Modeling Relationship Strength in Online Social Networks

Rongjing Xiang
Computer Science
Department
Purdue University
West Lafayette, IN 47907
rxiang@cs.purdue.edu

Jennifer Neville
Departments of Computer
Science and Statistics
Purdue University
West Lafayette, IN 47907
neville@cs.purdue.edu

Monica Rogati
LinkedIn
Mountain View, CA 94043
mrogati@linkedin.com

ABSTRACT

Previous work analyzing social networks has mainly focused on binary friendship relations. However, in online social networks the low cost of link formation can lead to networks with heterogeneous relationship strengths (e.g., acquaintances and best friends mixed together). In this case, the binary friendship indicator provides only a coarse representation of relationship information. In this work, we develop an unsupervised model to estimate relationship strength from interaction activity (e.g., communication, tagging) and user similarity. More specifically, we formulate a link-based latent variable model, along with a coordinate ascent optimization procedure for the inference. We evaluate our approach on real-world data from Facebook and LinkedIn, showing that the estimated link weights result in higher autocorrelation and lead to improved classification accuracy.

Categories and Subject Descriptors

H.4.m [**Information Systems**]: Miscellaneous

General Terms

Algorithms, Design

Keywords

Social networks, link prediction, latent variable models

1. INTRODUCTION

Recent research on analyzing social networks has demonstrated that relational patterns of *homophily* [14] can be exploited to improve predictive models of both link structure and behavior. For example, researchers have found that network connectivity and attribute similarity can improve link prediction models [13, 19]. Also, researchers have found that relational ties can improve behavior prediction in tasks such as as fraud detection [15] and viral marketing [6].

However, much of this past work has focused on social networks with *binary* relational ties (e.g., friends or not). These binary indicators provide only a coarse indication of the nature of the relationship. Due to the low-cost of friendship identification in online social networks and the variance of link information in electronic communication networks, the resulting networks often contain both strong and weak ties—with little or no information to differentiate between the two ends of the spectrum. Since pairs of individuals with *strong ties* (e.g., close friends) are likely to exhibit greater similarity than those with *weak ties* (e.g., acquaintances) [8], treating all relationships as equal will increase the level of noise in the learned models and likely lead to degradation in performance. 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 [17].

Fortunately, online social networks (OSNs) often consist of more than just a record of social network ties. Typically online communities contain ancillary *interaction* information among the users that can be used for modeling. Indeed, almost every OSN provides infrastructure for the formation and maintenance of communities over time by facilitating communication and transfer of information. The system thus keeps a record of low-level interactions among related people and can be used to identify which linked members are close friends/colleagues, as opposed to acquaintances. Facebook users, for instance, each have a *Wall* page as part of their profile, where friends can post messages. While a particular user may have hundreds of friends, due to resource constraints it is likely that she communicates more frequently with friends compared to acquaintances. Similarly, LinkedIn users can request and/or write recommendations for other users in the system. Although a given user might maintain connections to hundreds of professionals, they will only write recommendations for those with whom they are most familiar.

In this work, we propose a model to infer relationship strength based on profile similarity and interaction activity, with the goal of automatically distinguishing strong relationships from weak ones. Recently, interaction data has been used to predict relationship strength [11, 7] but this work only considered two levels of relationship strength, namely weak and strong relationships. In addition, this past work focused on supervised learning methods, which requires human annotation of link strength (e.g., friendship rating [7] or top friend nomination [11]). We focus instead on developing a richer model that can represent the full spectrum of relationship strength, from weak to strong, and propose an

WWW2010, April 26-30, 2010, Raleigh, North Carolina.
ACM 978-1-60558-799-8/10/04.
.

unsupervised method to infer a continuous-valued relationship strength for links.

More specifically, we formulate a latent variable model to infer (hidden) relationship strengths and develop a coordinate ascent optimization procedure for inference. From the modeling perspective, a unique characteristic of our model is that it distinguishes interaction activity from users' profile data, and integrates these two types of information by considering the relationship strength to be the hidden effect of user profile similarities, as well as the hidden cause of the interactions between users. This naturally leads to a latent variable model which captures the causality of the underlying social process. Furthermore, in view of the fact that the user profile data is relatively comprehensive, stable, and available to OSN service providers, while the interaction data is usually sparse, temporal and in many cases predictions of future interactions are of interest, we take a discriminative approach to modeling the profile similarities, and a generative approach to the interactions. The resulting hybrid model has the benefit of both accurate estimation of relationship strengths based on the discriminative modeling of profile data, and flexible handling/predicting missing interaction data.

In addition, our approach is also scalable. For the proposed statistical model, we implement a principled optimization scheme to infer the parameter values and the relationship strengths for a set of user pairs. The optimization algorithm proves to converge fast in our experiments. The learned parameter values can be applied to estimate the relationship strength for a new queried pair in constant time, which is suitable for real time application (e.g., online prediction).

Besides its immediate implications for social science applications, the estimation of relationship strength can also be used to improve the range and performance of various aspects of online social networks, including:

- **Link prediction:** In LinkedIn and Facebook, the system automatically suggests new connections to users. With the estimated relationship strengths for pairs of users within a certain distance (e.g., two-hop away in the network, working in the same company, etc.), this task could be easily improved by suggesting those people with top relationship strengths to users.

- **Item recommendation:** In general, any automatic recommendation service provided by OSNs could be improved with the estimated relation strengths, since a person's affinities and preferences are more likely to coincide with those who they are strongly related. For example, in LinkedIn, when recommending groups that a user might want to join, or news articles they might want to read, the activities of related people are highly predictive of a recommendation's success.

- **Newsfeeds:** Newsfeeds (i.e., real-time updates about status change, activities, new posts or other stories from contacts) is an important feature implemented in OSNs such as Facebook, LinkedIn and Twitter. When building an online member's personalized newsfeed about their connections, prioritizing the updates by relationship strength could be more beneficial to the user by removing or downplaying updates from spurious contacts.

- **People search:** By ranking search results according to relationship strengths between the query sender and the discovered people, the user is likely to find an accessible person more quickly.

- **Visualization:** The applications of visualizing people's local social network could be improved by scaling/shading links according to the estimated relationship strengths.

The rest of the paper is organized as follows. We outline the details of the model and the estimation algorithm in Section 2. In Section 3, we evaluate our approach on real-world data from LinkedIn and Facebook by showing that autocorrelation in the estimated relationship-strength graph is higher than any alternative graph formed from various aspects of the raw data. Furthermore, we demonstrate the utility of the inferred relationship strengths in a number of collective classification and recommendation tasks. In Section 5, we conclude and point out some directions for future work.

