Learning Interactions for Social Prediction in Large-scale Networks

Xiaofeng Yu
HP Labs China
Universal Business Park, 10 Jiu XianQiao Road
Chaoyang District, Beijing, China
xiaofeng.yu@hp.com

Junqing Xie
HP Labs China
Universal Business Park, 10 Jiu XianQiao Road
Chaoyang District, Beijing, China
jun-qing.xie@hp.com

ABSTRACT

Social networks have already emerged as inconceivably vast information repositories and have provided great opportunities for social connection and information diffusion. In light of these notable outcomes, social prediction is a critical research goal for analyzing and understanding social media and online social networks. We investigate underlying social theories that drive the characteristics and dynamics of social networks, including homophily, heterophily, and the structural hole theories. We propose a unified coherent framework, namely mutual latent random graphs (MLRGs), to exploit mutual interactions and benefits for predicting social actions (e.g., users' behaviors, opinions, preferences or interests) and discovering social ties (e.g., multiple labeled relationships between users) simultaneously in large-scale social networks. MLRGs introduce latent, or hidden factors and coupled models with users, users' actions and users' ties to flexibly encode evidences from both sources. We propose an approximate optimization algorithm to learn the model parameters efficiently. Furthermore, we speedup this algorithm based on the Hadoop MapReduce framework to handle large-scale social networks. We performed experiments on two real-world social networking datasets to demonstrate the validity and competitiveness of our approach.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*; I.5.1 [**Pattern Recognition**]: Models—*Statistical*

General Terms

Algorithms, Experimentation

Keywords

Social theories; social actions; social ties; mutual latent random graphs (MLRGs); latent factors; Hadoop MapReduce

CIKM'14, November 3–7, 2014, Shanghai, China.

http://dx.doi.org/10.1145/2661829.2662056.

1. INTRODUCTION

With the dramatically rapid growth of users worldwide, social networks provide great opportunities for social connection and we have access to unprecedented amounts of information about social interaction. Many large-scale online social networking services (e.g., Facebook and Twitter) are changing our daily physical life and emerging new possibilities for investigation [22]. In light of these notable outcomes, social prediction which aims at predicting social actions (e.g., users' behaviors) and inferring social ties (e.g., user-user connections) is an essential research goal. Predicting and analyzing social actions could gain insights into users' behaviors and experience, marketing analytics, etc. Similarly, social tie discovery can aid understanding characteristics and structures of social networks. Doubtlessly, social prediction has generated much theoretical and practical interest in social network analysis and has played an important role in a variety of world-wide-web applications such as online advertising and recommender systems.

Without loss of generality, we define *social actions* as activities, opinions, preferences or interests associated with users in socially connected networks. For example, a social action can be purchasing a product, posting a comment, specifying a particular favorite movie, etc. Traditional social network analysis only distinguishes between pairs of people that are linked vs. not-linked. However, user connections in social media are much richer. Correspondingly, a *social tie* or social relation is defined as any fine-grained relationship between pairs of people in a social network. More specifically, each social tie is explicitly tagged with a sign or label, such as the friend, family, or colleague relationship between linked user pairs. Social tie is the most basic unit to form the network structure, and relationships between users can be either directed or undirected. Note that our definitions for both social actions and social ties are much more general than prior research work, such as [18] for user behavior investigation and [11] for user interest modeling, thus they are more applicable to real-world scenarios.

Social networks are a media designed to be disseminated through social interaction. Information transmitted by the media interact with the personal influence arising from social networks [16, 8]. Real social networks are complex, both social actions and social ties are affected by a variety of factors. However, a fundamental mechanism that drives the characteristics and dynamics of networks is the underlying social theory of *homophily* [13]: people tend to follow the behaviors of their friends, and people tend to create relationships with other people who are already similar to them.

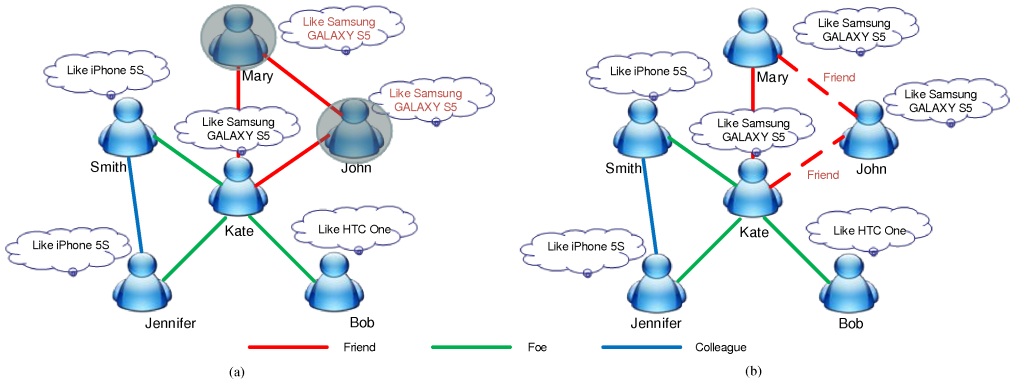

Figure 1: (a) An example social network illustrating social ties' influence on social actions. Since "Mary" and "John" are friends, it is likely that they have the same social action ("Like Samsung GALAXY S5") to follow the behaviors of their friends. (b) An example social network illustrating social actions' influence on social ties. Suppose "Mary", "John" and "Kate" have similar action ("Like Samsung GALAXY S5"), they tend to create the friend relationship with each other.

This phenomenon is often expressed in the adage "birds of a feather flock together" and it shows the power of strong ties[1]. On the other hand, the *heterophily* theory implies that numerous weak ties[2] can be important in seeking information and innovation. Many social structures are characterized by dense clusters, information within these clusters tends to be rather homogeneous and redundant. To find new information or insights, users of the social clusters will have to look beyond the clusters to other acquaintances. This is called "the strength of weak ties" [6]. The heterophily theory is related to the *structural hole* theory [2]: information diffuses and spreads across clusters, when two separate clusters have non-redundant information, opportunities and perspectives, there is said to be a structural hole between them. Thus, a network with many different structural holes will offer additive information benefits. Structural hole users control the information flow between different clusters and they are able to access information from diverse sources and clusters. For example, compared to strong ties, weak ties are more likely to bring information about job openings and opportunities.

Inspired by these social theories, we hypothesize that there exists high correlation and mutual interactions between social actions and social ties. In the following, we will give concrete motivating examples to illustrate and support our hypothesis.

1.1 Motivating Examples

Fig. 1 shows an social network example to illustrate the mutual interactions between social actions and social ties. This concrete social network consists of 6 users "Mary", "Smith", "John", "Kate", "Jennifer", and "Bob". The social actions can be "Like Samsung GALAXY S5", "Like IPhone 5S", and "Like HTC One", and the social ties can be "Friend", "Foe", and "Colleague", respectively. It is well known that different types of social ties have essentially different influence on social actions. Intuitively, a user's trusted friends on the web affect that user's online behavior. [11] and [12] claimed that one user's final behavior decision is the balance between his/her own taste and her/his trusted friends' favors. As shown in Fig. 1(a), since "Mary" and "John" are friends, it is likely that they have the same social action ("Like Samsung GALAXY S5") to follow the behaviors of their friends. "Kate" and "Smith" are foes, the probability that they have the same action is low. This is consistent with the social theories.

On the other hand, social actions also have important influence on social ties, and Fig. 1(b) demonstrates such influence. Suppose "Mary", "John" and "Kate" have similar action ("Like Samsung GALAXY S5"), they tend to create the friend relationship with each other. The influentials or opinion leaders bring in new information, ideas, and opinions, and disseminate them down to the masses through information diffusion [16]. Such users can influence a sufficient number of users in the network with their actions for a long period of time. Obviously, users with similar preferences or behaviors are more likely to be friends than others in social media. Users with momentous activities will attract many other users to be connected with. On the contrary, no body will be interested in users with trivial or insignificant behaviors. Interestingly, knowing the intimate relationship (e.g., friend) between "Mary" and "John" is helpful to discover the same action of them. Similarly, if two users have the same action, it will be a strong evidence indicating the intimate relationship between them. To summarize, exploring bidirectional interactions and rich interdependencies between social actions and social ties to capture mutual benefits is intuitively appealing.

