Using Transactional Information to Predict Link Strength in Online Social Networks

Indika Kahanda[1] **and Jennifer Neville**[2,3]
Departments of [1]Electrical and Computer Engineering, [2]Computer Science, and [3]Statistics
Purdue University
West Lafayette, Indiana
{ikahanda, neville} @purdue.edu

Abstract

Many scientific fields analyzing and modeling social networks have focused on manually-collected datasets where the friendship links are sparse (due to the costs of collection) but relatively noise-free (i.e. they indicate strong relationships). In online social networks, where the notion of "friendship" is broader than what would generally be considered in sociological studies, the friendship links are denser but the links contain noisier information (i.e., some weaker relationships). However, the networks also contain additional *transactional* events among entities (e.g., communication, file transfers) that can be used to infer the true underlying social network. With this aim in mind, we develop a supervised learning approach to predict *link strength* from transactional information. We formulate this as a link prediction task and compare the utility of attribute-based, topological, and transactional features. We evaluate our approach on public data from the Purdue Facebook network and show that we can accurately predict strong relationships. Moreover, we show that *transactional-network* features are the most influential features for this task.

Introduction

Recent research in machine learning, has demonstrated the utility of modeling social network information in domains such as fraud detection (Neville et al. 2005), citation analysis (McGovern et al. 2003), and marketing (Domingos and Richardson 2001). The presence of relational links in data from these domains offers a unique opportunity to improve model performance because inferences about one object can be used to improve inferences about related objects. For example, fraud and malfeasance exhibit homophily[1], thus if we know one person is involved in fraudulent activity, then his associates have an increased likelihood of being engaged in misconduct as well. Indeed, recent work in relational modeling has shown that collective inference over an entire dataset can result in more accurate predictions than conditional inference for each instance independently (e.g., (Chakrabarti, Dom, and Indyk 1998)) and that the gains over conditional models increase as homophily increases (Jensen, Neville, and Gallagher 2004).

The accuracy of these modeling techniques, however, is contingent on the presence of links in the data that confer homophily. Indeed, recent research that has attempted to prune away spurious relationships and highlight stronger relationships has been shown to improve the accuracy of relational models (Sharan and Neville 2008). These results are consistent with sociological research that has found pairs of individuals with *strong ties* (e.g., close friends) exhibit greater similarity than those with *weak ties* (e.g., acquaintances) (Granovetter 1983).

In small-scale social networks that have been manually collected through surveys, the resulting networks are often sparse but the links generally reflect strong relationships (due to the targeted collection process). On the other hand, the explosive growth of the Internet and electronic communication has recently facilitated the automatic collection of large-scale social networks. These networks are often more dense and contain more noise. This is due to the construction of the networks from transactional data (e.g., email, phone calls) or due to the low-cost of friendship identification (e.g., in online social networks). In both cases, the constructed networks contain both strong and weak ties with little or no information to differentiate between the two types of links. The goal of this work is to develop automated methods to differentiate between strong and weak relationships in these large-scale social networks.

In dynamic network domains, there is often ancillary data recording low-level interactions among the entities (e.g., emails, file transfers). For example, in online social networks such as Facebook, members continuously visit other members' pages, accessing content, posting comments and pictures, and sending messages. We believe that these *transactional* use patterns can be exploited to infer the nature and strength of relationships among members. More specifically, we conjecture that low-level interactions among entities provide evidence of the latent high-level social network structure, and that the patterns of interactions over time can be accurately and efficiently modeled to identify stronger relationships that confer a higher degree of homophily. For example, we may have communication events (e.g., phone calls, emails), data access/transfer events (e.g., web brows-

[1]The tendency of like to associate with like (McPherson, Smith-Lovin, and Cook 2001).

ing, file access), or localization events (e.g., meetings, conferences). These low-level transactions among individuals are easy to record electronically but a single event does not (necessarily) indicate a meaningful relationship between the participating parties. However, repeated interactions in multiple contexts do suggest a stronger relationship.

We formulate the problem of differentiating between strong and weak ties as a link prediction task where, given a pair of linked individuals, the aim is to predict whether or not the pair has a strong relationship. Our method is applicable to transactional network domains where relationships are either observed explicitly (e.g., friendship links in online social networks) or where relationships are constructed from interactions (e.g., email networks). In this paper, we take a supervised learning approach to the problem, using labeled examples of strong relationships. We analyze data from the public Purdue Facebook network, where we have friendship links, profile information, wall postings, picture postings, group memberships, and tags indicating "top friends." Under the assumption that "top friends" indicate a person's strongest relationships, we learn a model to predict which of a person's friends will be their "top friend". We use existing machine learning methods and consider features from four different categories: attribute similarity, topological connectivity, transactional connectivity, and network-transactional connectivity. We show experimentally, that we are able to predict "top friends" accurately and that the most influential features are those that consider transactional information in the context of the larger network structure (i.e., network-transactional features).