2. LATENT VARIABLE MODEL

One key assumption underlying our model is the theory of *homophily* from sociology [14]. The homophily principle postulates that people tend to form ties with other people who have similar characteristics (i.e., the tendency of like to associate with like). Moreover, it is likely that the stronger the tie, the higher the similarity [8]. Previous studies have shown that homophily is ubiquitous in social networks (see [14]). In online social networks therefore, we can model the relationship strength as a *hidden effect* of nodal profile similarities. Such profile attributes include, for instance, the schools and companies the users attended, the online groups that they joined, the geographic locations that they belong to, etc.

Furthermore, we assume that the relationship strength directly impacts the nature and frequency of online interactions between a pair of users. Since each user has a finite amount of resources (e.g., time) to use in the formation and maintenance of relationship, it is likely that they direct these resources towards the relationships that they deem more important [5]. Such interactions could be, for example, profile viewing activities between the pair of users, connection establishment, picture tagging, etc. The stronger the relationship, the higher likelihood that a certain type of interaction will take place between the pair of users. In this way, we model the relationship strength as the *hidden cause* of user interactions. Conditioning on the relationship strength, the interaction variables are independent of each other.

Formally, let $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ denote the profile vectors of two individuals i and j, and let $y_t^{(ij)}, t = 1, 2, \ldots, m$ be the occurrences of m different interactions considered between i and j. Then we define $z^{(ij)}$ to be the latent relationship strength between i and j and model the influence of $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ on $z^{(ij)}$, as well as the influence of $z^{(ij)}$ on $y_t^{(ij)}, t = 1, 2, \ldots, m$.

We illustrate the general model using the directed graphical model representation in Figure 1. The full model can be viewed as a hybrid of discriminative and generative models—the upper part is discriminative ($p(Z|X)$), while the lower part ($p(Y, Z)$) is generative. Our model represents the likely causal relationships among these variables by modeling the conditional dependencies, so that the joint distribution decomposes as follows:

$$
\begin{aligned}
&P(z^{(ij)}, \mathbf{y}^{(ij)}|x^{(i)}, x^{(j)}) \qquad (1)\\
=&P(z^{(ij)}|x^{(i)}, x^{(j)}) \prod_{t=1}^{m} P(y_t^{(ij)}|z^{(ij)})
\end{aligned}
$$

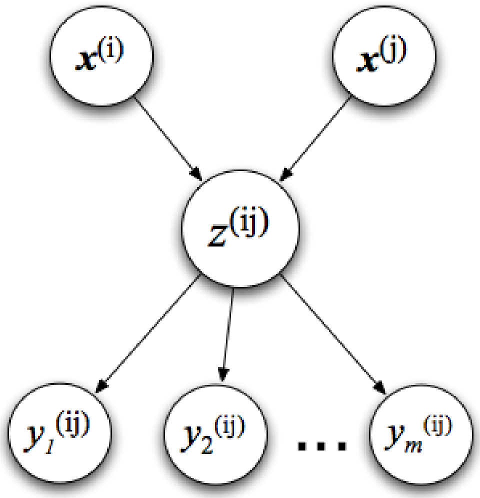


Figure 1: Graphical model representation of the general relationship strength model.

Although the relationship-strength variable z summarizes the similarities and interactions between a pair of people, its value is not directly observed in the data (and it would be difficult, if not impossible, to collect from online users). As such, it makes sense to treat z as a latent (i.e., hidden) variable in the model, which we will estimate for each pair of people (along with the values of model parameters) so as to maximize the overall observed data likelihood.

In general, our model can be applied to infer either directed or undirected relationship strengths, depending on the way how we specify the profile similarity and the interactions for each pair. In this work, we infer directed relationship strengths, i.e., the estimated $z^{(ij)}$ could possibly be different than $z^{(ij)}$, since we consider directed interactions in the data (e.g., while i has posted on j's wall, j might have not have posted on i's).

2.1 Model specification

The general latent variable model of relationship strength can be instantiated in an appropriate way, depending on domain-specific availability and interpretation of attributes and interactions. In this work, we adopt the widely-used Gaussian distribution to model the conditional probability of the relationship strength given profile similarities.

Let $s_k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ $(k = 1, 2, \ldots, n)$ denote a set of similarity measures taken on pairs of nodes i,j. Then the dependency between z^{ij} and $\mathbf{x}^{(i)}, \mathbf{x}^{(j)}$ is as follows:

$$P(z^{(ij)}|\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \mathcal{N}(\mathbf{w}^T \mathbf{s}(\mathbf{x}^{(j)}, \mathbf{x}^{(j)}), v) \qquad (2)$$

where $\mathbf{s}$ is a similarity vector calculated based on $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$; $\mathbf{w}$ is an n-dimensional weight vector to be estimated, and v is the variance in Gaussian model, which is configured to be 0.5 in our experiments. To reflect this, we replace $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ with $\mathbf{s}^{(ij)}$ in Figure 2.

In the proposed model, the probability distribution of each $y_t^{(ij)}$ is conditionally independent given $z^{(ij)}$. For this work, we model all interactions as binary variables, regardless of the frequency of interaction due to the sparsity of the data. For example, the variable may denote whether a user i has posted on j's wall.

Furthermore, to increase the accuracy of the model, we introduce a set of auxiliary variables $a_{t1}^{(ij)}, a_{t2}^{(ij)}, \ldots, a_{tl_t}^{(ij)}$ for each interaction t, as shown in Figure 2. Such variables capture auxiliary causes of the interactions which are independent of the relationship strength. For example, the total number of pictures that a user has tagged represents their intrinsic tendency to tag pictures, and hence it could moderate the effect of relationship strength on interactions with a specific user.

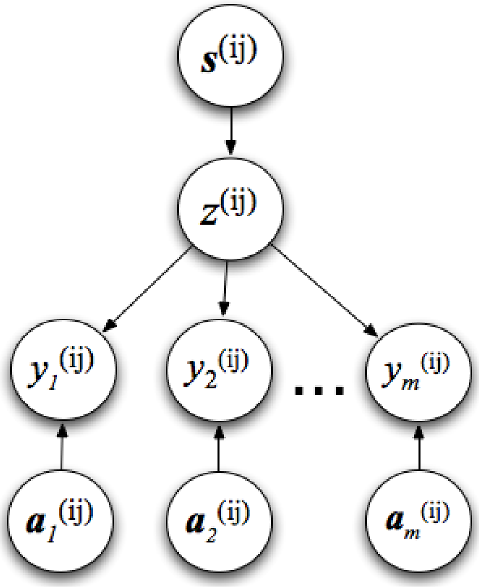


Figure 2: Graphical model representation of the specific instantiation described in Sec.2.1.

