A Unified Probabilistic Model of User Activities and Relations on Social Networking Sites

Xiaofeng Yu, Junqing Xie, and Shuai Wang
Networking and Mobility
HP Labs China
UBP, Chaoyang District, Beijing, China
{xiaofeng.yu,jun-qing.xie,shuai.wang}@hp.com

Abstract

In this work, we investigate the bidirectional mutual interactions (BMI) between users' activities and user-user relationships on social networking sites. We analyze and study the fundamental mechanism that drives the characteristics and dynamics of BMI is the underlying social influence. We make an attempt at a unified probabilistic approach, called joint activity and relation (JAR), for modeling and predicting users' activities and user-user relationships simultaneously in a single coherent framework. Instead of incorporating social influence in an ad hoc manner, we show that social influence can be captured quantitatively. Based on JAR, we learn social influence between users and users' personal preferences for both user activity prediction and user-user relation discovery through statistical inference. To address the challenges of the introduced multiple layers of hidden variables in JAR, we propose a new learning algorithm based on expectation maximization (EM) and we further propose a powerful and efficient generalization of the EM based algorithm for model fitting. We show that JAR exploits mutual interactions and benefits, by taking advantage of the learned social influence and users' personal preferences, for enhanced user activity prediction and user-user relation discovery. We further experiment with real world dataset to verify the claimed advantages achieving substantial performance gains.

1 Introduction

With the advent of social networking sites such as Facebook, an unprecedented number of users registered with these sites to engage in interesting activities such as commenting on, liking, and resharing posts and interact with each other to share thoughts. The exponential growth of information repositories and the diversity of users on these sites provide great opportunities and challenges for analyzing and understanding users' behaviors as well as user-user relationships. Doubtlessly, social activity prediction and social relationship discovery have become critical research goals in academia and industry recently in the field of social network analysis, and have played an essentially important role in a variety of world-wide-web applications [Wasserman *et al.*, 1994].

Information diffuses and spreads in socially connected networks [Friedkin, 1982; Bakshy *et al.*, 2012]. The widespread social phenomenon of *homophily* [McPherson *et al.*, 2001] suggests that users socially acquainted tend to behave similarly. The homophily social effect is also called the theory of "birds of a feather flock together" – people tend to follow the behaviors of their friends, and people tend to create relationships with other people who are already similar to them. This phenomenon illustrates high correlation and mutual interactions between users' activities and user-user relations [Wang *et al.*, 2011; Yang *et al.*, 2011; Yu and Xie, 2014a]. Fig. 1 shows an illustrative example of such bidirectional mutual interactions (BMI). On the one hand (Left), the user Kate's behavior (Like IPhone 6) can be influenced by her friend Bob's opinion. Since Kate and Bob are friends, it is likely that they have the same behavior. On the other hand (Right), Kate's behavior could in turn impact her relationships with others such as Bob. Suppose Kate and Bob have similar behaviors, they tend to create the friend relationship with each other. Thus it is highly desirable to leverage behavioral evidences to infer social relations and at the same time exploit relations to predict user behaviors in a unified framework.

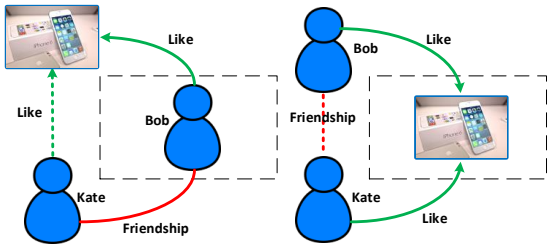

Figure 1: Illustration of bidirectional mutual interactions (BMI) between users' activities and user-user relationships.