Background

Online social networks

Online social network sites such as Facebook, Orkut, and MySpace allow members to maintain user profiles with basic information, interests, and friends, as well as interact with other users by posting comments, sending messages, tagging photos, etc. Friendships links are generally undirected (a friendship link is formed through agreement by both users and appears on both profiles) and due to the ease of electronic collection, more abundant than previous small-scale social network datasets. For example, in the Purdue Facebook network, the median and average degrees are 78 and 46 respectively, whereas the median and average degrees in the social networks collected by the National Longitudinal Study of Adolescent Health (Harris 2008) are 8 and 7 respectively. However, online social networks and other electronically-collected networks often contain ancillary *transactional* information that can be used in both descriptive and predictive models. For example, in Facebook, members can send each other email, write short comments on friends' profile pages (i.e., their wall), post photos and tag the members that appear in them, invite friends to join groups, etc. This transactional information records low-level interactions among related nodes and can be used to predict which linked members are close friends, as opposed to acquaintances.

Social network analysis

Social network analysis and link analysis are a collection of techniques for calculating descriptive models of networks (e.g., (Wasserman and Faust 1994)). Approaches range from estimating measures of node and link centrality based on the topology of the graph (e.g., (Brandes 2001)), to learning more advanced probabilistic models to describe the link structure of networks (e.g., (Robins et al. 2007)), to characterizing the structure and evolution of communities in the networks (e.g., (Girvan and Newman 2002; Kumar, Novak, and Tomkins 2006)). However, nearly all these methods focus on descriptive statistics and generative models of link structure, rather than predictive modeling of specific node/link attributes. Moreover, nearly all these methods focus on modeling the link structure in isolation and do not exploit the dependencies between the observed attributes and behaviors and the relational structure. As a result, they offer a means for calculating potential aggregate features to include in our statistical models (e.g., sender centrality) but they offer no support for developing predictive models of link strength on their own.

Data fusion

The broad area of data fusion is relevant to the task of combining information from multiple transactional networks (e.g., phone calls and emails) to predict the underlying social network (i.e., strong ties). *Data fusion* is the process of combining of data from multiple sources such that the resulting information is "better" than using the sources individually. Recent work on data fusion methods has focused on applications in biological (Lanckriet et al. 2004) and web-based datasets (Xu, King, and Lyu 2007). In these approaches, the data from each source is compiled into a matrix of similarity scores for each pair of entities (e.g., proteins, web pages) and then each source is weighted appropriately during aggregation of the matrices. This work has focused on fusing datasets with widely varying types of information (e.g., time-series expression data, DNA strings) to assess the similarity between all pairs of entities in the data. We are instead interested in modeling multiple sources of *transactional* information (e.g., phone calls, email) between pairs of *related* entities to predict which relationships are strong. We conjecture that larger relational context, within which the transactional information resides, will be critical to the fusion process. For example, if one network shows weak linkage from A to C and from B to C, but A and B are strongly linked in another network then this provides more evidence to infer strong links for $A - C$ and $A - B$.

Link Prediction

There has been a surge of interest in the link prediction task—which is a formulization of the problem of predicting future links in a social network, given a snapshot of the network at the current time step. This is the area of research that is most relevant to our work in this paper. Link prediction methods can be generally grouped into two approaches: those that use just the link structure of the network and those that use both the attributes on nodes in the

network. In the former category, the methods typically use *topological* features that measure the connectivity of nodes in the network (e.g., (Liben-Nowell and Kleinberg 2004; Kashima and Abe 2006)). In the latter category, the methods typically incorporate additional *similarity* features that measure the correspondence among the attributes of the nodes (e.g., (Taskar et al. 2003; Hasan et al. 2005)). The link prediction task is very challenging due to the extremely large class skew (in social networks the majority of node pairs are not linked) and as such researchers have recently investigated restricted problems that involve only the previously linked pairs, such as anomalous link discovery (Rattigan and Jensen 2005). We take the same approach but focus on predicting link strength rather than link existence. We differ from previous work in that we aim to exploit transactional information among nodes in order to improve prediction accuracy. O'Madadhain et al. (2005) also model transactional events, but they formulate a temporal link prediction task which tries to predict the occurrence of an event (e.g., coauthorship) in a time interval $[t, t + \delta]$ given the occurrence of events in the previous time interval (i.e., $< t$). Adamic and Adar (2003) also investigate the use of ancillary network information but with the goal of predicting social ties, instead of tie strength. They consider the web graph that connects students and faculty and use similarity-based features to predict link existence. The features consider similarity in homepage text and mailing list membership, as well as topological similarity in hyperlink structure.