1.2 Research Problems and Challenges

Consequently, we face several interesting but challenging research problems which will be delved:

- Can we predict social actions and discover social ties accurately and simultaneously? Among a variety of complicated factors, what factors should be exploited to capture the essential features for both social actions and social ties?
- Are there any dynamics or mutual interactions between social actions and social ties? If yes, to what extent do they influence each other?
- Popular social networking sites have gathered billions of acting users and are still attracting millions of newbies everyday. How do we deal with big data and large-scale social networks?

[1]Strong ties usually have frequent communication, but ties are redundant due to high clustering in social networks

[2]Weak ties reach far across network, but the communication is infrequent.

1.3 Our Contributions

To address the above research problems, we learn the bidirectional interactions between social actions and social ties for simultaneous social prediction. We further extend our proposed framework for large-scale social network analysis. In summary, we list our major contributions of this paper as follows:

- Firstly, we propose a single unified framework based on exponential-family random graph models [3, 23] for simultaneous social action prediction and social tie discovery. This mutual latent random graph (MLRG) framework incorporates shared latent factors with users, users' actions and users' ties, and defines coupled models to encode both social action and social tie information, to capture bidirectional dynamics and mutual benefits between them.
- Secondly, we propose an approximate algorithm exploiting variational approximation and mean-field theory for optimizing the parameters efficiently. We also perform theoretical analysis on this algorithm.
- Thirdly, we speedup the optimization algorithm based on the Hadoop MapReduce framework to handle big data and large-scale social networks. We perform extensive experiments on two real-world social networking datasets to illustrate the merits of our approach. Several interesting issues, such as latent factor contributions on mutual interactions and speedup efficiency are also discussed and analyzed.

2. OUR APPROACH

In this section we consider both social action prediction and social tie inference in the context of social media, where evidences for both actions and ties are available. We begin by necessary description of preliminaries and notations, we then present the mutual latent random graphs (MLRGs) model, upon which both sources of evidence could be exploited simultaneously to capture their mutual interactions. We also discuss the major difference and superiority of this model against several alternative models.

2.1 Preliminaries and Notations

Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a social network graph, where $\mathbf{V} = \{v_1, v_2, \ldots, v_N\}$ is the set of $|\mathbf{V}| = N$ users and $\mathbf{E} = \{e_{11}, e_{12}, \ldots, e_M\} \subset \mathbf{V} \times \mathbf{V}$ is the set of $|\mathbf{E}| = M$ connections between users. Let $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}(y_i \in \mathcal{Y})$ be the set of actions associated with N users, and $\mathbf{s} = \{s_{11}, s_{12}, \ldots, s_M\}(s_{ij} \in \mathcal{S})$ be the set of corresponding social tie labels associated with M connections. The connection $e_{ij}(1 \le i, j \le N, i \ne j)$ between v_i and v_j might be directed or undirected. To be consistent, both $s_{ij} \ne s_{ji}$ and $s_{ij} = s_{ji}$ are valid settings. Given the observed social network data $\mathcal{D}$ constructing the graph $\mathcal{G}$, our goal is to simultaneously detect the most likely types of actions $\mathbf{y}^*$ and ties $\mathbf{s}^*$ such that both of them are optimized.

2.2 Modeling Social Actions

To characterize the user action y_i, we assume that for the user v_i there exist observable attributes or properties $\mathbf{m}_i$, such as the user's registered information and historical actions. Without loss of generality, we further assume that there exist some hidden, or latent properties $\mathbf{x}_{ij}$ for v_i. These properties are implicit and cannot be observed directly [26], such as the influence from social ties. Consequently, we denote the observable factor $\phi(y_i, v_i, \mathbf{m}_i)$ for observable properties and latent factor $\phi_h(y_i, s_{ij}, \mathbf{x}_{ij})$ for hidden properties, respectively. Given the graph $\mathcal{G}$, the probability distribution of y_i depends on both observable and latent factors as:

$$P_{y_i|\mathcal{G}} \sim \phi(y_i, v_i, \mathbf{m}_i), \quad P_{y_i|\mathcal{G}} \sim \phi_h(y_i, s_{ij}, \mathbf{x}_{ij}),$$
$$P_{y_i|\mathcal{G}} \sim \phi(y_i, v_i, \mathbf{m}_i)\phi_h(y_i, s_{ij}, \mathbf{x}_{ij}). \quad (1)$$

This modeling integrates two types of factors for both observable and latent properties. It captures not only the user-action dependencies, but also the influence from social ties, for exploring social actions.

2.3 Modeling Social Ties

To characterize the social tie s_{ij} between user pair (v_i, v_j), we also assume that there exist observable properties $\mathbf{w}_{ij}$, such as the posterior probability of the social tie s_{ij} assigned to (v_i, v_j). We denote the observable factor $\phi'(s_{ij}, v_i, v_j, \mathbf{w}_{ij})$ for $\mathbf{w}_{ij}$. Similarly, we further assume that there exist some latent properties to incorporate the social action influence on social ties. To be consistent, we still use the vector $\mathbf{x}_{ij}$ to represent the latent properties and the latent factor $\phi_h(y_i, s_{ij}, \mathbf{x}_{ij})$ to capture the social action influence on social ties. Note that both $\mathbf{x}_{ij}$ and $\phi_h(y_i, s_{ij}, \mathbf{x}_{ij})$ now play double duties in encoding social action dependency and social tie connection simultaneously. On the one hand, $\phi_h(y_i, s_{ij}, \mathbf{x}_{ij})$ exploits influence from social ties for modeling social actions. On the other hand, this factor exploits influence from social actions for modeling social ties. By doing so, the latent factor $\phi_h(y_i, s_{ij}, \mathbf{x}_{ij})$ is bi-directionally coupled, encoding both sources of evidence and exploring mutual interactions and dynamics between social actions and social ties. Such mutual influence and dynamics are crucial and modeling them often leads to improved performance. Given the user action y_i and the graph $\mathcal{G}$, we devise the following model for the probability distribution of s_{ij} depending on both observable and latent factors as:

$$P_{s_{ij}|(y_i,\mathcal{G})} \sim \phi'(s_{ij}, v_i, v_j, \mathbf{w}_{ij}), \quad P_{s_{ij}|(y_i,\mathcal{G})} \sim \phi_h(y_i, s_{ij}, \mathbf{x}_{ij}),$$
$$P_{s_{ij}|(y_i,\mathcal{G})} \sim \phi'(s_{ij}, v_i, v_j, \mathbf{w}_{ij})\phi_h(y_i, s_{ij}, \mathbf{x}_{ij}). \quad (2)$$

2.4 Modeling Mutual Interactions

The mutual correlation between social actions and social ties advocates joint modeling of both sources of evidence in a single unified framework. Based on the above descriptions, we define our mutual latent random graph (MLRG) based on exponential-family random graph models (ERGMs) [3, 23], which have gained tremendous successes in social network analysis and have even become the current state-of-the-art [15]. To design a concrete model, one needs to specify distributions for the dependencies for MLRGs. According to the celebrated Hammersley-Clifford theory, the joint conditional distribution $P_{(y_i,s_{ij})|\mathcal{G}}$ is factorized as a product of potential functions over all cliques in the graph $\mathcal{G}$ and we summarize the MLRG in the following table. In summary, our model consists of three factors: the factor $\phi(y_i, v_i, \mathbf{m}_i)$ measuring dependencies of the social action y_i conditioned on $\mathcal{G}$, the factor $\phi'(s_{ij}, v_i, v_j, \mathbf{w}_{ij})$ measuring the social tie s_{ij} between two arbitrary users v_i and v_j in $\mathcal{G}$, and the latent factor $\phi_h(y_i, s_{ij}, \mathbf{x}_{ij})$ exploiting mutual interactions between the social action y_i and social tie s_{ij}.