We use a logistic function to model the conditional probability of $y_t^{(ij)}$ given $z^{(ij)}$ and $\mathbf{a}_t^{(ij)}$:

$$
\begin{aligned}
&P(y_t^{(ij)} = 1|z^{(ij)}, \mathbf{a}_t^{(ij)}) \qquad (3)\\
=&\frac{1}{1 + e^{-(\theta_{t1} a_{t1}^{(ij)} + \theta_{t2} a_{t2}^{(ij)} + \ldots + \theta_{tl} a_{tl}^{(ij)} + \theta_{tl+1} z^{(ij)} + b)}}
\end{aligned}
$$

where $\boldsymbol{\theta}_t = [\theta_{t1}, \theta_{t2}, \ldots, \theta_{tl}, \theta_{tl+1}]^T$ are the set of parameters to be estimated. To make the notation more compact, we define $\mathbf{u}_t^{(ij)} = \begin{bmatrix} \mathbf{a}_t^{(ij)} \\ z^{(ij)} \end{bmatrix}$, so that:

$$P(y_t^{(ij)} = 1|\mathbf{u}_t^{(ij)}) = \frac{1}{1 + e^{-(\boldsymbol{\theta}_t^T \mathbf{u}_t^{(ij)} + b)}} \qquad (4)$$

In general, we could apply other appropriate generalized linear models for interaction variables without adding difficulty to the inference (the objective function will still be concave and the optimization procedure will remain unchanged). For example, poisson regression could be used if the interaction is represented as count data.

Finally, to avoid over-fitting, we put L2 regularizers on the parameters $\mathbf{w}$ and θ, which can be regarded as Gaussian priors:

$$P(\mathbf{w}) \propto e^{-\frac{\lambda_w}{2}\mathbf{w}^T\mathbf{w}} \tag{5}$$

$$P(\boldsymbol{\theta}_t) \propto e^{-\frac{\lambda_\theta}{2}\boldsymbol{\theta}_t^T\boldsymbol{\theta}_t}, t = 1, 2, \ldots, m \tag{6}$$

The data are represented as N samples of user pairs, denoted by $\mathcal{D} = \{(i_1, j_1), (i_2, j_2), \ldots, (i_N, j_N)\}$. During training, the variables $\mathbf{x}^{(ij)}, \mathbf{y}^{(ij)}$ and $\mathbf{a}_t^{(ij)}$, $((ij) \in \mathcal{D}, t = 1, 2, \ldots, m)$ are all visible. Since the attribute similarities are pre-calculated based on the $\mathbf{x}'s$, to simplify the notation, we define $s^{(ij)} = s(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$. Given all the observed variables, based on Eq. (1), the joint probability is as follows:

$$\begin{aligned}
&P(\mathcal{D}|\mathbf{w}, \boldsymbol{\theta})P(\mathbf{w}, \boldsymbol{\theta}) \qquad (7)\\
&= \left(\prod_{(i,j)\in\mathcal{D}} P(z^{(ij)}, \mathbf{y}^{(ij)}|x^{(i)}, x^{(j)}, \mathbf{w}, \boldsymbol{\theta})\right) P(\mathbf{w})P(\boldsymbol{\theta})\\
&= \prod_{(i,j)\in\mathcal{D}} P(z^{(ij)}|x^{(i)}, x^{(j)}, \mathbf{w}) \prod_{t=1}^{m} P(y_t^{(ij)}|z^{(ij)}, \boldsymbol{\theta}_t)P(\mathbf{w})P(\boldsymbol{\theta}_t)\\
&\propto \prod_{(i,j)\in\mathcal{D}} \left(e^{-\frac{1}{2v}\left(\mathbf{w}^T\mathbf{s}^{(ij)} - z^{(ij)}\right)^2} \prod_{t=1}^{m} \frac{e^{-(\boldsymbol{\theta}_t^T\mathbf{u}^{(ij)}+b)(1-y_t^{(ij)})}}{1 + e^{-(\boldsymbol{\theta}_t^T\mathbf{u}_t^{(ij)}+b)}}\right)\\
&\quad \cdot e^{-\frac{\lambda_w}{2}\mathbf{w}^T\mathbf{w}} \prod_{t=1}^{m} e^{-\frac{\lambda_\theta}{2}\boldsymbol{\theta}_t^T\boldsymbol{\theta}_t}
\end{aligned}$$

2.2 Inference

In general, estimation of a latent variable model can be done in two different ways. First, we can infer the distribution of the latent variable z, and find point estimates of the parameters $\hat{\mathbf{w}}, \hat{\boldsymbol{\theta}}$ so as to maximize the joint likelihood $P(\mathbf{y}, \hat{\mathbf{w}}, \hat{\boldsymbol{\theta}}|\mathbf{x})$ (i.e., the latent variable $\mathbf{z}$ is integrated out). This type of approach usually involves an iterative expectation maximization (EM) algorithm. Second, we can treat the latent variable as a parameter—namely, find point estimates $\hat{\mathbf{w}}, \hat{\boldsymbol{\theta}}, \hat{\mathbf{z}}$ that maximize the likelihood $P(\mathbf{y}, \hat{\mathbf{z}}, \hat{\mathbf{w}}, \hat{\boldsymbol{\theta}}|\mathbf{x})$. In this work, we will use the latter approach since integration over the latent variable z involved in the E-step of an EM algorithm would be intractable. We leave the investigation of an approximate EM algorithm, to estimate the distribution of the latent variables z, as future work.

Taking the logarithm of Eq. (7), we get the data log-likelihood:

$$\begin{aligned}
&\mathcal{L}(z^{(\{(i,j)\in\mathcal{D}\})}, \mathbf{w}, \boldsymbol{\theta}_t) \qquad (8)\\
&= \sum_{(ij)\in\mathcal{D}} -\frac{1}{2v}\left(\mathbf{w}^T\mathbf{s}^{(ij)} - z^{(ij)}\right)^2\\
&+ \sum_{(ij)\in\mathcal{D}} \sum_{t=1}^{m} -(1 - y_t^{(ij)})(\boldsymbol{\theta}_t^T\mathbf{u}_t^{(ij)} + b) - \log\left(1 + e^{-\left(\boldsymbol{\theta}_t^T\mathbf{u}_t^{(ij)}+b\right)}\right)\\
&- \frac{\lambda_{\mathbf{w}}}{2}\mathbf{w}^T\mathbf{w} - \sum_{t=1}^{m} \frac{\lambda_\theta}{2}\boldsymbol{\theta}_t^T\boldsymbol{\theta}_t + C
\end{aligned}$$

Note that in Eq. (8), both the quadratic terms and the logarithm of logistic function are concave. Since the sum of concave functions is concave, the function $\mathcal{L}$ is concave. Therefore, a gradient-based method will allow us to optimize over the parameters $\mathbf{w}, \boldsymbol{\theta}_t$ $(t = 1, 2, \ldots, m)$, and the latent variables $z^{(ij)}, (ij) \in \mathcal{D}$ to find the maximum of $\mathcal{L}$. Below, we derive a coordinate ascent method for the optimization.