Methodology

We define the link strength prediction problem as follows. Given a network graph $G = (V, E)$ with nodes V representing users and undirected edges E representing relationships (e.g., friendships) between pairs of users (e_{ij} : v_i and v_j are friends). Each edge e_{ij} is associated with a link weight l_{ij}, which indicates the strength of relationships between nodes v_i and v_j. The goal is to learn a predictive model of link weight l from labeled training data. In addition to the network graph G, we also have directed multigraphs $T = \{T_k = (V_k, E_k)\}$, where $V_k = V$ and $E_k \subseteq E$ (i.e., the edges represent transactions among pairs of linked nodes in E). The data also contain attributes on nodes V (e.g., gender, political views) and edges E_k (e.g., email subject). In this work, we consider the simpler binary task of predicting whether or not a relationship is strong.

Data

We evaluated our approach on data from the public Purdue Facebook network. Facebook is a popular online social network site with over 150 million members worldwide. Members create and maintain a personal profile page, which contains information about their views, interests, and friends, and can be listed as private or public. Friendship links are undirected and are formed through an invitation by one user along with a confirmation by the other. One key aspect of Facebook that we exploited for this work is the popularity of the "Top Friends" application. The application, which has more than 15 million users, allows users to nominate some of their friends as *top friends*. Among the users with the "Top Friends" application listed on their profile page, we can use *top friend* nominations as indicators of strong friendships.

We considered the set of 56061 Facebook users belonging to Purdue University network in March 2008. To be affiliated with a University network, users must have a valid email account within the appropriate domain (e.g., purdue.edu), thus the members consist of students, faculty, staff, and alumni. The public Purdue network comprised more than 3 million public friendship links among the members. Users had an average and median degree of 46 and 81 respectively (see Table 1).

In addition to the friendship graph, we considered three *transactional* graphs recording interactions among friends. First, the *wall graph* consists of links from users' public message boards on their profile pages. This message board is called the "wall" and is a place where other users can write small messages to their friends. From the wall postings in the period 03/01/07-03/01/08, we constructed directed links in the wall graph from the sender to the receiver. Second, the *picture graph* consists of links from users' public photo pages. The photo page can contain both photos of the member and their photo albums. The section that displays the photos of the member, consist of both photos posted by the member herself and photos posted in other users' albums that are tagged as containing the member. From these tagged photos, we constructed directed links from the album owner to the member. Third, the *group graph* consists of links calculated from the group membership information posted in the users' profile pages. Each "group" maintains a separate page reflecting some interest (e.g., friends of AAAI), and users who share that interest can become members of the group. If two users are members of the same group, we add an undirected link between the pair in the group graph.

From these data, we selected a random sample of 500 public users with top friends nominations. From this set of users, we considered all friendship links to other users. We restricted attention to links between pairs of users with values for ≥ 4 common attributes (to facilitate the attribute-similarity features used below). The final sample contained 8766 linked friends. Each pair (v_i, v_j) is labeled with a positive class label (*isTopFriend*) if node v_i has nominated node v_j as a top friend, and negative otherwise. The resulting target class contained 896 (10.2%) positive examples.

Features

Based on profile and graph information, we constructed 50 features to use for classification. The features can be grouped into four categories based on the information they consider in the data: *attribute-based*, *topological*, *transactional*, and *network-transactional*. The details of the features in those categories are given below.

Attribute-Based Features In the first category we constructed nine features which measure the similarity of the profile attributes on the pair of users. We created boolean match features on single-valued attributes like gender and relationship-status (i.e., 1 if pair of user values match, 0

Graph	Nodes	Edges	Median In Degree	Median Out Degree	Size of Largest Conn. Component	Avg Clustering Coefficient
Friendship	56,061	3,138,644	81	81	56,061	0.193
Wall	51,143	430,241	27	7	49893	0.195
Picture	35180	100,666	14	0	30938	0.333

Table 1: Graph statistics for public Purdue Facebook network

otherwise). On multi-valued attributes like networks and interested-in, we created integer features that counted the number of matches across the pair of lists (e.g., number of networks common to the pair of users). Finally, we constructed an aggregate similarity measure that summed the number of matches found across any of the eight profile attributes.

Topological Features In the second category we constructed six features that measure the connectivity of the users in the friendship graph. These features (along with the attribute-based features) are designed to be similar to the features used in current link prediction models. Two features are used to record the clustering coefficients of the pair of users. The remaining four features measure the degree and number of shared neighbors in the friendship graph, including the Jacquard coefficient and Adamic/Adar coefficient. For example:

$$Jacquard_{ij} = \frac{|N(i) \cap N(j)|}{|N(i) \cup N(j)|}$$

$$Adamic/Adar_{ij} = \sum_{k \in N(i) \cap N(j)} \frac{1}{log\ N(k)}$$

where $N(i)$ refers to a function that returns the neighbors of node v_i.

