Papers 00-12-064, Santa Fe Institute, Dec. 2000.

[12] M. Newman. Coauthorship networks and patterns of scientific collaboration. *NAS*, 101:5200–5205, 2004.

[13] A. Popescul and L. H. Ungar. Statistical relational learning for link prediction. In *IJCAI Workshop on Learning Stat. Models from Relational Data*, 2003.

[14] M. J. Rattigan and D. Jensen. The case for anomalous link discovery. *SIGKDD Explor.*, 7(2):41–47, 2005.

[15] B. Taskar, M.-F. Wong, P. Abbeel, and D. Koller. Link prediction in relational data. In *NIPS*, 2003.