The coordinate-wise gradients are:

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial z^{(ij)}} =& \frac{1}{v}(\mathbf{w}^T\mathbf{x}^{(ij)} - z^{(ij)})\\
&+ \sum_{t=1}^{m} \left(y_t^{(ij)} - \frac{1}{1 + e^{-(\boldsymbol{\theta}_t^T\mathbf{u}_t^{(ij)}+b)}}\right)\theta_{t,l_t+1}
\end{aligned} \tag{9}$$

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}_t} = \sum_{(ij)\in\mathcal{D}} \left(y_t^{(ij)} - \frac{1}{1 + e^{-(\boldsymbol{\theta}_t^T\mathbf{u}_t^{(ij)}+b)}}\right)\mathbf{u}_t^{(ij)} - \lambda_{\boldsymbol{\theta}}\boldsymbol{\theta}_t \tag{10}$$

$$\frac{\partial\mathcal{L}}{\partial\mathbf{w}} = \frac{1}{v} \sum_{(ij)\in\mathcal{D}} (z^{(ij)} - \mathbf{w}^T\mathbf{s}^{(ij)})\mathbf{s}^{(ij)} - \lambda_w\mathbf{w} \tag{11}$$

A coordinate ascent optimization scheme will update $\mathbf{w}$, $z^{(ij)}$ and $\boldsymbol{\theta}_t$ iteratively until convergence. For $z^{(ij)}$ and $\boldsymbol{\theta}_t$, since the root of Eq. (9) and Eq. (10) cannot be found analytically, we use the following Newton-Raphson updates in each iteration:

$$z^{(ij)\text{new}} = z^{(ij)\text{old}} - \frac{\partial\mathcal{L}}{\partial z^{(ij)}} / \frac{\partial^2\mathcal{L}}{\partial\left(z^{(ij)}\right)^2} \tag{12}$$

$$\boldsymbol{\theta}_t^{\text{new}} = \boldsymbol{\theta}_t^{\text{old}} - \frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}_t} / \frac{\partial^2\mathcal{L}}{\partial\boldsymbol{\theta}_t\partial\boldsymbol{\theta}_t^T} \tag{13}$$

where the 2nd order derivatives are given by:

$$\frac{\partial^2\mathcal{L}}{\partial\left(z^{(ij)}\right)^2} = -\frac{1}{v} - \sum_{t=1}^{m} \frac{\theta_{t,l_i+1}^2 e^{-(\boldsymbol{\theta}_t^T\mathbf{u}_t^{(ij)}+b)}}{\left(1 + e^{-(\boldsymbol{\theta}_t^T\mathbf{u}_t^{(ij)}+b)}\right)^2} \tag{14}$$

$$\frac{\partial^2\mathcal{L}}{\partial\boldsymbol{\theta}_t\partial\boldsymbol{\theta}_t^T} = -\sum_{(ij)\in\mathcal{D}} \frac{e^{-(\boldsymbol{\theta}_t^T\mathbf{u}_t^{(ij)}+b)}}{\left(1 + e^{-(\boldsymbol{\theta}_t^T\mathbf{u}_t^{(ij)}+b)}\right)^2}\mathbf{u}_t^{(ij)}\mathbf{u}_t^{(ij)T} - \lambda_\theta\mathbf{I} \tag{15}$$

For $\mathbf{w}$, the root of (11) can be found analytically as in usual ridge regression:

$$\mathbf{w}^{\text{new}} = \left(\lambda_w\mathbf{I} + \mathbf{S}^T\mathbf{S}\right)^{-1}\mathbf{S}^T\mathbf{z} \tag{16}$$

where $\mathbf{S} = \begin{bmatrix} \mathbf{s}^{(i_1j_1)} \\ \mathbf{s}^{(i_2j_2)} \\ \vdots \\ \mathbf{s}^{(i_Nj_N)} \end{bmatrix}$, and $\mathbf{z} = \begin{bmatrix} z^{(i_1j_1)} \\ z^{(i_2j_2)} \\ \vdots \\ z^{(i_Nj_N)} \end{bmatrix}$.

An overview of the optimization procedure is given in Table 1.

During testing, for a new pair of users (i, j), the learned model can be applied in two ways. First, if both user attributes $\mathbf{x}^{(i)}, \mathbf{x}^{(j)}$ and their interactions $y_1^{(ij)}, \ldots, y_t^{(ij)}$ are known, we can estimate the relationship strength $z^{(ij)}$ in the same way as step 2 in the learning algorithm. Second, when the interaction data are unobserved, we just apply Eq. (2) to infer $z^{(ij)}$. Since the interaction data are usually sparse, temporal and difficult to obtain, the second scenario

```
While not converged:
    1. For each Newton-Raphson step:
           For $t = 1, \ldots, m$:
               update $\boldsymbol{\theta}_t$ according to Eq. (13).
    2. For each Newton-Raphson step:
           For $(i, j) \in \mathcal{D}$:
               update $\mathbf{z}^{(ij)}$ according to Eq. (12).
    3. Update $\mathbf{w}$ according to Eq. (16).
```

Table 1: The learning algorithm

is more common in real online social networks. This in fact demonstrates a strength of our hybrid model: the lower part of the model is generative so that the overall model will not suffer much from missing interaction data during training. Once the model is learned, for new data the latent variables can be inferred using only the upper level of variables in the model. In addition, the generative lower part also facilitates application of the learned model for predicting future interactions(e.g., predicting new connections). On the other hand, fully discriminative models which treat user background information and interactions equally, will not be easy to apply in these situations.

3. EXPERIMENTS

3.1 LinkedIn Data

Our first set of experiments evaluate the utility of the proposed model on proprietary data from LinkedIn (www.linkedin.com). LinkedIn is a business-oriented social networking site with more than 50 million users worldwide. Each member can maintain a business profile and establish *connections* with colleagues or other business contacts that they know and trust. Members can search member profiles and job postings, communicate with other members, request/write recommendations, and form/join groups.

3.1.1 Dataset

In our first experiment, we randomly selected 100 LinkedIn users as seed nodes and from these sampled pairs of nodes as follows. To select both connected and unconnected pairs, we considered each seed node and all its neighbors up to two links away in the connection graph. From these direct and indirect neighbors, we sampled a set of around 100,000 pairs. In other words, each sample pair consists of a seed node and another user within its two-hop neighborhood.