Transactional Features In the third category we constructed seven features that consider the transactional information between users (i.e., wall postings, picture postings and groups). These features only consider single edges in the transactional graphs; they do not consider the larger relational context of those transactions. For example, one feature counts the number of posts from node v_j on node v_i's wall; another counts the number of photos posted by node v_j and tagged as containing node v_i. However, the features do not consider the other transactional activity of nodes v_i and v_j. See Figure 1(a) for an illustration.

Network-Transactional Features For the last category we constructed 28 features that considered the transactional information between users, moderated by additional information in the local transactional network. See Figure 1(b) for an illustration. The idea here is to capture the transaction information between nodes, but represent it within the context of the larger network structure. For example, instead of just counting the number of wall posts from v_j to v_i, a network-transactional feature would also consider the number of posts made by node v_j to other nodes in the network (e.g., $\frac{|posts_{ji}|}{\sum_{k \in V} |posts_{jk}|}$).

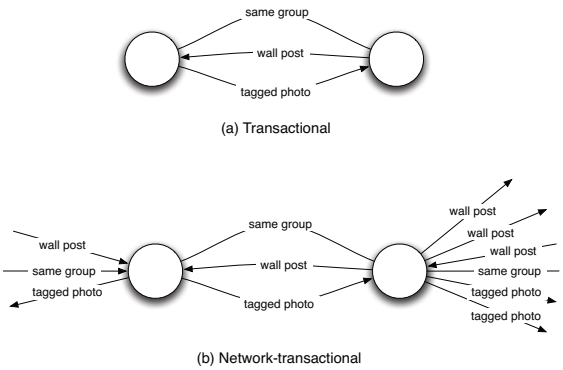

Figure 1: Example illustrating the network views considered by (a) transactional and (b) network-transactional features.

In this category, we also include features that record-clustering coefficients calculated from the transactional graphs (because they require more than just local knowledge about a single edge).

Models

For classification, we considered three supervised learning algorithms: logistic regression (LR), bagged decision trees (BDT), and naive Bayesian classifiers (NBC). The logistic regression model is an additive model used to predict the probability of a discrete event (e.g., the class label) given a set of explanatory variables. The model weights the impact of each feature with an estimated coefficient and is non-selective (i.e., it uses all possible features in the final model). Our bagged decision trees consisted of an ensemble of ten decision trees, each learned from a different random *pseudosample* drawn with replacement from the training set (see e.g., (Breiman 1996)). Decision trees are selective models that greedily choose a subset of the features that are deemed to be most relevant to the prediction task. Naive Bayesian classifiers model the target class probabilities using the class conditional distribution of each attribute and assuming conditional independence among the attributes. Naive Bayes classifiers are also non-selective models—all 50 features are used in the final model. For all three models, we used the algorithms from the Weka machine-learning library (Witten and Frank 2005) with default parameter settings.

Experimental Results

The experiments in this section demonstrate the utility of our method for automatically predicting strong friendships (e.g., *top friends*) based on attribute, network, and transactional information. We evaluate the models on real-world

Facebook data, using 10-fold cross validation, and report performance with area under the ROC curve (AUC). AUC measures the quality of rankings (by probability) produced by the model and is a more reasonable estimate of performance than accuracy on problems with skewed class distributions. We investigate overall performance of the models and use ablation studies to assess the influence of different features and graph data.

Overall Classification

In our first experiment, we used all 50 features during classification in order to measure the overall performance of each modeling technique. Figure 2 graphs the average AUC calculated from the 10 fold cross-validation trials. All three models achieve more than 80% average AUC, which indicates the model rankings are quite accurate (a random ranking would correspond to 50% AUC). Although the logistic regression (LR) and the Naive Bayes (NBC) models perform well, bagged decision trees (BDT) achieved the highest AUC of 87%. Bagged decision trees also exhibit the highest AUC in the ablation studies reported next.

Figure 3 graphs an example ROC curve, selected randomly from the 10 trials, to further illustrate the AUC results—bagged decision trees dominate in ROC space, indicating that the improvement is consistent throughout the ranking.