We argue that the fundamental mechanism that drives the characteristics and dynamics of BMI is the underlying *social influence*. Social influence here refers to the phenomenon that a user follows an opinion from others, which may or may not deviate from his own interests. The user's ac-

tivities are not solely depend on his own preferences but also influenced by the tastes of other people. Similarly, the social relationship between two users depends not only on their prior impressions to each other but also their behavior agreement. More specifically, we study the social influence quantitatively as the probability that a user follows an opinion from others, for both this user's activities and this user's relationships to others. Intuitively, the influences from different people are essentially different. Furthermore, some people with different tastes may be very influential to a user, while some other people with very similar interests may not contribute too much to this user. Noticeably, our investigated social influence is fundamentally different from social correlation between users' behaviors and relations [Wang *et al.*, 2011], and trust intensities or similarities among users and their friends [Ma *et al.*, 2009; 2011]. Since such approaches only consider social influence by heuristics for coarse and limited measurement.

To address the aforementioned problems, we exploit social influence systematically and quantitatively to a user from others. We propose a unified probabilistic framework to capture BMI between users' activities and user-user relationships. The major contributions in this research are summarized as follows.

- We study the social influence among users on social networking sites systematically and quantitatively, which is the underlying mechanism of bidirectional mutual interactions (BMI) between users' activities and user-user relationships. We propose joint activity and relation (JAR) for modeling and predicting users' activities and user-user relationships simultaneously in a single coherent framework.
- We propose a new learning algorithm based on expectation maximization (EM) to optimize two layers of hidden variables jointly including influential users and latent topics in the JAR model, and we further propose a powerful and efficient generalization of the EM based algorithm for model fitting.
- We demonstrate that the learned social influence and users' personal preferences in the JAR model are very useful for boosting both user activity prediction and user-user relation discovery. We conduct a comprehensive performance evaluation on real world social networking dataset to illustrate the validity and competitiveness of our approach.

2 Our Approach

In this section, we propose the joint activity and relation (JAR) model to explore several important factors contributing to both users' activities and user-user relationships. Besides user preferences, an important factor that JAR captures is the social influence. We aim to learn quantified social influence among individuals for both users' activities and user-user relationships. We first briefly review some notations and research problem formulation. We then present the JAR model in detail.

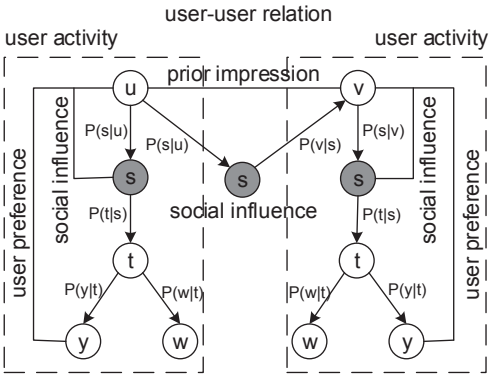

Figure 2: Graphical representation of our JAR model.

2.1 Notations and Problem Description

Let $G = (V, E)$ be a social network graph representing $|V| = N$ users and $|E| = M(E \subset V \times V)$ connections between these users. Let u and v be two different users in the user set V, and e_{uv} be the corresponding connection in E for the user pair (u, v). Let $y = \{y_1, y_2, \cdots, y_N\}(y_i \in \mathcal{Y})$ be activities of the N users, and let $r_{uv}(r_{uv} \in \mathcal{R})$ be the relationship of e_{uv} from u to v. Without loss of generality, both $r_{uv} = r_{vu}$ and $r_{uv} \neq r_{vu}$ are valid settings for undirected and directed networks, respectively. Let $P(y|u)$ be the conditional probability of user u's activity y, and let $P(v|u)$ be the conditional probability of user u's relationship r_{uv} to user v. The joint task of user activity prediction and user-user relationship discovery is to find the most likely activity and relationship for user u such that the joint conditional probability $P(y, v|u)$ is optimized.

2.2 Joint Activity Relation

As stated in the introduction, a user u performs an activity y probabilistically to his own tastes or the preferences from others due to social influence. For user activities, several important factors are needed to take into consideration, including u's own interests and others' possible preferences, u's independence in performing the activity, and social influence from others for the activity. Ultimately, the activity of u is performed implicitly in accordance with the overall collective interests of u's and others'. In summary, activity performing procedure of user u is to draw a preference from the collection based on u's independence and others' influence. As a result, a preference similar to u has a higher probability to be drawn than a different preference from others. Similarly, u creates relationship with v probabilistically to his prior impressions to v or the agreement from others. For user-user relationships, we may conclude similar preference drawing process which captures influence from others.