For each pair of users (i, j), we computed nine features to capture the similarity among their profiles and their connections (i.e., contacts). We define overall similarity as: $\mathbf{s}^{(ij)} = [s_1^{(ij)} s_2^{(ij)} \ldots s_8^{(ij)}]^T$ and describe each of the eight features in Table 2.

s_1	1 if i and j went to the same school, 0 otherwise
s_2	1 if i and j work in the same company, 0 otherwise
s_3	1 if i and j are in the same geographical region, 0 otherwise
s_4	1 if i and j are in the same industry, 0 otherwise
s_5	1 if i and j have the same job title, 0 otherwise
s_6	1 if i and j are in the same functional area, 0 otherwise
s_7	logarithm of the normalized counts of common groups that i and j join
s_8	logarithm of the normalized counts of common connections that i and j share

Table 2: LinkedIn profile and connection similarity features.

In addition to similarity features, we also considered three types of user interactions in the model. We computed four interaction features $y_1^{(ij)}, y_2^{(ij)}, y_3^{(ij)}, y_4^{(ij)}$ based on connections, recommendations, profile viewing, and address book entries. Table 3 describes each of the features. For each type of interaction, we include an auxiliary variable in the model that denotes the total number of people that i has interacted with in the specified manner (e.g., for y_1 the auxiliary feature counts the total number of nodes k with which i has established a connection).

y_1	1 if i and j have established a connection, 0 otherwise
y_2	1 if i has written a recommendation for j, 0 otherwise
y_3	1 if i has viewed j's profile, 0 otherwise
y_4	1 if i has included j in his or her online LinkedIn address book, 0 otherwise

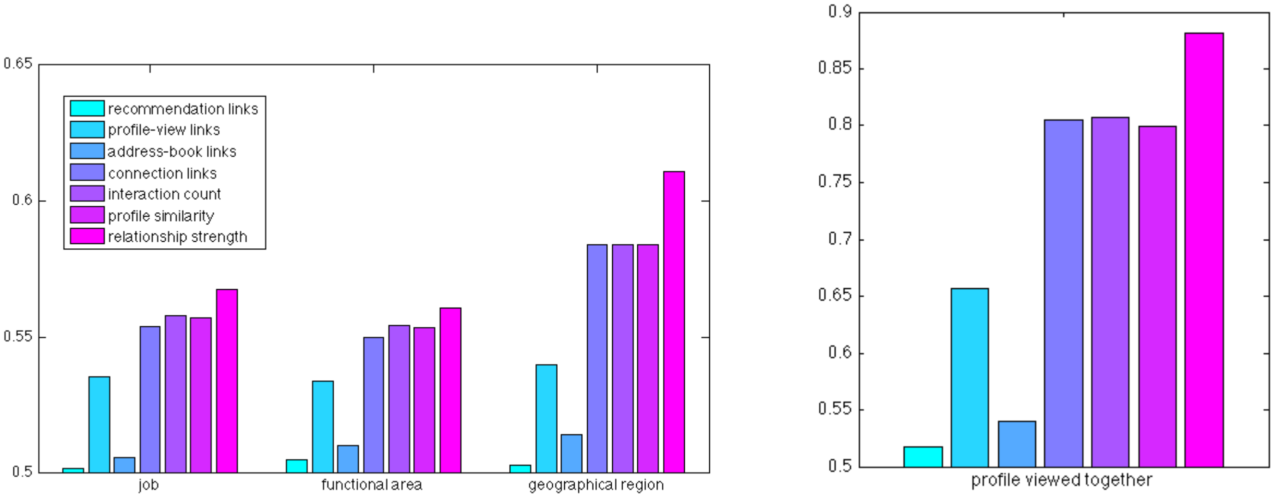
Table 3: LinkedIn interaction features.

3.1.2 Evaluation

We used the proposed model to estimate the relationship strengths for the 100,000 pairs of users and evaluated the results on different recommendation tasks. One important task for the LinkedIn system is the recommendation of people *related* to specific users. For example, when a hiring manager looks for a candidate for a particular job through LinkedIn, a typical situation is that the recruiter knows person A and regards A as a perfect candidate for the job, but A is not available so the recruiter wants to find people who are closely related to A.

For our first set of evaluations, we held out the job, functional area, geographical region similarity features while learning the model and estimating relationship strength. Then we measured how well the estimated relationship strengths identify pairs of users who have the same job title, work in the same functional area, or live in the same region. To do this, we rank the pairs of users by relationship strength and measure the area under the ROC curve (AUC) based on the feature values for the ranked pair (e.g., 1 if they are in the same industry, 0 otherwise). We compare the rankings using relationship strength to several alternative rankings:

1. **Recommendation links**, which ranks pairs that have recommended each other higher than those that do not.
2. **Profile-view links**, which ranks pairs of users according to the number of times that one has viewed the other's profile.
3. **Address-book links**, which ranks pairs of users that list each other in their online LinkedIn address book higher than those that do not.



(a)

(b)

Figure 3: AUC results comparing inferred relationship-strengths to other types of links in LinkedIn data.

4. **Connection links**, which simply ranks pairs of connected users higher than unlinked pairs.

5. **Interaction count**, which ranks pairs of users according to the total count over all four types of interaction links listed above.

6. **Profile similarity**, which ranks pairs according to their overall similarity $\sum_k s_k^{(ij)}$.

Note that (1)-(4) correspond to the different types of observable links in the data. On the other hand, (5)-(6) represent natural heuristics to utilize the profile similarities and interactions separately. These are included to illustrate the utility of the upper portion (profile similarity) and lower portion (interaction occurrence) of the proposed model in isolation. The results are shown in Figure 3(a). For all three tasks, the ranking based on relationship strengths results in a clear gain in AUC, indicating that the model can automatically identify pairs of similar users based on the *combination* of interaction and profile similarity.

Next, we evaluated the quality of the estimated relationship strengths using historical data. We estimated the model using all features and then measured how well the relationship strengths correlate with historical profile *co-viewing*. Again we ranked the pairs of users by the estimated relationship strength and evaluated AUC using a profile co-viewing variable, which is 1 if the pair of users have been viewed by the same person, and 0 otherwise. The results are shown in Figure 3(b). The ranking based on relationship strengths outperforms all the other methods by a large margin, showing that it can provide more relevant recommendations for identifying *related* people. This is evidence that the relationship strengths modeled by our approach, are approximating the way that humans perceive relationships among people.

3.2 Facebook Data

Our next set of experiments evaluate our proposed model on data from Facebook (www.facebook.com). Facebook is a popular online social network site with over 250 million members worldwide. Members create and maintain a personal profile page, which contains information about their views, interests, group memberships, and friends. Friendship links are undirected and are formed through an invitation by one user along with a confirmation by the other. Users can interact with their friends, among other ways, by posting on each others' walls and tagging each other in pictures.

3.2.1 Dataset

For these experiments, we randomly selected five public Purdue Facebook users as seed nodes and considered all nodes within three hops of the seed nodes, in the Purdue network friendship graph, which resulted in a total sample of 4500 nodes. From this sample, we constructed a training set of all the directly linked users, which amounts to 144,712 pairs.