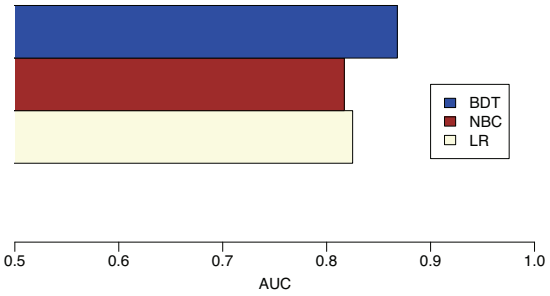

Figure 2: Classification results for logistic regression (LR), Naive Bayesian classifier (NBC), and bagged decision trees (BDT), using all 50 features.

Feature Category Comparison

Our second set of experiments consisted of ablation studies where we varied the sets of features available to the models for classification. We evaluated performance of each of the models with the features from each category separately: *attribute-based* (ATT), *topological* (TOP), *transactional* (TR), and *network-transactional* (NTR). Figure 4 graphs the average performance of each model for each category of features.

The three models performed similarly in each category. The *attribute-based* features result in the worst performance of all feature categories, with AUC close to random (0.5). Both the *topological* (based on the friendship graph) and the *transactional* (based on interaction between users) feature sets result in average performance, with AUCs in the range of $0.65 - 0.75$. However, the *network-transactional* features result in AUCs > 0.8 for all three models. On the bagged decision trees, the performance using only the network-transactional features accounts for 97% of the performance we observe using all features. These results indicate the influence of the network-transactional features for our classification task. Moreover, their improvement over the transactional features indicates that is important to consider the interactions in the context of the node behavior in the larger transactional network.

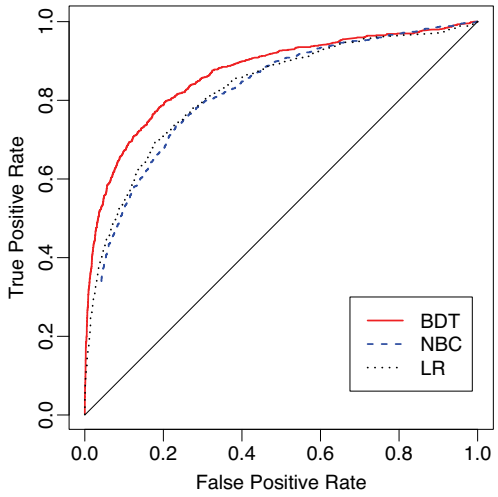

Figure 3: Example ROC curve comparing logistic regression (LR), Naive Bayesian classifier (NBC), and bagged decision trees (BDT), using all 50 features.

Figure 5 graphs an example ROC curve, selected randomly from the 10 trials, showing BDT model performance using features from each category. The model learned with NTR features dominates in ROC space, indicating again that the gains are due to consistent improvement throughout the ranking.

Network comparison

Our third set of experiments comprised another set of ablation studies, where in this case we varied the network graph available for classification. We evaluated performance of each of the models with all features types but only generating those features which apply to a particular network: the friendship network, the group membership network, the photo tagging network, and the wall posting network. In each case, we included the same set of attribute-based features, which consider the profile information on pairs of users. Figure 6 graphs the average performance of each model for each category of feature. Again the models perform similarly in each ablation case. Overall, the wall network appears to offer the most information, resulting in AUCs close to 80%. The friendship graph results in average performances (i.e., around 70% AUC). The picture and group graph produce the lowest performances, with AUCs of less than 65%.

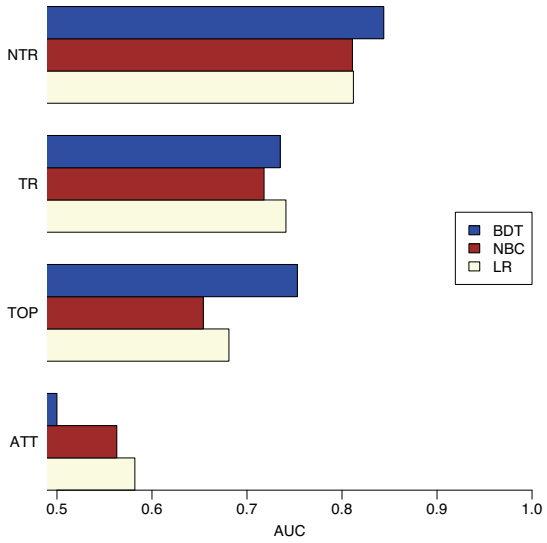

Figure 4: Classification results when using the features from each category separately: attribute-based (ATT), topological (TOP), transactional (TR), and network-transactional (NTR).