Fig. 2 shows the graphical representation of our JAR framework, which models activity of user u (left part in the dashed box), activity of user v (right part in the dashed box), and user-user relationship between u and v (middle part). Since the left and right parts have the same structure, we use

the left to illustrate user activity. To capture u's own preferences, we introduce a set of latent topics t and we correlate activities y and their tags w with u through t. Each activity is associated with a set of tags $w = \{w_1, w_2, \cdots, w_m\}(w \in W)$ representing the content of this activity. The latent topic set $t = \{t_1, t_2, \cdots, t_l\}(t \in T)$ is introduced to capture user u's own interests or preferences, and also to characterize activities y and their content w. As can be seen, the activity $y(y \in \mathcal{Y})$ (y is associated with a tag w) of user $u(u \in V)$ is subject to u's own preference (the latent topic t) for u's independent selection and the social influence s from others. Similarly, in the middle part of Fig. 2, u's relationship to v is attributed to u's own prior impression to v and the social influence s.

Without loss of generality, the social influence variable s represents any user directly connected to user u in the social graph G, including both u's close friends and u's loose acquaintances. Let $S(u)(S(u) \subseteq V)$ be the set of all users directly connected to user u. The social influence from s in $S(u)$ to u is measured as the probability of s's own preferences that contributes to u's activity and/or u's relationship to v. It is obvious that sometimes u's independent choice and u's own tastes play more important role for activity y and/or u's relationship to v than the social influence and preferences from $S(u)$. We describe the JAR model in detail as follows. Both the activity of user u and the relationship of user u to user v are probabilistically determined based on u's own interests or the preferences of $s(s \in S(u))$ via social influence. We define the social influence dependency $P(s|u)$ as the probability of user u to be influenced by user s. For user activity, once s is selected based on $P(s|u)$, we randomly draw a topic t from s's interests based on the conditional probability $P(t|s)$. The topic t finally generates an activity y and a tag w based on t's activity distribution $P(y|t)$ and t's content distribution $P(w|t)$. Otherwise, we directly draw a topic t for u's own tastes based on $P(t|u)$. For user-user relationship, once s is selected based on $P(s|u)$, the relationship probability from u to v is measured as $P(s|u)P(v|s)$. Otherwise, the relationship probability from u to v is measured based on u's prior impression as $P(v|u)$.

As can be seen, both user activities and user-user relationships are modeled in a single coherent framework via social influence to capture the BMI between them. We can therefore measure social influence from $S(u)$ quantitatively and efficiently and investigate the effect of social influence for both user activities and user-user relations.

Based on the above discussions and considerations, we now formally define our JAR model. Recall that our goal is to optimize the joint conditional probability $P(y, v|u)$ for most likely activity prediction and relationship discovery. $P(y, v|u)$ could be calculated as

$$P(y, v|u) = \frac{P(y, v, u)}{P(u)} \propto P(y, v, u) \propto \sum_{s \in S(u)} \sum_{t \in T} \sum_{w \in W} P(y, w, s, t, u, v). \quad (1)$$

Based on the JAR model, we assume that activities y and tags w are independently conditioned on the topics t. Consequently the joint probability dependency $P(y, w, s, t, u, v)$ over all factors is decomposed as

$$P(y, w, s, t, u, v) = P(v)P(u|v)P(s|u)P(t|s)P(y|t)P(w|t). \quad (2)$$

From the graphical structure of JAR in Fig. 2, we observe that u and t are independently conditioned on s, and s, y and w are independently conditioned on t, Eq.(2) can be rewritten as follows:

$$\begin{aligned} P(y, w, s, t, u, v) &= P(u|v)P(v)\big[P(s|u)P(t|s)P(y|t)P(w|t)\big] \\ &= P(v|u)P(u)\big[P(s|u)P(t|s)P(y|t)P(w|t)\big] \\ &= P(v|u)\big[P(t)P(u|s)P(s|t)P(y|t)P(w|t)\big]. \end{aligned} \quad (3)$$