For each pair of users (i, j), we computed three features to capture the similarity among their profiles and their connections (i.e., friends). We define overall similarity as: $\mathbf{s}^{(ij)} = [s_1^{(ij)} s_2^{(ij)}, s_3^{(ij)}]^T$ and describe each of the three features in Table 4.

s_1	logarithm of the normalized counts of common networks for which i and j are both members
s_2	logarithm of the normalized counts of common groups that i and j join
s_3	logarithm of the normalized counts of common friends that i and j share

Table 4: Facebook profile and connection similarity features.

For evaluation purposes, we do not consider similarity features computed from user profile attributes (e.g., gender, relationship status, political and religious views) and use only information about network and group membership, and connection topology in our similarity features $\mathbf{s}^{(i)}$. We do this for two reasons. First, we will later use profile similarity to evaluate the quality of the estimated link strengths, so for accurate evaluation we cannot use these features during learning. Second, in the publicly visible Facebook data, the majority of users either do not list these profile attribute or they do not make the information public (e.g., only 44% have gender and 27% have political views listed in their public profiles).

In addition to similarity features, we also considered two types of user interactions in the model. We computed the interaction features $y_1^{(ij)}, y_2^{(ij)}$ based on wall postings and picture tagging. Table 5 describes each of the features. Furthermore, as auxiliary variables for each of the corresponding interaction variables, we also include the total number of people on whose wall i has posted and the total number of people that i has tagged in pictures in the model.

y_1	1 if i has posted on j's wall, 0 otherwise
y_2	1 if i has tagged j in a picture, 0 otherwise

Table 5: Facebook interaction features.

3.2.2 Evaluation

For evaluation, we used the proposed model to estimate the relationship strengths for the 144,712 pairs of users and compare the weighted graph formed by the estimated relationship strengths to the following four graphs formed from the observed data:

1. **Friendship graph:** The graph consists of all friendship links between users. This network can be viewed as a graph of both "strong" and "weak" relationships.

2. **Top-friend graph:** The graph consists of all *top-friend* nominations. Facebook has a "Top Friends" application which allows users to nominate a small portion of their friends as *best* friends. Such top-friend links can be regarded as "strong" relationships but the resulting network is quite sparse.

3. **Wall graph:** The graph consists of edges corresponding to wall posting activities. Every link correspond to a pair of users (i, j) such that i has posted on j's wall. This network can be viewed as an interaction network.

4. **Picture graph:** The graph consist of edges corresponding to picture-tagging activities. Every link correspond to a pair of users (i, j) such that i has tagged j in his or her uploaded pictures. This network can also be viewed as an interaction network.

We evaluated the resulting graphs in two different ways. First, we measure the increase in autocorrelation on the induced graph, and second we measure the accuracy improvement over several collective classification tasks.

Autocorrelation improvement

In relational data, autocorrelation is the statistical dependency of the same attribute on related instances [10]. To measure the autocorrelation of an attribute with K values in a graph, we first construct a contingency table which consists of K rows and K columns where each row (or column) corresponds to a value of the categorical attribute, and then the autocorrelation is calculated based on the chi-square statistic:

$$\chi^2 = \sum_{i \in K} \sum_{j \in K} \frac{O_{ij} - E_{ij}}{E_{ij}}$$

where E_{ij} is the expected occurrence of the attribute value pair (i, j), and O_{ij} is the observed occurrence of the attribute value pair (i, j) across pairs of linked nodes in the graph. We scale the chi-square statistic to the range $[-1, 1]$ by using the corrected contingency coefficient:

$$CC = \sqrt{\frac{K\chi^2}{(K-1)(N+\chi^2)}}$$

where N is the number of linked pairs in the data. For the weighted graph, we scale the counts in each cell by the link strength of the corresponding link (N is scaled as well).

In Figure 4 we graph the autocorrelation of the withheld profile attributes (i.e., gender, relationship status, political and religious views) on the five networks. For the sparse networks (i.e., top-friend, wall, picture), we plot a single point for the observed autocorrelation. For the relationship-strength graph, we vary the number of links in the network by thresholding the link strength and plot the associated autocorrelation as the number of links is increased. For comparison, on the friendship graph we randomly drop links to assess the autocorrelation on networks with the same density. Note, that the maximum value on the x-axis reflects the autocorrelation with *all* the friendship-links/relationship-strengths in the network.

The graphs show that, with the exception of religious views in the picture network, the relationship-strength network has autocorrelation greater than, or equal to, all other networks (for the same number of links). In particular, the relationship-strength network has significantly higher autocorrelation than the friendship graph in all cases—even though these four profile attributes were not used to learn the model. This demonstrates the utility of the model to identify relationships that reflect the natural similarity among people. Also, we note that as the number of links increase, for both the friendship and relationship-strength networks, the autocorrelation decreases. This indicates a tradeoff between link density and autocorrelation—as we restrict the number of relationships it is likely that similarity among users increase, however increased sparsity will hamper our ability to use the autocorrelation to improve predictive models. We will revisit this issue in the classification discussion below.

Classification improvement

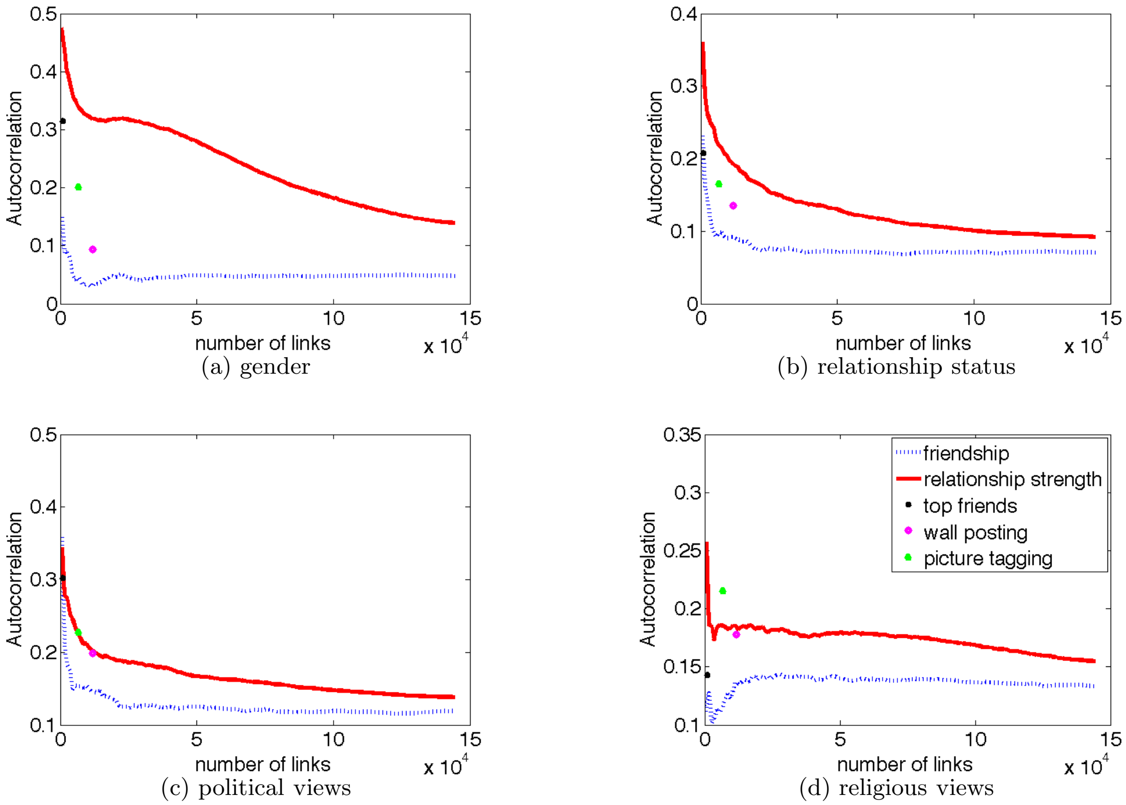
In this section, we further explore the utility of the weighted graph formed from relationship strength by applying it in a number of collective classification tasks. For each of the four profile attributes, we considered a binary classification task based on its most frequent value.