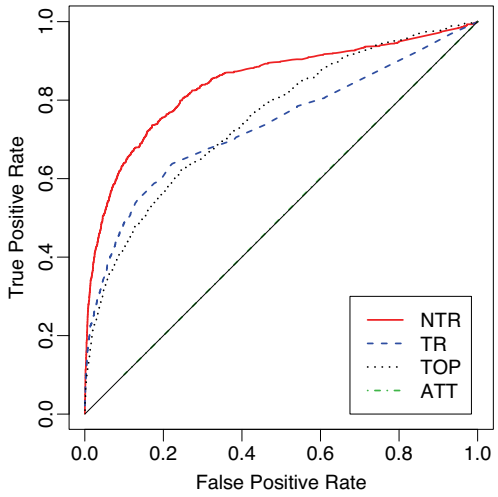

Figure 5: Example ROC curve comparing BDT models with features from each category separately: attribute-based (ATT), topological (TOP), transactional (TR), and network-transactional (NTR).

Figure 7 graphs an example ROC curve, selected randomly from the 10 trials, showing BDT model performance for each network. The model learned from the wall features mostly dominates the ROC space, indicating that the improvement is consistent through all but the bottom of the ranking.

We hypothesized that the group interactions would be least useful (because there is often no direct interaction among group members) and that the picture interactions would be most informative to the models (since tagged photos indicate not only that the two users were physically together at the time of the photo, but also that one of them had taken the time to post, view, and tag the photo). Consequently, we hypothesized that the wall interactions would be important but somewhat less informative than the picture interactions. The poor performance of the models when we restrict attention to the picture network appears to contradict this hypothesis. However, one explanation for the poor performance is the sparsity of the picture graph compared to the wall graph. Although 27.9% of user pairs in our sample have at least one wall link between them (i.e., a posting in either direction), only 3.7% of the user pairs have a picture link between them. This means that there will be few non-zero values for the picture-based features and could be a reason that the models perform poorly. Indeed, it indicates that at most 36% of the positive examples would have non-zero values for the features. Given the feature ranking results in the next section, if there were more picture data available, it is likely that performance would improve significantly.

Feature ranking

Our final set of experiments investigated the relative importance of each of the 50 features. We considered each feature independently and calculated their ability to distinguish between the positive and negative examples using Information Gain (IG) and the Chi-Square statistic (χ^2). For each feature, we recorded the ranking assigned by each measure and computed the average overall ranking. Table 2 lists the top 15 features, along with their category, description of the feature, and the resulting rankings. Note that the function $N_W(i)$ returns the set of unique users to which node v_i has posted wall comments. We define $N_P(i)$ analogously for the picture taggings.

Of the top 15 features, twelve are *network-transactional* features, and the other three are*transactional* features. These results lend further support to the claim that considering transactional information in the context of the larger network structure is important when designing features. We also note, that of the top 15 features, twelve features use the information in the wall graph and three features use information from the picture graph. This indicates the importance of the using the transactional network data, as opposed to the social network recorded in the friendship graph.

Based on the feature rankings, we conducted additional experiments where the models were only supplied the top 10 and top 20 features as determined by the ranking. The restricted feature set resulted in overall model performance around 80% AUC. This indicates that the bagged decision trees, which achieved 87% AUC with all the features and 84% AUC with all the NTR features, used additional features to make fine grained distinctions among the users pairs and make more accurate predictions.

Conclusion

In this paper we formulate and investigate a new task in social network mining: *link strength prediction*. To date, work on link prediction has focused primarily on the task of

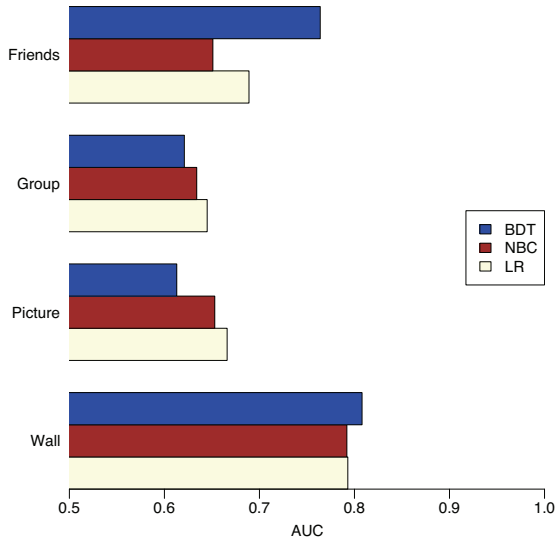

Figure 6: Classification results when using the features from each network graph separately: friendship, group membership, picture tagging, and wall postings.