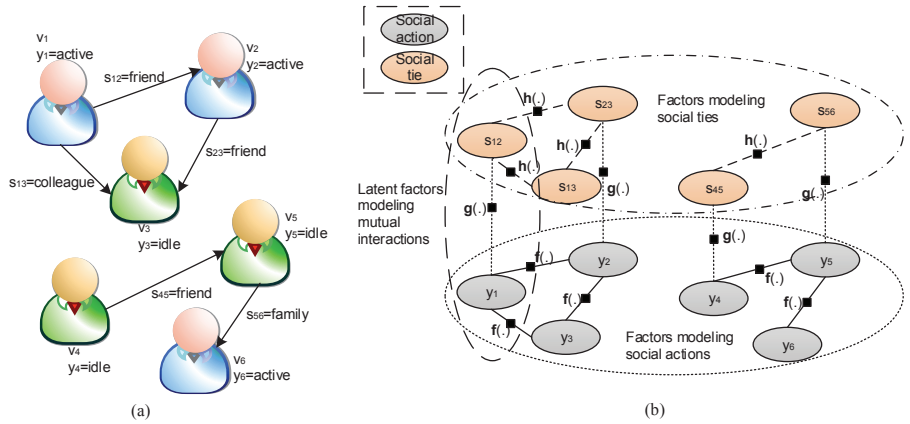

Figure 2: (a) A social network containing 6 users and 5 social ties. The social action can be active or idle, and the social tie can be friend, colleague, or family. (b) The three-dimensional graphical representation of the corresponding MLRG model of (a). The bottom part consists of factors modeling social actions, and the upper part consists of factors modeling social ties. The latent factors modeling mutual interactions are also described. We use different lines to represent functions $\mathbf{f}(\cdot)$, $\mathbf{g}(\cdot)$, and $\mathbf{h}(\cdot)$.

The Mutual Latent Random Graph (MLRG) model	
$\forall y_i \in \mathcal{Y}$	$P_{y_i\|\mathcal{G}} \sim \phi(y_i, v_i, \mathbf{m}_i)\phi_h(y_i, s_{ij}, \mathbf{x}_{ij})$
$\forall s_{ij} \in \mathcal{S}$	$P_{s_{ij}\|(y_i,\mathcal{G})} \sim \phi'(s_{ij}, v_i, v_j, \mathbf{w}_{ij})\phi_h(y_i, s_{ij}, \mathbf{x}_{ij})$
$\forall y_i \in \mathcal{Y}, \forall s_{ij} \in \mathcal{S}$	$P_{(y_i,s_{ij})\|\mathcal{G}} \sim \phi(y_i, v_i, \mathbf{m}_i)\phi_h(y_i, s_{ij}, \mathbf{x}_{ij})\phi'(s_{ij}, v_i, v_j, \mathbf{w}_{ij})$

The three factors $\phi(\cdot)$, $\phi_h(\cdot)$, and $\phi'(\cdot)$ can be instantiated in different ways. In this paper, each factor is defined as the exponential family of an inner product over sufficient statistics (feature functions) and corresponding parameters. Each factor is a clique template whose parameters are tied. More specifically, we define these factors as $\phi(y_i, v_i, \mathbf{m}_i) = \exp\{\sum_{y_i \in \mathcal{Y}} \alpha \mathbf{f}(y_i, v_i, \mathbf{m}_i)\}$, $\phi_h(y_i, s_{ij}, \mathbf{x}_{ij}) = \exp\{\sum_{y_i \in \mathcal{Y}} \sum_{s_{ij} \in \mathcal{S}} \beta \mathbf{g}(y_i, s_{ij}, \mathbf{x}_{ij})\}$, and $\phi'(s_{ij}, v_i, v_j, \mathbf{w}_{ij}) = \exp\{\sum_{s_{ij} \in \mathcal{S}} \gamma \mathbf{h}(s_{ij}, v_i, v_j, \mathbf{w}_{ij})\}$, respectively. α, β, and γ are real-valued weighting vectors and $\mathbf{f}(\cdot)$, $\mathbf{g}(\cdot)$, and $\mathbf{h}(\cdot)$ are corresponding vectors of sufficient statistics. The joint probability distribution shown in the above table can be rewritten as

$$P_{(\mathbf{y},\mathbf{s})|\mathcal{G}} = \frac{1}{\mathcal{Z}} \exp\{\sum_{y_i \in \mathcal{Y}} \alpha \mathbf{f}(y_i, v_i, \mathbf{m}_i) + \sum_{y_i \in \mathcal{Y}} \sum_{s_{ij} \in \mathcal{S}} \beta \mathbf{g}(y_i, s_{ij}, \mathbf{x}_{ij}) + \sum_{s_{ij} \in \mathcal{S}} \gamma \mathbf{h}(s_{ij}, v_i, v_j, \mathbf{w}_{ij})\}, \quad (3)$$

where α, β, and γ are real-valued weighting vectors and $\mathbf{f}(\cdot)$, $\mathbf{g}(\cdot)$, and $\mathbf{h}(\cdot)$ are corresponding vectors of sufficient statistics, which capture social network features of interest, its postulated dependence structure, or both. $\mathcal{Z}$ is the normalization factor to make all probabilities sum to one. In the following, we use $P(\mathbf{y}|\mathcal{G})$, $P(\mathbf{s}|\mathbf{y}, \mathcal{G})$, and $P(\mathbf{y}, \mathbf{s}|\mathcal{G})$ to denote $P_{\mathbf{y}|\mathcal{G}}$, $P_{\mathbf{s}|(\mathbf{y},\mathcal{G})}$, and $P_{(\mathbf{y},\mathbf{s})|\mathcal{G}}$, respectively.

Fig. 2 shows an example social network ($\mathbf{V} = \{v_1, v_2, v_3, v_4, v_5, v_6\}$, $\mathcal{Y} = \{\text{active}, \text{idle}\}$, $\mathcal{S} = \{\text{friend}, \text{colleague}, \text{family}\}$) and the corresponding 3D graphical representation of the MLRG model. The functions $\mathbf{f}(\cdot)$ model dependencies of social actions in the bottom part, and the functions $\mathbf{h}(\cdot)$ model dependencies of social ties in the upper part. More importantly, the functions $\mathbf{g}(\cdot)$ capture mutual influence and dependencies between social actions and social ties. The superiority of MLRG is its ability to represent a large number of complicated factors, and capture rich interdependencies between them. Thus both homophily and heterophily theories could be exploited in this modeling. As we will see, this modeling offers a natural formalism for exploiting bidirectional dependencies and interactions between social actions and social ties to capture their mutual benefits, as well as a great flexibility to incorporate a large collection of arbitrary, overlapping and nonindependent features.

2.5 Discussion

Noticeably, our proposed MLRG model is essentially different from the standard exponential-family random graph models (ERGMs) and the prior models discussed in Sec. 5 mainly in two aspects. Firstly, compared to the standard ERGMs, the MLRG model defines latent factors to assume mutual and dynamical interactions between social ties and social actions. Secondly, compared to the prior models such as [11] and [20], MLRG provides a single unified framework to address both social action prediction and social tie inference simultaneously while enjoying the resources of both sources of evidence.

Importantly, we give an analytical explanation on the mutual nature of our model in terms of a random walk [10] perspective. A random walk on the graph $\mathcal{G}$ is a reversible Markov chain on the vertexes $\mathbf{V}$. The social influence propagation procedure occurs through information diffusion in the social graph $\mathcal{G}$. More specifically: (1) a user v_i will propagate her/his influence to other related users, and will propagate more to the user which has a stronger relation (e.g., friendship) with v_i. This kind of propagation expresses the power of strong ties. Interestingly, this process is consistent with the homophily theory that a user in the social network tends to be similar to their connected neighbors; (2) according to the heterophily theory, a user v_i will also be influenced by other users who have weak ties with the user v_i. The influence propagation will stop when the social graph reaches an equilibrium state, in which both social actions and social ties are mutually reinforced. Both strong ties and weak ties affect this social influence propagation procedure.

3. LEARNING INTERACTIONS

The objective of learning the MLRG model is how to find a configuration of the whole set of model's parameters $\boldsymbol{\theta} = \{\alpha, \beta, \gamma\}$ to maximize the log-likelihood objective function $\mathcal{O}(\boldsymbol{\theta})$ of the observation given the graph $\mathcal{G}$ as $\log P_{\boldsymbol{\theta}}(\mathcal{G}) = \log \sum_{\mathbf{y},\mathbf{s}} P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}, \mathcal{G})$.