Since s and t are hidden variables which are unobserved in the social data, to model the probability $P(y, w, s, t, u, v)$ in terms of s and t, we transform Eq.(3) into the following equation as

$$\begin{aligned} P(s, t|u, v, y, w) &= \frac{P(y, w, s, t, u, v)}{\sum_s \sum_t P(y, w, s, t, u, v)} \\ &= \frac{P(t)P(u|s)P(s|t)P(y|t)P(w|t)}{\sum_s \sum_t P(t)P(u|s)P(s|t)P(y|t)P(w|t)}. \end{aligned} \quad (4)$$

According to Eq.(4), it is required to estimate several model parameters of JAR, including $P(u|s)$, $P(s|t)$, $P(y|t)$, $P(w|t)$ and $P(t)$ to calculate probabilities $P(y, w, s, t, u, v)$ and $P(y, v, u)$.

In summary, JAR jointly incorporates several important factors for simultaneous user activity prediction and user-user relationship inference, including the distribution of social influence $P(u|s)$ from s to u; the distribution of u's own preference $P(t|u)$ over the latent topics t; the distribution of s's preference $P(s|t)$ over the latent topics t; the distribution of activity $P(y|t)$ for each topic t; and the distribution of generated content $P(w|t)$ for each topic t.

3 Learning

Due to the introduced multiple hidden variables s and t in JAR, we meet new challenges for optimization. We perform detailed mathematical derivation and we propose a new and efficient learning algorithm based on expectation maximization (EM) to address this issue. For optimizing the parameters in JAR, we aim to maximize the complete log-likelihood of the social data as follows:

$$\log P(G; \theta) = \log \sum_s \sum_t P(s, t, u, v, y, w; \theta), \quad (5)$$

where s and t represent social influence variables and latent topic variables respectively, and θ represents all parameters of JAR model, including $P(u|s)$, $P(s|t)$, $P(y|t)$, $P(w|t)$, and $P(t)$.

According to Jensen's inequality, we try to maximize the lower bound $\mathcal{L}(\theta)$ of the log-likelihood in Eq.(5) as

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_s \sum_t \log P(s, t, u, v, y, w; \theta) \\ &= \sum_s \sum_t \log \big[P(s, t|u, v, y, w)P(u, v, y, w)\big]. \end{aligned} \quad (6)$$

In the EM based algorithm, the E-step is to compute the expected value of the log-likelihood objective function $\mathcal{L}(\theta)$

with respect to the conditional distribution of latent variables s and t, under the current estimate of the model parameters. This step can be performed according to Eq.(4) and Eq. (6), in which s and t are computed simultaneously such that $P(s|t)$ can be estimated in the M-step.

In the M-step, all the parameters θ are estimated to maximize $\mathcal{L}(\theta)$ in the E-step. Now we maximize $\mathcal{L}(\theta)$ with its parameters by the Lagrangian multiplier method. For example, take the derivation of $\mathcal{L}(\theta)$ with respect to $p(u|s)$ for the following equation:

$$\frac{\partial}{\partial(P(u|s))}\left[\mathcal{L}_{[P(u|s)]} + \lambda(1 - P(u|s))\right] = 0, \quad (7)$$

where $\mathcal{L}_{[P(u|s)]}$ represents terms containing $P(u|s)$ in the objective function $\mathcal{L}$. We can therefore have

$$\lambda = \frac{\sum_{v',y',w'} \sum_{t'} P(s,t'|u,v',y',w')}{\sum_{u,v,y,w} \sum_{s,t} P(s,t|u,v,y,w)}. \quad (8)$$

Finally the update formula of $P(u|s)$ can be obtained as

$$P(u|s) = \frac{\sum_{v',y',w'} \sum_{t'} P(s,t'|u,v',y',w')}{\sum_{u,v,y,w} \sum_{s,t} P(s,t|u,v,y,w)}. \quad (9)$$