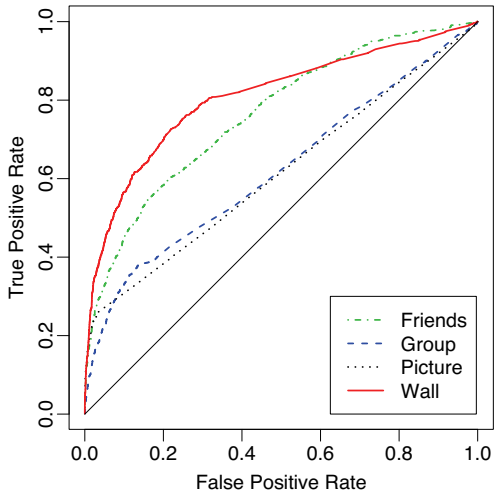

Figure 7: Example ROC curve comparing BDT models with features from each network graph separately: friendship, group membership, picture tagging, and wall postings.

predicting link existence. However, in domains where the underlying social network is collected automatically (e.g., online friendship networks) the underlying graph generally contains more spurious (e.g., acquaintance) relationships than previous data that was collected in a targeted manner. When there is additional data from *transactional* networks that contains low-level interactions among the users (e.g., text messages), this information can be used to predict which social ties are strongest and identify possibly spurious ties. This is the focus of our work.

We outlined a supervised learning approach for this task and evaluated our methods on real-world (public) data from the Purdue Facebook network. We compared three models and showed that bagged decision trees perform best overall, achieving 87% AUC. We evaluated the importance of features from four different categories and showed that *network-transactional* features had the largest impact on the overall performance of the models. The experimental results indicate that (1) transactional events are useful for predicting link strength, and (2) it is necessary to consider the transactional events in the context of user behavior within the larger social network. This success of network-transactional features is likely due to the same reasons that term-frequency-inverse document-frequency (TF-IDF; Salton & Buckley 1988) is a useful measure for ranking words in documents—a word that occurs frequently in a document is less discriminative if it occurs in many documents. Similarly, a transaction link between two users is less likely to indicate a strong relationship when the users have interacted with many other users.

In addition, we evaluated the influence of each of the different networks (friendship, wall, picture, group) on prediction accuracy and showed that the wall network had the largest impact on model performance. This is additional evidence in support of using transactional information for prediction. A ranking of individual features by their ability to discriminate the class also showed a preponderance of wall features in the top of the ranking (12 out of the top 15). The picture network was not as useful for predicting strong relationships, but this is likely due to the relative sparsity of these links in the data.

This work presents our initial attempts to use transactional information to predict link strength. Although transactional events generally occur over time, the features we used in this work did not consider the temporal aspect of the data (e.g., time stamps on wall postings). Our future work will consider ways to incorporate temporal patterns of interaction among the users in our models. We will focus on identifying influential temporal motifs (e.g., a burst of transactions in particular time window) for use as relational features. In addition, we will address the more general link-strength prediction task by formulating a latent variable model where link weights between pairs of nodes are hidden variables that change over time and affect the strength of relationships between the incident nodes.

References

Adamic, A., and Adar, E. 2003. Friends and neighbors on the web. social networks. *Social Networks* 25(2):211230.

Brandes, U. 2001. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology* 25:163–177.

Breiman, L. 1996. Bagging predictors. *Machine Learning* 24(2):123–140.

Chakrabarti, S.; Dom, B.; and Indyk, P. 1998. Enhanced hypertext categorization using hyperlinks. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 307–318.

Domingos, P., and Richardson, M. 2001. Mining the network value of customers. In *Proceedings of the 7th ACM*