Approximating objective function $\mathcal{O}(\boldsymbol{\theta})$. Exactly maximizing this objective function is prohibitively intractable in our model. Recall that the target probability distribution of our MLRG model $P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}|\mathcal{G})$, we employ another probability distribution $Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s})$ parameterized by the vector $\boldsymbol{\theta'} = \{\alpha', \beta', \gamma'\}$, for an approximation of $P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}|\mathcal{G})$. Instead of maximizing the objective function $\mathcal{O}(\boldsymbol{\theta})$ directly, we would like to find the best approximated distribution $Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s})$ such that $Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s})$ and $P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}|\mathcal{G})$ are as close as possible. More specifically, we exploit the Kullback-Leibler (KL) divergence $D_{\mathrm{KL}}(Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s})||P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}|\mathcal{G}))$ between $Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s})$ and $P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}|\mathcal{G})$ as the choice of dissimilarity function to minimize. Consequently, we have the following objective function to maximize:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta'}) &= \log P_{\boldsymbol{\theta}}(\mathcal{G}) - D_{\mathrm{KL}}(Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s})||P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}|\mathcal{G})) \\ &= \log P_{\boldsymbol{\theta}}(\mathcal{G}) - \sum_{\mathbf{y},\mathbf{s}} Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s}) \log \frac{Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s})}{P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}|\mathcal{G})} \\ &= -\sum_{\mathbf{y},\mathbf{s}} Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s}) \log Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s}) + \sum_{\mathbf{y},\mathbf{s}} Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s}) \log P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}, \mathcal{G}) \\ &= \mathbf{H}(Q_{\boldsymbol{\theta'}}) + \mathbf{E}_{Q_{\boldsymbol{\theta'}}}\{\log P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}, \mathcal{G})\}, \end{aligned} \tag{4}$$

where $\mathbf{H}(Q_{\boldsymbol{\theta'}})$ is the entropy of distribution $Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s})$ and $\mathbf{E}_{Q_{\boldsymbol{\theta'}}}\{\cdot\}$ represents the expectation with respect to $Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s})$.

Computing $\mathcal{L}(\boldsymbol{\theta'})$. Now we describe how to compute $\mathcal{L}(\boldsymbol{\theta'})$ efficiently. We explore substructures or subgraphs which form a factorized distribution: factor graphs provide a method to factorize the global probability as a product of local factor functions (e.g., marginal probabilities), each of which depends on a subset of the variables in the graph [7, 21]. Let the original probability $P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}, \mathcal{G})$ factorize into a product of pairwise potentials depending only on the variables associated with each undirected edge as $P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}, \mathcal{G}) = \prod_{e \in E} \psi(\mathbf{y}_e, \mathbf{s}_e, \mathcal{G}_e)$. According to the mean-field variational theory [7], we assume that $Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s})$ forms a feasible factorized distribution as

$$Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s}) = \prod_{y_i \in \mathcal{Y}} Q_{\boldsymbol{\theta'}_i}(y_i) \prod_{s_{ij} \in \mathcal{S}} Q_{\boldsymbol{\theta'}_{ij}}(s_{ij}), \tag{5}$$

where $\mathbf{y} = \{y_i\}(y_i \in \mathcal{Y})$ and $\mathbf{s} = \{s_{ij}\}(s_{ij} \in \mathcal{S})$.

Having the factorization graph structure of the original probability and the structure imposed on the distribution $Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s})$, $\mathcal{L}(\boldsymbol{\theta'})$ can be evaluated feasibly as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta'}) = &\sum_{y_i \in \mathcal{Y}} \mathbf{H}(Q_{\boldsymbol{\theta'}}(y_i)) + \sum_{s_{ij} \in \mathcal{S}} \mathbf{H}(Q_{\boldsymbol{\theta'}}(s_{ij})) \\ &+ \sum_{e \in E} \sum_{\mathbf{r}_{e \cap h}} Q_{\boldsymbol{\theta'}}(\mathbf{r}_{e \cap h}) \log \psi(\mathbf{y}_e, \mathbf{s}_e, \mathcal{G}_e), \end{aligned} \tag{6}$$

where $Q_{\boldsymbol{\theta'}}(\mathbf{r}_{e \cap h})$ is the marginal probability over variables $\mathbf{r} = \{\mathbf{y}, \mathbf{s}\}$ associated with edge e and $\mathcal{Y} \cup \mathcal{S} = h$.

Updating distribution $Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s})$ and learning $\boldsymbol{\theta'}$. Note that $\mathcal{L}(\boldsymbol{\theta'})$ is a concave function of the distribution Q. We need to update the marginal probabilities to optimize $\boldsymbol{\theta'}$ based on feasibility calculation of $\mathcal{L}(\boldsymbol{\theta'})$ via Eq. (6). We resort to an iterative procedure for maximizing $\mathcal{L}(\boldsymbol{\theta'})$ within the class of factored distributions via Eq. (5). The marginal probabilities in Eq. (5) can be updated independently, and we can optimize $\mathcal{L}(\boldsymbol{\theta'})$ one marginal component at a time. Recall that $\mathbf{E}_{Q_{\boldsymbol{\theta'}}}\{\cdot\}$ denotes the expectation with respect to Q, let $\mathbf{E}_{Q_{\boldsymbol{\theta'}}}\{\cdot|r_m\}, m \in h$ be the conditional expectation with respect to Q. We explicate in more detail the feasibility of computation the conditional expectation in the updates as

$$\begin{aligned}&\mathbf{E}_{Q_{\boldsymbol{\theta'}}}\{\log P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}, \mathcal{G})|r_m\} \\ &= \sum_{\{r_i\}_{i \in h \backslash m}} [\prod_{i \in h \backslash m} Q_{\boldsymbol{\theta'}_i}(r_i)] \log P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}, \mathcal{G}) \qquad (7) \\ &= \sum_{e \in E} \sum_{\mathbf{r}_{e \cap \{h \backslash m\}}} \sum_{\mathbf{r}_{h \backslash m}} Q_{\boldsymbol{\theta'}}(\mathbf{r}_{e \cap \{h \backslash m\}}) \log \psi(\mathbf{y}_e, \mathbf{s}_e, \mathcal{G}_e), \qquad (8)\end{aligned}$$

where $h \backslash m$ is the set of nodes other than m, the variable r_m can be either y_i or s_{ij}. $e \cap \{h \backslash m\}$ is either an empty set or refers to a single node associated with edge e. Note that this expectation does not depend on the marginal $Q_{\boldsymbol{\theta'}_m}(r_m)$ but a function of the conditioning variable r_m.

To update the m^{th} marginal, we view $\mathcal{L}(\boldsymbol{\theta'})$ as a function of $Q_{\boldsymbol{\theta'}_m}(\cdot)$ while keeping the remaining marginals fixed, thus the dependency of $\mathcal{L}(\boldsymbol{\theta'})$ on the marginal $Q_{\boldsymbol{\theta'}_m}(\cdot)$ is explicit. Through straightforward calculation, it is easy to verify that maximizing $\mathcal{L}(\boldsymbol{\theta'})$ with respect to $Q_{\boldsymbol{\theta'}_m}(\cdot)$ forms the following mean field equations [21] as

$$Q_{\boldsymbol{\theta'}_m}(r_m) \leftarrow \frac{1}{\mathcal{Z}_m} \exp\{\mathbf{E}_{Q_{\boldsymbol{\theta'}}}\{\log P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}, \mathcal{G})|r_m\}\}, \tag{9}$$

$$\mathcal{Z}_m \leftarrow \sum_{r_m} \exp\{\mathbf{E}_{Q_{\boldsymbol{\theta'}}}\{\log P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}, \mathcal{G})|r_m\}\}. \tag{10}$$

In summary, we describe the whole parameter estimation procedure of $\boldsymbol{\theta'}$ as follows: we compute the marginal expectation $\mathbf{E}_{Q_{\boldsymbol{\theta'}}}\{\log P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}, \mathcal{G})|r_m\}$ via Eq. (8) and we update the marginal $Q_{\boldsymbol{\theta'}_m}(r_m)$ via Eq. (9) and Eq. (10) to maximize the objective function $\mathcal{L}(\boldsymbol{\theta'})$ via Eq. (6). After we compute and maximize $\mathcal{L}(\boldsymbol{\theta'})$, we can have a configuration of the parameter vector $\boldsymbol{\theta'}$. Such optimization procedure runs iteratively until it converged to an equilibrium state in which both $\mathcal{L}(\boldsymbol{\theta'})$ and $\boldsymbol{\theta'}$ are optimized. Let $\boldsymbol{\theta'_t}$ and $\boldsymbol{\theta'_{t+1}}$ be the values of parameter vector $\boldsymbol{\theta'}$ at iteration t and $t+1$, respectively. At iteration $t+1$, each update is carried out in a closed form, the optimization procedure increases $\mathcal{L}(\boldsymbol{\theta'})$ such that $\mathcal{L}(\boldsymbol{\theta'_{t+1}}) \geq \mathcal{L}(\boldsymbol{\theta'_t})$ to maximize $\mathcal{L}(\boldsymbol{\theta'})$. We have a configuration of the parameter vector $\boldsymbol{\theta'}$ at iteration $t+1$ as $\boldsymbol{\theta'_{t+1}} = \arg\max \mathcal{L}(\boldsymbol{\theta'_t})$.