Similarly, we can derive the update formulas of $P(s|t)$, $P(y|t)$, $P(w|t)$, and $P(t)$ as follows:

$$P(s|t) = \frac{\sum_{u',y',w'} \sum_{v'} P(s,t|u',v',y',w')}{\sum_{u,v,y,w} \sum_{s,t} P(s,t|u,v,y,w)}, \quad (10)$$

$$P(y|t) = \frac{\sum_{u',v',w'} \sum_{s'} P(s',t|u',v',y,w')}{\sum_{u,v,y,w} \sum_{s,t} P(s,t|u,v,y,w)}, \quad (11)$$

$$P(w|t) = \frac{\sum_{u',v',y'} \sum_{s'} P(s',t|u',v',y',w)}{\sum_{u,v,y,w} \sum_{s,t} P(s,t|u,v,y,w)}, \quad (12)$$

$$P(t) = \frac{\sum_{u',v',y',w'} \sum_{s'} P(s',t|u',v',y',w')}{\sum_{u,v,y,w} \sum_{s,t} P(s,t|u,v,y,w)}. \quad (13)$$

3.1 EM Generalization

The proposed EM-based algorithm optimizes the model parameters iteratively via E-step and M-step until converges to a local optimum. To obtain better model fitting such that it generalizes well on the unseen testing social data, we propose a generalization of the EM-based algorithm, which is known as annealing and is based on an entropic regularization term. Following [Neal and Hinton, 1998], the EM procedure in JAR could be obtained by minimizing a common objective function (also called the free energy) as follows:

$$\mathcal{F}_\gamma = -\gamma \sum_{s,t} \tilde{P}(s,t|u,v,y,w) \log P(u,v,y,w|s,t)P(s,t) + \sum_{s,t} \tilde{P}(s,t|u,v,y,w) \log \tilde{P}((s,t|u,v,y,w), \quad (14)$$

where $\tilde{P}(s,t|u,v,y,w)$ is the variational distribution and γ is the control parameter called inverse computational temperature. In the case of $\tilde{P}(s,t|u,v,y,w) = P(s,t|u,v,y,w)$ minimizing $\mathcal{F}$ w.r.t. the probabilities defining $P(u,v,y,w|s,t)P(s,t)$ amounts to the standard M-step. It is straightforward to verify that the posteriors are obtained by minimizing $\mathcal{F}$ w.r.t. $\tilde{P}(s,t|u,v,y,w)$ at $\gamma = 1$. Essentially, we modify the E-step in Eq.(4) as follows [1]:

$$\tilde{P}(s,t|u,v,y,w) = \frac{\left[P(t)P(u|s)P(s|t)P(y|t)P(w|t)\right]^\gamma}{\sum_s \sum_t \left[P(t)P(u|s)P(s|t)P(y|t)P(w|t)\right]^\gamma}. \quad (15)$$

If $\gamma = 1$ it becomes the standard E-step, while for $\gamma < 1$ the likelihood in Eq.(15) is discounted. We propose a held-out data technology by first performing EM iterations and then decreasing γ until held-out performance deteriorates. We now summarize the overall learning procedure of our JAR model as follows: (1) set $\gamma \leftarrow 1$ and perform EM iterations with early stopping; (2) decrease γ by $\gamma \leftarrow \eta\gamma$ with $\eta < 1$ and perform Eq.(15); (3) as long as the performance on held-out data improves continue to perform Eq.(15) at this γ, otherwise goto (2); (4) stop when decreasing γ does not produce further improvements; (5) perform final iterations using both training and held-out social data. In our experiments, we find that this procedure accelerates the model fitting significantly.

4 Inference

4.1 User Activity Prediction

According to the JAR model, the activity of user u is probabilistically determined based on u's personal interests or the preferences of s via social influence. This implies that both user's personal preference and social influence are critical to accurate activity prediction. Thus the conditional probability of u's activity y can be calculated as follows:

$$P(y|u) = \alpha P_{\text{self}}(y|u) + (1 - \alpha)P_{\text{socinf}}(y|u), \quad (16)$$

where $P_{\text{self}}(y|u)$ is u's activity based on the personal preference, and $P_{\text{socinf}}(y|u)$ is u's activity based on social influence. More importantly, we introduce the factor $\alpha(0 \leq \alpha \leq 1)$ to weight the probability of u's own tastes on his activity. Thus the probability of social influence on his activity is measured as $1 - \alpha$.