1. **Gender**: *Male?*

2. **Relationship Status**: *Single?*

3. **Political Views**: *Conservative?*

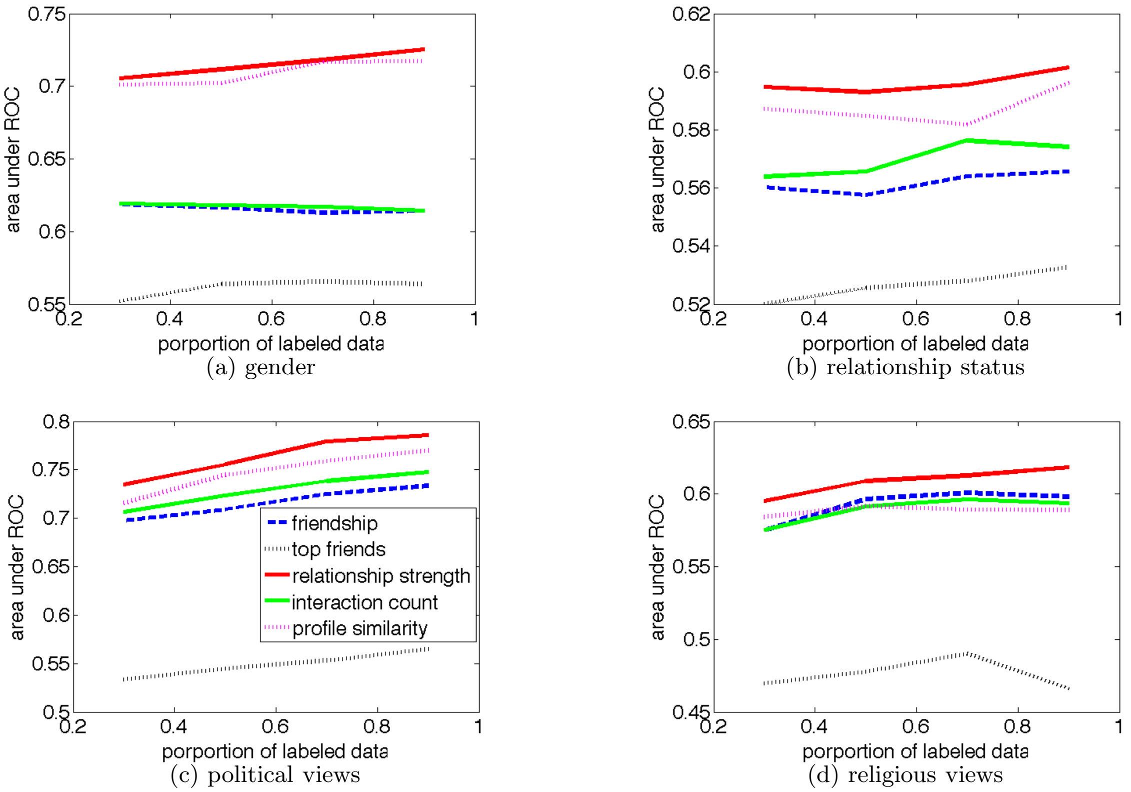
4. **Religious Views**: *Christian?*



(a) gender (b) relationship status

(c) political views (d) religious views

Figure 4: Autocorrelation on various Facebook graphs, as link density is varied.



(a) gender (b) relationship status

(c) political views (d) religious views

Figure 5: Collective classification performance on various Facebook graphs.

We apply a widely-used semi-supervised classification algorithm—the Gaussian random field (GRF) model [20], which assumes autocorrelation is present in the graph and propagates information from the labeled portion of the graph to infer the values for unlabeled nodes. Since the GRF assumes undirected graphs as input, we modify each link $w(i,j)$ in the 4 directed graphs (relationship strength, top-friend, wall posting and picture tagging graphs) to be $\max\{w(i,j), w(j,i)\}$. We vary the proportion of labeled nodes in the graph from 30% to 90% and measure the resulting classification rankings using area under the ROC curve.

Classification using the relationship-strength graph is compared with the four observed Facebook graphs (friendship, top-friend, wall, picture), as well as two additional graphs: (1) the profile-similarity graph, which weights each link by $\sum_{k=1}^{N} s_k^{(ij)}$, and (2) the interaction-count graph which sums the links in the wall, top-friend, and picture graphs. Recall that the profile-similarity and interaction-count graphs are natural heuristics to illustrate the utility of the upper portion (profile similarity) and lower portion (interaction occurrence) of our proposed model in isolation.

The classification results are shown in Figure 5. All results are averaged over five runs with different random selections of labeled instances. The performance curves for the wall graph and the picture graph lie well below the interaction-count graph for all classification tasks so we omit them in the plot for clarity. Note the poor performance of the top-friend graph—this occurs despite the fact that the top-friend network had the highest autocorrelation of the observed graphs for all but religious views, which indicates that high autocorrelation is only helpful for classification if the network has sufficient density to exploit the correlation among neighbors. This illustrates one strength of our proposed approach, since the model can maintain the density of the full friendship graph but significantly increase the autocorrelation levels.

Indeed, the relationship-strength graph results in the highest classification performance for all tasks, suggesting that our approach to summarizing the rich profile and interaction information in online social networks leads to a single meaningful *relationship* graph which can improve subsequent knowledge discovery and prediction tasks. We note that neither the profile-similarity nor the interaction-count graphs perform well across all the tasks. This illustrates another strength of our proposed approach, since we combine both these sources of relationship information together in a single representation of overall relationship strength.