3.1 Theoretical Analysis

We perform theoretical analysis on this parameter estimation algorithm, and we show that it offers guarantees in the form of a lower bound on the true objective function $\mathcal{O}(\boldsymbol{\theta})$. Based on Jensen's inequality and given the non-negativity property of the KL divergence, we have

$$\begin{aligned}\mathcal{O}(\boldsymbol{\theta}) &= \log \sum_{\mathbf{y},\mathbf{s}} P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}, \mathcal{G}) = \log \sum_{\mathbf{y},\mathbf{s}} Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s}|\mathcal{G}) \frac{P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}, \mathcal{G})}{Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s}|\mathcal{G})} \\ &= D_{\mathrm{KL}}(Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s})||P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}|\mathcal{G})) + \mathcal{L}(\boldsymbol{\theta'}) \geq \mathcal{L}(\boldsymbol{\theta'}). \end{aligned} \tag{11}$$

Definitely, $\mathcal{L}(\boldsymbol{\theta'})$ is the lower bound on the observed data $\mathcal{O}(\boldsymbol{\theta})$ (with equality when $Q^*_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s}) = P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}|\mathcal{G})$)). Maximizing the lower bound $\mathcal{L}(\boldsymbol{\theta'})$ with respect to Q will always recover the log-marginal probability $\log P_{\boldsymbol{\theta}}(\mathcal{G})$.

We now analyze the computational complexity of this algorithm as follows. Let $|h|(\mathcal{Y} \cup \mathcal{S} = h)$ be the number of social action and social tie variables $\mathbf{r} = \{\mathbf{y}, \mathbf{s}\}$, and let ν be the distinct values each variable can take. In Eq. (6), the computation of the first two summations of $Q_{\boldsymbol{\theta'}_m}(\cdot)$ takes $O(|h|\nu)$. Similarly, the computation of the third summation takes $O(|E|\nu^2)$ since each expectation $\mathbf{E}_{Q_{\boldsymbol{\theta'}}}\{\cdot\}$ involves two variables and there are $|E|$ edges. Let T be the iteration number, thus the overall computational complexity of this algorithm is $O((|h|\nu + |E|\nu^2) \cdot T)$. As can be seen, this algorithm provides a fast, deterministic approximation to otherwise unattainable posteriors. Also its convergence time is independent of dimensionality, especially in the case that Markov chain Monte Carlo (MCMC) algorithms are too slow to converge [21].

3.2 Speedup for Large-scale Data

Large-scale data analysis is a significant challenge in social networks. We explore such challenge and we present a distributed implementation of the learning algorithm based on the Hadoop MapReduce framework[3] to scale up to large-scale networks. The Apache Hadoop is a framework that allows for the scalable parallel and distributed computing of large data sets across clusters of computers using simple programming models such as MapReduce. This framework adopts a master-slave architecture which consists of one master node and multiple slave nodes in the clusters. The master node is generally served as `JobTracker` and each slave node is served as `TaskTracker`.

Generally speaking, our proposed learning algorithm contains two iterative steps: (1) compute the marginal expectation $\mathbf{E}_{Q_{\boldsymbol{\theta'}}}\{\log P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}, \mathcal{G})|r_m\}$ and update the marginal $Q_{\boldsymbol{\theta'}_m}(r_m)$ to maximize $\mathcal{L}(\boldsymbol{\theta'})$; (2) optimize the parameter vector $\boldsymbol{\theta'}$ based on (1). The most expensive part is the first step, and we develop a distributed algorithm to speed up. Based on the Hadoop MapReduce framework, one master node is responsible for optimizing parameters (Step (2)), and the other multiple slave nodes are responsible for calculating the marginal probabilities (Step (1)). Since our algorithm exploits factorized distribution of the real probability, the substructures or subgraphs are then distributed and parallelized over slave processors. In other words, each slave node calculates its own "local" marginal expectation $\mathbf{E}_{Q_{\boldsymbol{\theta'}}}\{\log P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}, \mathcal{G})|r_m\}$ via Eq. (8) and the marginal $Q_{\boldsymbol{\theta'}_m}(r_m)$ via Eq. (9) and Eq. (10) through the *Map* stage. In the *Map* stage, each slave node receives a subset of data as input and produces a set of intermediate key/value pairs for the marginal probabilities. Finally we combine and aggregate these marginal probabilities on each slave node together to calculate and maximize $\mathcal{L}(\boldsymbol{\theta'})$ via Eq. (6) for optimizing the parameter vector $\boldsymbol{\theta'}$ on the master node in the *Reduce* stage. In the *Reduce* stage, the master node merges all intermediate marginal values associated with the same intermediate key, calculates $\mathcal{L}(\boldsymbol{\theta'})$ and optimizes $\boldsymbol{\theta'}$ for the final computation results. We summarize this distributed learning algorithm in Algorithm 1.

3.3 Social Prediction

The optimized parameter vector $\boldsymbol{\theta'}$ can be used to infer social actions and social ties. More specifically, given the testing network data, the task of social prediction is to find the most likely types of actions $\mathbf{y}^*$ and corresponding social tie labels $\mathbf{s}^*$ that have the maximum posterior probability:

$$(\mathbf{y}^*, \mathbf{s}^*) = \arg\max_{(\mathbf{y},\mathbf{s})} P(\mathbf{y}, \mathbf{s}|\mathcal{G}). \quad (12)$$

This can be accomplished by performing the model inference which is straightforward. After parameter estimation via Algorithm 1, the marginal expectation $\mathbf{E}_{Q_{\boldsymbol{\theta'}}}\{\log P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}, \mathcal{G})|r_m\}$ and marginal probability $Q_{\boldsymbol{\theta'}_m}(r_m)$ are obtained. We predict the labels of social actions and social ties by finding the maximum a posterior (MAP) social action labeling assignment and corresponding social tie labeling assignment that have the largest marginal probability.

Algorithm 1: The distributed learning algorithm for MLRGs.