Based on u's personal preference, we directly draw a topic t for u's own tastes based on $P(t|u)$ in the JAR model. Consequently $P_{\text{self}}(y|u)$ can be calculated as

$$P_{\text{self}}(y|u) = \sum_t \sum_w P(t|u)P(y|t)P(w|t), \quad (17)$$

where $P(t|u)$, $P(y|t)$, and $P(w|t)$ can be computed efficiently from the learned model parameters. And $P_{\text{socinf}}(y|u)$ is calculated as

$$P_{\text{socinf}}(y|u) = \sum_s \sum_t \sum_w P(s|u)P(t|s)P(y|t)P(w|t). \quad (18)$$

4.2 User-User Relation Inference

Similarly, the conditional probability $P(v|u)$ of user u's relationship to user v is calculated as

$$P(v|u) = \beta P_{\text{self}}(v|u) + (1 - \beta)P_{\text{socinf}}(v|u), \quad (19)$$

[1] We omitted the detailed derivation for space. Please refer to [Hofmann *et al.*, 1999] for more details.

where $P_{\text{self}}(v|u)$ measures the relationship from u to v based on u's prior impressions to v, $P_{\text{socinf}}(v|u)$ measures the relationship from u to v based on social influence, and $\beta(0 \leq \beta \leq 1)$ is another weighting factor. For a given social data, $P_{\text{self}}(v|u)$ is easy to compute. And $P_{\text{socinf}}(v|u)$ can also be computed efficiently as

$$P_{\text{socinf}}(v|u) = \sum_s P(s|u)P(v|s) \propto \sum_s P(u|s)P(v|s) \quad (20)$$

5 Performance Evaluation

5.1 Data

Our dataset is crawled from Twitter.com[2], a widely used microblogging system. It is an online social networking service that enables users to send and read short 140-character messages called tweets. Twitter rapidly gained worldwide popularity. As of December 2014, Twitter has more than 500 million users, out of which more than 284 million are active users. This dataset is comprised of 5,275 users, 22,382 friendship/followership links among them, and 120,285 tweets posted by these users. This dataset also contains 282,450 activities including authoring, commenting on, liking, resharing tweets. We aim to predict these activities and friendship/followership relations among users.

5.2 Experimental Setup

Evaluation metrics. To quantitatively evaluate the proposed model, we use Precision (P), Recall (R), and F1-measure for both activity prediction and relation discovery. We also perform case study on learned social influence and personal preference to illustrate the effectiveness of our model.

Comparison methods. We compare our JAR model with one decoupled method and two joint methods. The decoupled method is Logistic Regression (LR) employed in [Leskovec *et al.*, 2010], which performs activity prediction and relation inference independently. The two joint methods include (1) correlation (CORR) based classifier [Wang *et al.*, 2011] and (2) Friendship-Interest Propagation (FIP) [Yang *et al.*, 2011]. For the CORR method, we use Pearson correlation with optimal threshold.

Methodology. We perform five-fold cross-validation on the Twitter data, and take the average performance. For the generalization of EM algorithm in JAR, we use 60% of the data for initial training and 20% of the data for held-out fitting each time in the cross-validation procedure, and we set parameter $\eta = 0.8$ for model fitting. Thus all the models use the same dataset for both training and testing. All the models exploit the same set of information for features, including users' activities and friendship/followership relations among them. For the JAR model, the contents of tweets are used as tags w.

For our JAR model, we set the latent topic size as $[10, 20, 30, 40, 50, 60]$ to investigate its effect on the performance. For simplicity, the two weighting factors α and β are set to the same values from 0.0 to 1.0, with an incremental step of 0.1. Thus we have $11 \times 6 = 66$ different parameter settings for the JAR model. The best performances of different parameter settings are reported to compare with the baseline models.