Rnk	Cat.	Description	IG	χ^2	Avg
1	NTR	$\begin{cases} \frac{1}{\lvert N_W(i)\rvert+\lvert N_W(j)\rvert} & if\ \lvert posts_{ij}\rvert>0 \\ & \wedge \lvert posts_{ji}\rvert>0 \\ 0 & otherwise \end{cases}$	1	1	1.0
2	NTR	$\begin{cases} \frac{1}{\lvert N_W(j)\rvert} & if\ \lvert posts_{ji}\rvert > 0 \\ 0 & otherwise \end{cases}$	2	2	2.0
3	NTR	$\frac{\lvert posts_{ij}\rvert}{\sum_{k\in V}\lvert posts_{ik}\rvert}$	3	3	3.0
4	TR	$\lvert posts_{ji}\rvert$	5	4	4.5
5	TR	$\lvert posts_{ij}\rvert$	4	5	4.5
6	NTR	$\begin{cases} \frac{1}{\lvert N_W(i)\rvert} & if\ \lvert posts_{ij}\rvert > 0 \\ 0 & otherwise \end{cases}$	7	6	6.5
7	NTR	$\frac{\lvert posts_{ji}\rvert}{\sum_{k\in V}\lvert posts_{jk}\rvert}$	6	8	7.0
8	NTR	$\frac{\lvert posts_{ij}\rvert+\lvert posts_{ji}\rvert}{\sum_{m\in V}\lvert posts_{im}\rvert+\sum_{k\in V}\lvert posts_{jk}\rvert}$	8	7	7.5
9	TR	$\lvert posts_{ij}\rvert+\lvert posts_{ji}\rvert$	9	9	9.0
10	NTR	$\begin{cases} \frac{1}{\lvert N_P(j)\rvert} & if\ \lvert pics_{ji}\rvert > 0 \\ 0 & otherwise \end{cases}$	12	10	11.0
11	NTR	$\frac{\lvert posts_{ij}\rvert}{\sum_{k\in V}\lvert posts_{kj}\rvert}$	10	13	11.5
12	NTR	$\begin{cases} \frac{1}{\lvert N_P(i)\rvert+\lvert N_P(j)\rvert} & if\ \lvert pics_{ij}\rvert>0 \\ & \wedge \lvert pics_{ji}\rvert>0 \\ 0 & otherwise \end{cases}$	13	11	12.0
13	NTR	$\frac{\lvert posts_{ji}\rvert}{\sum_{k\in V}\lvert posts_{kj}\rvert}$	11	14	12.5
14	NTR	$\begin{cases} \frac{1}{\lvert N_P(i)\rvert} & if\ \lvert pics_{ij}\rvert > 0 \\ 0 & otherwise \end{cases}$	15	12	13.5
15	NTR	$\frac{\lvert posts_{ij}\rvert+\lvert posts_{ji}\rvert}{\sum_{m\in V}\lvert posts_{mj}\rvert+\sum_{k\in V}\lvert posts_{ki}\rvert}$	14	16	15.0

Table 2: Feature rankings.

SIGKDD International Conference on Knowledge Discovery and Data Mining, 57–66.

Girvan, M., and Newman, M. E. J. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99(12):7821–7826.

Granovetter, M. 1983. The strength of weak ties: A network theory revisited. *Sociological Theory* 1:201–233.

Harris, K. 2008. The national longitudinal study of adolescent health (add health), waves i & ii, 1994 1996; wave iii, 20012002 [machine-readable data file and documentation]. *Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill.*

Hasan, M.; Chaoji, V.; Salem, S.; and Zaki, M. 2005. Link prediction using supervised learning. In *In Proceedings of the Workshop on Link Discovery: Issues, Approaches and Applications*.

Jensen, D.; Neville, J.; and Gallagher, B. 2004. Why collective inference improves relational classification. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 593–598.

Kashima, H., and Abe, N. 2006. A parameterized probabilistic model of network evolution for supervised link prediction. In *In Proceedings of the Sixth IEEE International Conference on Data Mining*.

Kumar, R.; Novak, J.; and Tomkins, A. 2006. Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 611–617. New York, NY, USA: ACM.

Lanckriet, G.; Bie, T. D.; Cristianini, N.; Jordan, M.; and Noble, W. 2004. A statistical framework for genomic data fusion. *Bioinformatics* 20(16):2626–2635.

Liben-Nowell, D., and Kleinberg, J. 2004. The link prediction problem for social networks. In *In Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM)*.

McGovern, A.; Friedland, L.; Hay, M.; Gallagher, B.; Fast, A.; Neville, J.; and Jensen, D. 2003. Exploiting relational structure to understand publication patterns in high-energy physics. *SIGKDD Explorations* 5(2):165–172.

McPherson, M.; Smith-Lovin, L.; and Cook, J. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27:415–445.

Neville, J.; Şimşek, O.; Jensen, D.; Komoroske, J.; Palmer, K.; and Goldberg, H. 2005. Using relational knowledge discovery to prevent securities fraud. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 449–458.

O'Madadhain, J.; Hutchins, J.; and Smyth, P. 2005. Prediction and ranking algorithms for event-based network data. *SIGKDD Explorations* 7(2):2330.

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

Robins, G.; Snijders, T.; Wang, P.; Handcock, M.; and Pattison, P. 2007. Recent developments in exponential random graph (p*) models for social networks. *Social Networks* 29:192–215.

Sharan, U., and Neville, J. 2008. Temporal-relational classifiers for prediction in evolving domains. In *Proceedings of the 8th IEEE International Conference on Data Mining*.

Taskar, B.; Wong, M. F.; Abbeel, P.; and Koller, D. 2003. Link prediction in relational data. In *In Proceedings of the Neural Information Processing Systems Conference (NIPS03)*.

Wasserman, S., and Faust, K. 1994. *Social Network Analysis: Methods and Applications*. Cambridge, UK: Cambridge University Press.

Witten, I., and Frank, E. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann.

Xu, Z.; King, I.; and Lyu, M. 2007. Web page classification with heterogeneous data fusion. In *Proceedings of the World Wide Web Conference*.