Input: The social graph $\mathcal{G}$, number of iterations T.
Output: Optimized parameters $\boldsymbol{\theta'} = \{\alpha', \beta', \gamma'\}$.
Initialize pairwise potentials $\psi(\mathbf{y}_e, \mathbf{s}_e, \mathcal{G}_e)$;
Initialize factorized distribution for $Q_{\boldsymbol{\theta'}}(\mathbf{y}, \mathbf{s})$;
repeat
  Map process: Calculate marginals on all slave nodes in parallel;
  Each slave node calculates the marginal expectation $\mathbf{E}_{Q_{\boldsymbol{\theta'}}}\{\log P_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{s}, \mathcal{G})|r_m\}$ according to Eq. (8);
  Each slave node calculates the marginal probability $Q_{\boldsymbol{\theta'}_m}(r_m)$ according to Eq. (9) and Eq. (10);
  Reduce process: Calculate objective function and update all parameters on the master node;
  Master node calculates and maximizes objective function $\mathcal{L}(\boldsymbol{\theta'})$ according to Eq. (6);
  Master node updates and optimizes all parameters $\boldsymbol{\theta'} = \{\alpha', \beta', \gamma'\}$ such that $\boldsymbol{\theta'} = \arg\max \mathcal{L}(\boldsymbol{\theta'})$;
until *converge*;

4. EXPERIMENTS

4.1 Data Collection and Statistics

Since our prediction task involves both social actions and social ties, several publicly available social network datasets are not suitable for us. We collected the following two real-world datasets for our experimental study:

Epinion+. This dataset is extracted by crawling Epinions[4], a well known knowledge sharing site and consumer review site. Users of this site can assign integer ratings from 1 to 5 as their personal opinions or reviews on products, companies, movies, etc. These ratings and reviews will influence future customers when they are about to decide whether a product is worth buying or a movie is worth watching. Users can also decide whether to trust or distrust each other. Thus Epinions is a who-trust-whom online social network. The social actions here are users' review ratings, and we use *very dislike*, *dislike*, *neural*, *like*, *very like* to represent the ratings from 1 to 5 on products, respectively. The social ties are the *trust* and *distrust* relationships in this network of product reviewers. Our dataset consists of 126,785 users who have rated a total of 102,807 different items from 5 categories: electronics, health & beauty, computers, media, and other.

[3]http://hadoop.apache.org/

[4]http://www.epinions.com/

Table 1: Statistics of Epinion+ and Mobile datasets.

Statistics	Epinion+	Mobile
Num. of Nodes	126,785	3,268
Num. of Actions	816,946	30,776
Num. of Ties	620,159	18,489
Avg. Num. of Actions	6.44	9.42
Max. Num. of Actions	932	1278
Avg. Num. of Ties	4.89	5.66
Max. Num. of Ties	956	247
ACC	0.14	0.45
Diameter	15	11

The total number of ratings is 816,946. Thus the number of distinct action labels is $5 \times 5 = 25$ (We view the same ratings in the same category as the same labels to alleviate the sparsity problem). This dataset also contains 620,159 relationships, of which more than 80% are trust relationships. A more detailed statistics is shown in Table 1.

Mobile. Mobile phones have become an important tool for communication. They are an ideal platform for understanding social influence and social dynamics. We collected a social network dataset containing 3,268 mobile phone users. The social actions are formed by *calling* or *sending short messages* between each other during a few months. The social ties we investigated are *friend*, *family*, and *colleague* relationships. Table 1 also lists detailed statistics of this dataset, where ACC denotes the average clustering coefficient, and Diameter is the longest shortest path in the network. The Avg. Num. of item and Max. Num. of item are the average number of item per user and maximal number of item per user, respectively.

4.2 Experimental Setup

For quantitative performance evaluation, we employed the standard measures of Precision (P), Recall (R), and F-measure for both social action prediction and social tie inference. We also investigated the efficiency issue of the models.

- **Precision (P)**: the number of correctly predicted actions (or ties) divided by the total number of predicted actions (or ties).
- **Recall (R)**: the number of correctly predicted actions (or ties) divided by the total number of actions (or ties) in the dataset.
- **F-measure**: the weighted harmonic mean of precision and recall, and we used $F_{\beta=1} = \frac{2PR}{P+R}$, the balanced F-score which weights precision and recall evenly.
- **Efficiency**: the running time of the compared models, including learning time and inference time.

We performed four-fold cross-validation on the two datasets, and took the average performance. We compared our approach with the following alternative methods for predicting social actions and inferring social ties:

- **SVM**: This model views social action prediction and social tie inference as two separate classification problems, and solves them independently without considering their bidirectional interactions.

Table 2: Comparative performance of social action prediction on the Epinion+ dataset.

Method	Precision	Recall	F1-score
SVM	71.48	69.05	70.24
ERGM	72.08	69.76	70.90
DCRF	73.78	71.22	72.48
MLRG	**74.96**	**73.89**	**74.42**

- **ERGM**: This is the traditional exponential-family random graph model for social action prediction and social tie inference independently. Note that this model has no latent factor $\phi_h(\cdot)$ incorporated.
- **DCRF**: This model is a dynamical and factorial CRF [17] used to jointly solve the two tasks. This model was originally proposed for labeling and segmenting sequence data.

We exploited a wide range of important features for all the models. For the Epinion+ dataset, the features include in-degree, out-degree, total-degree, common neighbors, etc. For the Mobile dataset, the features include temporal and social features. For the latent factor, we incorporated social tie evidences and hypotheses as features to capture social actions, and we also incorporated social action evidences and hypotheses as features to leverage social ties.

All these models exploited standard parameter learning and inference algorithms in our experiments. For ERGM, we employed a standard gradient based method for parameter estimation. To avoid over-fitting, penalization techniques on likelihood were also performed. All experiments were performed on the Linux blade server, with 18 2.5GHz Intel Xeon E5-2640 CPU processors and 16×18 GB of memory.

4.3 Performance

Table 2 and Table 3 show the social action prediction performance of different methods on the Epinion+ and Mobile datasets, respectively. Table 4 and Table 5 list the social tie discovery performance of different methods on the Epinion+ and Mobile datasets, respectively. We only compared the overall social action prediction performance on the Epinion+ dataset in Table 2, since the number of possible action labels is large (25 in total). The best overall Precision, Recall and F1-measure of these results are highlighted in boldface. Our proposed MLRG consistently achieves better performance on F1-measure than other comparison methods. We performed McNemar's paired tests on these results, which confirm that all the improvements of our proposed models over the baseline methods are statistically significant.

The SVM and ERGM methods generally produce poor prediction performance on the two datasets. They perform social action prediction and social tie discovery independently without considering the mutual correlations between them, thus leading to reduced performance. By modeling interactions between actions and ties, boosted performance can be achieved, as illustrated by the DCRF model. However, the DCRF model applies loopy belief propagation (LBP) for approximate learning and inference, which is inherently unstable and may cause convergence problems. Another shortcoming of DCRF is that the graphical structure of this model is not well suited for social networks, as DCRF was originally proposed for sequence labeling problems. As

Table 3: Comparative performance of social action prediction on the Mobile dataset.

Actions	SVM			ERGM			DCRF			MLRG		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Calling	81.07	76.13	78.52	81.99	80.57	81.27	90.12	82.73	86.27	90.66	89.24	89.94
Sending msg	82.76	78.34	80.49	79.32	78.64	78.98	88.78	80.44	84.40	88.75	87.91	88.33
Overall	82.45	76.56	79.40	80.69	79.70	80.19	**89.67**	81.32	85.29	89.55	**88.70**	**89.12**

Table 4: Comparative performance of social tie discovery on the Epinion+ dataset.

Ties	SVM			ERGM			DCRF			MLRG		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Trust	83.77	76.68	80.07	75.93	83.62	79.59	83.45	76.99	80.09	84.26	82.91	83.58
Distrust	83.01	75.49	79.07	75.03	83.05	78.84	82.78	76.16	79.33	82.95	80.44	81.68
Overall	83.19	76.02	79.44	75.33	**83.22**	79.08	83.25	76.58	79.78	**83.45**	82.69	**83.07**

can be seen, our model achieves stronger interactions via appropriate graphical structures for social networks, and coupled latent factors to exploit bidirectional mutual interactions between social actions and social ties. We also noticed that the prediction performance on the two datasets varies differently. The prediction performance on the Mobile dataset is much better than that on the Epinion+ dataset. This is due to the difference of the social network properties.

4.4 Latent Factor Contributions

We examined the nature and effectiveness of the associated latent factors of our MLRG on exploiting bidirectional interactions and mutual benefits between social actions and social ties. Fig. 3 and Fig. 4 demonstrate their feasibility on the Epinion+ and Mobile datasets, respectively. It shows that the latent factors consistently enhance Precision, Recall, and F1-measure for both social action prediction and social tie inference tasks. For example, the latent factors significantly improve the F-measure by 7.92% (from 81.20 to 89.12) for social action prediction, and improve the F-measure by 8.24% (from 80.24 to 88.48) for social tie discovery on the Mobile dataset, respectively. However, the latent factor contributions on the Epinion+ dataset are significantly less: 2.9% F-measure improvement on social action prediction and 3.89% F-measure improvement on social tie discovery. We present an in-depth analysis on this phenomenon. By carefully investigating the two datasets, we found that in the Mobile dataset, the mobile phone users tend to cluster together to create tightly knit groups characterized by a relatively high density of ties. However, in the Epinion+ dataset, the social network is loose and many users seldom connect with each other through the trust or distrust relationships. This is also implied by the average clustering coefficient (ACC) shown in Table 1. The ACC of Mobile dataset is significantly higher than that of Epinion+ dataset (0.45 vs. 0.14). These results not only illustrate that there exists high correlations and mutual interactions between social actions and social ties, but also demonstrate the feasibility and effectiveness of our latent factors for exploring them, especially when the nodes in the network cluster together tightly with high density of ties.