[2] https://twitter.com/

Table 1: Comparative performance of user activity prediction

Models	Precision	Recall	F1-measure
LR	69.44	67.37	68.39
CORR	70.89	68.82	69.84
FIP	75.25	73.69	74.46
JAR (EM)	79.36	76.67	78.00
JAR (GEM)	84.68	82.97	83.82

Table 2: Comparative performance of user-user relationship discovery

Models	Precision	Recall	F1-measure
LR	74.55	73.19	73.86
CORR	72.48	69.85	71.14
FIP	79.65	76.33	77.95
JAR (EM)	85.46	83.27	84.35
JAR (GEM)	90.82	88.83	89.81

5.3 Experimental Results

Prediction performance. Table 1 lists the performance of user activity prediction and Table 2 lists the performance of user-user relationship discovery, respectively. The best activity prediction performance of JAR is obtained when the topic size is set to 50 and $\alpha = \beta = 0.8$, and the best performance of relationship discovery is obtained when the topic size is set to 50 and $\alpha = \beta = 0.7$. Compared to JAR (EM), JAR (GEM) is the model exploring generalized EM algorithm for learning, which accelerates the model fitting significantly. In our experiments, the typical number of GEM iterations is less than 100. As can be seen, our proposed JAR model consistently achieves better performance than baseline methods. We perform the paired t-test over the F-measures to verify that all the improvements of our proposed model over the baseline models are statistically significant.

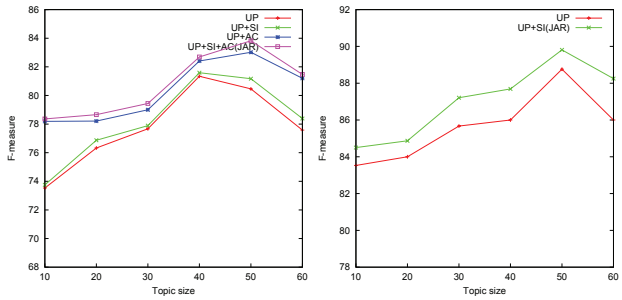

Figure 3: Effect of topic size and factor on the performance of user activity prediction (Left) and user-user relationship discovery (Right).

Effect of topic size and factor contributions. Different topic sizes may effect the performance remarkably. To study the effectiveness of different factors incorporated in the JAR model, we compare the following factor configurations for activity prediction when $\alpha = \beta = 0.8$: (1) users' preferences (UP); (2) both users' preferences and social influence (UP+SI); (3) both users' preferences and activity content (UP+AC); (4) the complete JAR considering all factors

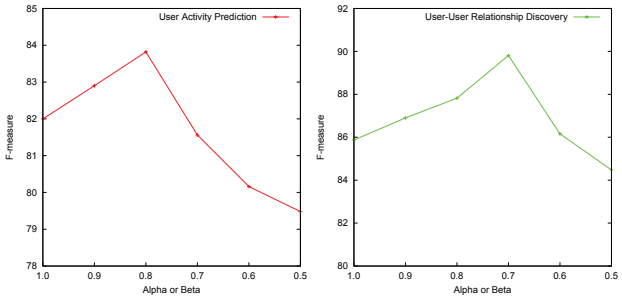

Figure 4: Effect of α and β on the performance of user activity prediction (Left) and user-user relationship discovery (Right).

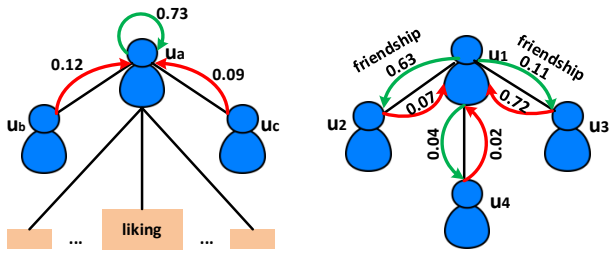

Figure 5: Contributions of personal preference (green arrows with probabilities) and social influence (red arrows with probabilities) on user activities (Left) and user-user relationships (Right).

(UP+SI+AC). For relationship discovery, we also compare the following configurations when $\alpha = \beta = 0.7$: (a) users' preferences (UP); (2) the complete JAR incorporating all factors (UP+SI). As shown in Fig. 3, social influence indeed improves the performance. The complete JAR model consistently performs the best among all configurations.