4. RELATED WORK

The recent growth and popularity of online social networks (OSNs) such as Facebook, MySpace, and LinkedIn has lead to a surge in research focused on modeling networks and their properties. Much of this work has focused on the analysis of network structure and growth patterns. For example, Backstrom et al. [3] investigated the evolution of network structure and group membership in MySpace and LiveJournal and showed that homophily can be used to improve predictive models of group membership. Singla and Richardson [18] investigate the correlation between individual search topics among people that interact using instant messaging, and show that not only does a correlation exist but that it increases with the amount of time the users communicate. Crandall et al. [4] study the temporal evolution of link structure and attribute similarity in Wikipedia and propose a mathematical model that includes both influence and homophily effects to predict future behavior in the network. However, nearly all these methods focus on descriptive analysis and generative models of link structure, based on the observed structure in the network—they do not attempt to model the latent properties of the networks.

Another direction of related research has focused on link prediction—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 *topological* features to capture the link structure of the network (e.g., [13, 12]) and those that use attribute similarity features in addition to topological features (e.g., [19, 9]).

We differ from past work on link prediction in that we focus on modeling link strength rather than link existence. We also aim to exploit interaction information among nodes in order to improve model accuracy. O'Madadhain et al. [16] model interaction events, but they formulate a temporal link prediction task which tries to predict the occurrence of events (e.g., co-authorship) in future time intervals. Adamic and Adar [1] also investigate the use of ancillary network information but with the goal of predicting social ties, instead of tie strength. More recently, interaction data has been utilized in predicting relationship strength [11, 7]. However, this work considered the binary prediction task of distinguishing strong ties from weak ties. Our work instead uses a richer representation that can span the full spectrum from weak to strong ties. Moreover, unlike the previous work, we treat interaction data differently from profile features, which leads to a more interpretable model. Finally, this previous work has focused on supervised methods, which usually involve efforts on human annotation, e.g., friendship rating [7] or top friend nomination [11]. Our method takes an unsupervised approach instead, inferring a continuous-valued measure of relationship strength for online social networks.

5. CONCLUSION AND FUTURE WORK

In this paper, we proposed a latent variable model for the task of relationship strength estimation in online social networks. The model attempts to represent the intrinsic causality of social interactions via statistical dependencies. Our experiments show that the weighted graph formed by the estimated relationship strengths gives rise to higher autocorrelation and better classification performance than the graphs formed from various aspects of the raw data.

Besides the natural interest from the social science perspective in estimating relationship strengths from indirect indicators, our model can also be used to improve the performance of online social network systems in a number of ways. Since the estimated relationship strengths result in a weighted graph where the spurious links have been down-weighted and the important links have been highlighted—this could be used to increase the accuracy of many graph learning and social network mining tasks, including link prediction, collaborative filtering, product recommendations, and personalization.

There are a number of modifications to our proposed model that can be explored to improve performance. Since the current inference scheme uses point estimation for the latent variable, we plan to develop alternative inference procedures which maximize the observed data likelihood by integrating over all possible values of the latent variable. This will involve the use of a mathematically convenient approximate distribution on the latent variable. Furthermore, we will consider alternative ways to specify the model—for instance, we could apply kernels in defining profile similarities and learn the functions automatically. Also, nonlinear classification or regression could be used instead of the current choice of a generalized linear model for the interaction dependencies. In both cases, however, the difficulty of inference will increase.

There are also several natural extensions of this work that we will investigate as future work. First, the current model considers the relationship strength on each edge independently. Although inference will be more complex, it may improve the accuracy of the model if we consider the dependencies between adjacent edges. For example, the relationship strengths associated with the same person are likely to be dependent, given the resource constraints on relationship formation and maintenance.

Second, it would be interesting to investigate the evolutionary aspect of relationship strengths over time. We could extend the current model to a temporal setting by smoothing the relationship strengths over adjacent time steps through tied parameters, as long as the interactions at each time step are not too sparse. Alternatively, we could estimate our proposed model on a sequence of network snapshots over several time steps, and analyze how relationship strengths evolve with time.

Finally, we will consider other interesting applications which use relationship strength to understand human behavior (e.g., studying the effect of relationship strength on social influence and diffusion of information). In this work, we attempted to model the causal influence of *homophily*—that similar people tend to interact with each other—on link formation and link strength. We choose this approach because the profile attributes remain relatively stable compared to the interactions among users. However, the process of *social influence*—when people who interact frequently become more similar—is another cause of relational autocorrelation that may affect link formation differently than homophily (see e.g., [2]). An important direction for future work is to model these two effects in a joint model of link strength, particularly in domains where the attribute values change over time.

Acknowledgments

This research is supported by DARPA and NSF under contract numbers NBCH1080005 and SES-0823313, and by a generous gift from Microsoft Research. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements either expressed or implied, of DARPA, NSF, Microsoft, or the U.S. Government.

6. REFERENCES

[1] L. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(2):211–230, 2003.

[2] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–15, 2008.

[3] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD '06*, 2006.

[4] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *KDD '08*, 2008.

[5] K. Dindia and D. Canary. Definitions and Theoretical Perspectives on Maintaining Relationships. *Journal of Social and Personal Relationships*, 10(2):163–173, 1993.

[6] P. Domingos and M. Richardson. Mining the network value of customers. In *KDD '01*, 2001.

[7] E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *CHI '09*, 2009.

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

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

[10] D. Jensen and J. Neville. Linkage and autocorrelation cause feature selection bias in relational learning. In *ICML '02*, 2002.

[11] I. Kahanda and J. Neville. Using transactional information to predict link strength in online social networks. In *ICWSM '09*, 2009.

[12] H. Kashima and N. Abe. A parameterized probabilistic model of network evolution for supervised link prediction. In *ICDM '06*, 2006.

[13] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM '03*, 2003.

[14] M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[15] J. Neville, O. Simsek, D. Jensen, J. Komoroske, K. Palmer, and H. Goldberg. Using relational knowledge discovery to prevent securities fraud. In *KDD '05*, 2005.

[16] J. O'Madadhain, J. Hutchins, and P. Smyth. Prediction and ranking algorithms for event-based network data. *SIGKDD Explorations*, 7(2):23–30, 2005.

[17] U. Sharan and J. Neville. Temporal-relational classifiers for prediction in evolving domains. In *ICDM '08*, 2008.

[18] P. Singla and M. Richardson. Yes, there is a correlation: - from social networks to personal behavior on the web. In *WWW '08*, 2008.

[19] B. Taskar, M. F. Wong, P. Abbeel, and D. Koller. Link prediction in relational data. In *NIPS '03*, 2003.

[20] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML '03*, 2003.