4.5 Speedup Efficiency

We empirically compared the effectiveness and efficiency of the distributed learning algorithm of our model against other baselines. Table 6 shows the total running time of

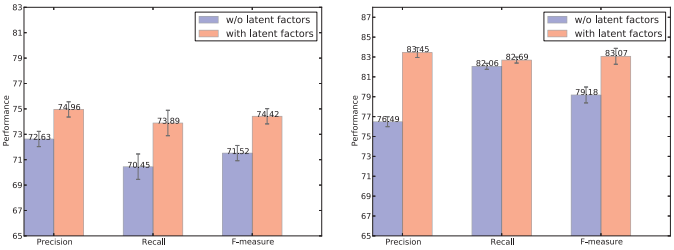

Figure 3: Contribution of latent factors on social action prediction (left) and social tie inference (right) on the Epinion+ dataset.

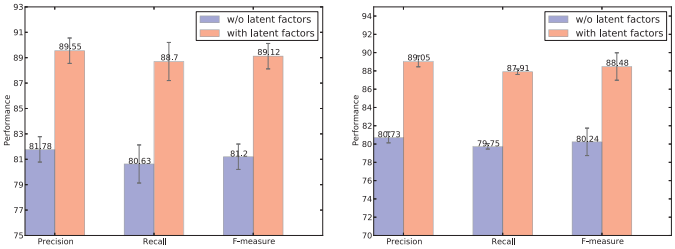

Figure 4: Contribution of latent factors on social action prediction (left) and social tie inference (right) on the Mobile dataset.

different models on the two datasets in our experiments. We used single machine node for the models SVM, ERGM, DCRF and MLRG. It is particularly notable that our MLRG model without speedup takes much less time than the joint model DCRF for running. As stated before, when the graph has large tree-width as in our case, the LBP algorithm in DCRF is inefficient, and is slow to converge. Table 6 also summarizes the running time of the distributed learning algorithm for our model. Here we used 16 processor nodes for the speedup of the distributed learning algorithm. In particular, our model (MLRG with Speedup) is over orders of magnitude faster than MLRG with single processor node. As can be seen, the speedup ratios are 14.2 (9.25hrs/39.10mins) and 13.5 (65.27mins/4.83mins) on the Epinion+ and Mobile datasets, respectively.

Table 5: Comparative performance of social tie discovery on the Mobile dataset.

Ties	SVM			ERGM			DCRF			MLRG		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Friend	81.92	76.73	79.24	80.93	79.25	80.08	89.46	79.75	84.33	89.73	88.78	89.25
Family	81.04	75.72	78.29	79.81	78.03	78.91	88.82	78.42	83.30	88.71	87.52	88.11
Colleague	81.15	75.59	78.27	79.87	78.14	79.00	89.01	78.67	83.52	88.64	87.39	88.01
Overall	81.37	76.06	78.63	80.19	78.44	79.31	**89.19**	78.96	83.76	89.05	**87.91**	**88.48**

Table 6: Efficiency comparison of different models on the Epinion+ and Mobile datasets. The running time includes both learning time and inference time.

Method Running Time	Epinion+ Dataset	Mobile Dataset
SVM	3.86hrs	29.25mins
ERGM	2.40hrs	18.20mins
DCRF	15.48hrs	1.69hrs
MLRG	9.25hrs	65.27mins
MLRG with Speedup	39.10mins	4.83mins

Fig. 6 demonstrates the speedup ratios of the distributed learning algorithm with different numbers of computer nodes. We increased the number of nodes from 2 to 16, with an incremental step of 2 on both Epinion+ and Mobile datasets. The two speedup curves are near linear. This finding is particularly interesting, suppose we have a commercial distributed system with thousands of computer nodes, our learning algorithm can be further accelerated over hundreds of or even thousands of times.

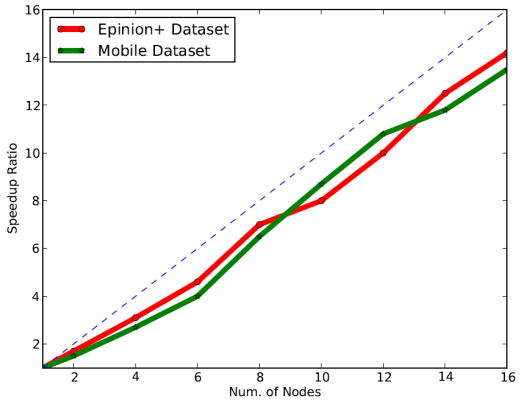

Figure 6: Speedup efficiency for the distributed learning algorithm of MLRGs.

4.6 Case Study

We performed a representative case study to further demonstrate the effectiveness of the proposed model, and Fig. 5 shows a social prediction example from our Epinion+ dataset of the computers category. All the users here are anonymized. Both SVM and ERGM are prone to many prediction errors. For example, the SVM model mis-predicted 3 social ties and 2 social actions, and the ERGM model mis-predicted 2 social ties and 2 social actions. Since they model social actions and social ties as two separate tasks. The DCRF model can alleviate some of the errors by solving the two tasks jointly. Our proposed MLRG model can correct most of the errors made by SVM and ERGM models. This illustrative example shows the validity and competitiveness of our model.

5. RELATED WORK

In online social media, social action prediction and social tie discovery are two fundamental tasks for social network analysis. Traditionally, they were considered as separate tasks and solved independently without taking into consideration the reciprocities or mutual interactions between them. [18] proposed a noise tolerant time-varying model to track social behaviors, which are affected by various kinds of factors, such as users' attributes, users's historical behaviors, social influence and social network structures. Aiming at modeling user actions more accurately and realistically, [11] and [12] considered connections among users and proposed social trust ensemble to fuse the users' tastes and their trusted friends' favors together. In this modeling, only trusted friend relationship was considered. [4] investigated users' social behaviors from a spatio-temporal-social aspect in location-based mobile social networks. For social tie prediction, [20] proposed a semi-supervised framework, the partially labeled factor graph model to infer the type of social relationships. [19] inferred social ties across multiple heterogeneous networks via transfer learning. [24] investigated labeling the edges of a social network graph as either positive or negative relations to show that it is possible to infer signed social ties. Thus we can turn an unsigned acquaintance network into a signed trust-distrust network. However, the social tie label in [24] is only binary (positive/negative). As stated in Sec. 1, our definitions for social actions and social ties are much more general than the above-mentioned research work, thus our modeling is more applicable to real-world problems. Our approach involving both actions and ties modeling is more challenging than [18, 20]. Most importantly, we introduce latent factors and coupled models to incorporate both sources of evidences, and we capture bidirectional interactions between social actions and social ties to predict them simultaneously.

Exploring interplay and mutual benefits between relevant tasks has proven to be highly desirable in NLP [27, 29], data mining and information extraction research communities [31, 28, 25, 30, 32]. We are also aware of several research work attempting to explore joint or unified models in social media and social network analysis [14, 9, 1, 5]. For ontology emergence and semantics, [14] extended a bipartite model of ontologies to a tripartite model of actors, concepts and instances. [9] proposed a unified approach, the latent user preference model (LUPM), to employ mixed membership stochastic block models and topic models for users and the posted content on social networking sites. This unified modeling captures the interactions between users and posts, but it is only loosely coupled. Moreover, the LUPM is a generative model, whereas our proposed model is discriminative. Compared to discriminative models, generative models may have difficulties to capture complex dependencies and rich

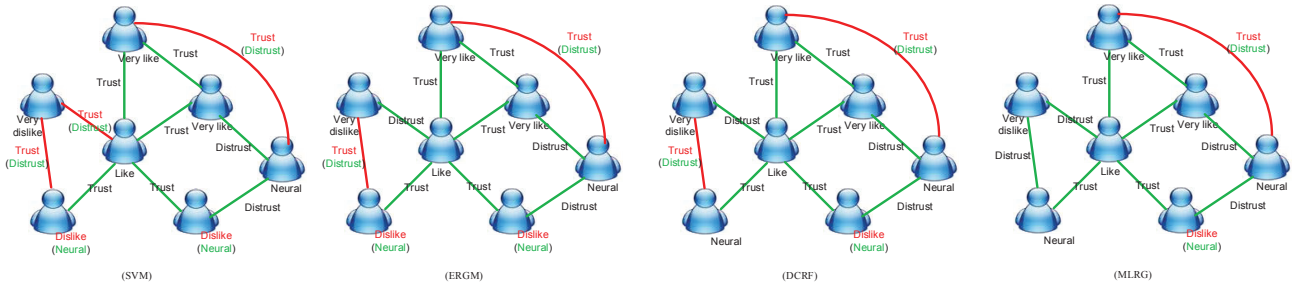

Figure 5: A case study from the Epinion+ dataset to compare social prediction performance. The mis-predicted social action and social tie labels are in red color. The labels in green color are correctly predicted.

features. Due to these reasons, discriminative models are generally more accurate than generative models. [1] proposed a CRF based approach to user identity resolution which combines user profile attributes and social linkage. [5] extended the social-attribute network (SAN) framework with several link prediction algorithms. The SAN framework integrates network structure and node attributes to perform both link prediction and attribute inference. However, due to the computational complexity problem, both [1] and [5] require considerable engineering and they are not scaled up to large-scale social graphs. Since individual users are socially connected, social influence occurs through information diffusion in social networks. Social influence happens when one's opinions or behaviors are affected by others, intentionally or unintentionally. We efficiently exploit mutual interactions and benefits via MLRGs, and we speedup our framework based on Hadoop MapReduce to scale it up to big data and large-scale social networks.

6. CONCLUSION AND FUTURE WORK

In this paper we investigated underlying social theories, and we presented a coherent unified framework, mutual latent random graphs (MLRGs), for simultaneous social action prediction and social tie discovery. This framework incorporates shared latent factors and coupled models with users, users' actions and users' ties to exploit bidirectional mutual interactions and interdependencies between social actions and social ties. We proposed an approximated algorithm which explores factor graph substructures and factorized distribution to optimize model parameters efficiently, and this algorithm offers a theoretical lower bound on the real objective function. Furthermore, we scaled it up to large-scale social network process based on the Hadoop MapReduce architecture. Extensive experimental results on two real-world datasets Epinion+ and Mobile exhibit that our model significantly outperforms several state-of-the-art models while also running much faster. Several interesting issues were analyzed and discussed as well. For the future work, we plan to further investigate and extend our framework to more general scenarios including semi-supervised and unsupervised learning settings. We also plan to apply and test our approach on other large-scale social network datasets.

7. REFERENCES

[1] S. Bartunov, A. Korshunov, S.-T. Park, W. Ryu, and H. Lee. Joint link-attribute user identity resolution in online social networks. In *Proceedings of The 6th SNA-KDD Workshop*, Beijing, China, 2012.

[2] R. S. Burt. *Structural holes: The social structure of competition*. Harvard University Press, Cambridge, MA, 1992.

[3] O. Frank and D. Strauss. Markov graphs. *Journal of the American Statistical Association (JASA)*, 81(395):832–842, 1986.

[4] H. Gao, J. Tang, X. Hu, and H. Liu. Modeling temporal effects of human mobile behavior on location-based social networks. In *Proceedings of CIKM-13*, pages 1673–1678, San Francisco, CA, USA, 2013.

[5] N. Z. Gong, A. Talwalkar, L. Mackey, L. Huang, E. C. R. Shin, E. Stefanov, E. Shi, and D. Song. Joint link prediction and attribute inference using a social-attribute network. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(2), 2014.

[6] M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[7] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical methods. *Machine Learning*, 37:183–233, 1999.

[8] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proceedings of WWW-10*, pages 591–600, Raleigh, North Carolina, USA, 2010.

[9] H. Lakkaraju and A. Rai. Unified modeling of user activities on social networking sites. In *Proceedings of NIPS Workshop on Computational Social Science and the Wisdom of Crowds*, 2011.

[10] L. Lovász. Random walks on graphs: A survey. *Combinatorics, Paul Erdős is Eighty*, 2:353–398, 1996.

[11] H. Ma, I. King, and M. R. Lyu. Learning to recommend with social trust ensemble. In *Proceedings of SIGIR-09*, pages 203–210, Boston, MA, USA, 2009.

[12] H. Ma, I. King, and M. R. Lyu. Learning to recommend with explicit and implicit social relations. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–19, 2011.

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

[14] P. Mika. Ontologies are us: A unified model of social networks and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(1):5–15, 2007.

[15] G. Robins, T. A. B. Snijders, P. Wang, M. Handcock, and P. Pattison. Recent developments in exponential random graph (p^*) models for social networks. *Social Networks*, 29(2):192–215, 2007.

[16] M. G. Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Proceedings of KDD-10*, pages 1019–1028, Washington, DC, USA, 2010.

[17] C. Sutton, A. McCallum, and K. Rohanimanesh. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. *Journal of Machine Learning Research*, 8:693–723, 2007.

[18] C. Tan, J. Tang, J. Sun, Q. Lin, and F. Wang. Social action tracking via noise tolerant time-varying factor graphs. In *Proceedings of KDD-10*, pages 1049–1058, Washington, DC, USA, 2010.

[19] J. Tang, T. Lou, and J. Kleinberg. Inferring social ties across heterogeneous networks. In *Proceedings of WSDM-12*, pages 743–752, Seattle, WA, USA, 2012.

[20] W. Tang, H. Zhuang, and J. Tang. Learning to infer social ties in large networks. In *Proceedings of ECML/PKDD-11*, pages 381–397, Athens, Greece, 2011.

[21] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.

[22] S. Wasserman, K. Faust, D. Iacobucci, and M. Granovetter. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[23] S. Wasserman and P. Pattison. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^*. *Psychometrika*, 61(3):401–425, 1996.

[24] S.-H. Yang, A. J. Smola, B. Long, H. Zha, and Y. Chang. Friend or frenemy?: predicting signed ties in social networks. In *Proceedings of SIGIR-12*, pages 555–564, 2012.

[25] X. Yu, I. King, and M. R. Lyu. Towards a top-down and bottom-up bidirectional approach to joint information extraction. In *Proceedings of CIKM-11*, pages 847–856, Glasgow, Scotland, UK, 2011.

[26] X. Yu and W. Lam. Hidden dynamic probabilistic models for labeling sequence data. In *Proceedings of AAAI-08*, pages 739–745, Chicago, USA, 2008.

[27] X. Yu and W. Lam. An integrated probabilistic and logic approach to encyclopedia relation extraction with multiple features. In *Proceedings of COLING-08*, pages 1065–1072, Manchester, United Kingdom, 2008.

[28] X. Yu and W. Lam. Bidirectional integration of pipeline models. In *Proceedings of AAAI-10*, pages 1045–1050, Atlanta, Georgia, USA, 2010.

[29] X. Yu and W. Lam. Jointly identifying entities and extracting relations in encyclopedia text via a graphical model approach. In *Proceedings of COLING-10*, pages 1399–1407, Beijing, China, 2010.

[30] X. Yu and W. Lam. Probabilistic joint models incorporating logic and learning via structured variational approximation for information extraction. *Knowledge and Information Systems (KAIS)*, 32:415–444, 2012.

[31] X. Yu, W. Lam, and B. Chen. An integrated discriminative probabilistic approach to information extraction. In *Proceedings of CIKM-09*, pages 325–334, Hong Kong, China, 2009.

[32] X. Yu and J. Xie. Modeling mutual influence between social actions and social ties. In *Proceedings of COLING-14*, pages 848–859, Dublin, Ireland, 2014.