Effect of parameters α and β. We examine the impact of parameters α and β on the performance, and we fix the topic size to 50 for both activity prediction and relationship discovery. As can be seen from Fig. 4, the curves are typically in a inverted U-shape with the optimal performance achieved at around 0.8 and 0.7, respectively. From these curves, we conclude that (1) both users' preferences and social influence contribute to users' final decision for activities and user-user relationships, (2) both users' activities and user-user relationships are mainly based on their own tastes, although opinions from social influence may affect them to a certain degree.

Table 3: Examples of discovered latent topics by JAR

Topic No.	Representative Words
Topic 5	movie, theater, coffee, bar, clubs, cafe, museum, hotel
Topic 12	store, shop, grocery, service, retails, discount, clothes, food
Topic 27	job, opening, manage, employee, company, recruit, layoff, develop
Topic 33	education, university, school, course, homework
Topic 36	news, newspaper, radio, media, multimedia, broadcast, web
Topic 42	music, artist, song, mtv, play, jazz, hip-hop, rap, fashion

Case study. Both user preference and social influence can be learned quantitatively in our JAR model. We perform case study for a deep understanding of user activities and user-user relationships, and Fig. 5 shows the learned user preference and social influence from our dataset. As can be seen, u_a's activity of liking (a tweet) is mainly based on his personal preference (0.73) rather than the social influence from u_b (0.12) or u_c (0.09). The friendship relation between u_1 and u_2 is attributed to u_1's prior impression (0.63) on u_2, while the friendship between u_1 and u_3 is attributed to social influence (0.72) from u_3. Moreover, Table 3 shows some examples of discovered latent topics by our JAR. These topics are exploited by the joint distributions of activities and contents.

6 Related Work

Exploring joint models to capture mutual benefits between relevant tasks has proven to be highly desirable in data mining and information extraction research communities [Yu *et al.*, 2009; Yu and Lam, 2010; Yu, 2010; Yu *et al.*, 2011; Yu and Lam, 2012]. We review some highly related approaches to joint social actions and relations emphasizing on the differences and advantages of our approach against others. [Yang *et al.*, 2011] proposed joint friendship and interest propagation (FIP), where the behaviors of two socially-connected users and the friendship relations are tied and reinforced by each other. [Bartunov *et al.*, 2012] proposed a CRF based approach to user identity resolution which combines user profile attributes and social linkage. More recently, [Gong *et al.*, 2014] 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. [Yu and Xie, 2014a] and [Yu and Xie, 2014b] presented a coherent unified framework for simultaneous social action prediction and social tie discovery. All the above mentioned approaches only consider social influence by heuristics for coarse and limited measurement. The social influence in these models is not measured and captured quantitatively to take into account core factors for prediction tasks. Thus the effect of social influence on the performance is unclear and largely unexplored.

Considerable work has been conducted for studying social influence. However, most of the existing methods focus on qualitatively study the existence of social influence in different networks. For example, [Wang *et al.*, 2011] investigated high correlation between two individuals' movement similarity and their proximity in the social network. [Bond *et al.*, 2012] utilized a randomized controlled trial to demonstrate the social influence on political voting behavior. [Crandall *et al.*, 2008] investigated the correlation between social similarity and influence. For social influence quantification, [Tang *et al.*, 2009] presented a topical affinity propagation (TAP) approach to quantify the topic-level social influence in large networks. [Goyal *et al.*, 2010] proposed a method to learn the influence probabilities by counting the number of correlated social actions. [Tang *et al.*, 2013] investigated the problem of conformity influence analysis. Despite the fundamental difference in the social influence investigated, we have different research goals from the above mentioned work. We learn

social influence quantitatively to capture BMI for joint and enhanced user activity prediction and user-user relationship discovery. We also systematically study how social influence affect the performance.

7 Conclusions and Future Work

We propose a unified probabilistic framework JAR, which captures social influence quantitatively and incorporates several important factors, for modeling and predicting users' activities and user-user relationships simultaneously. We propose a new and efficient generalization of the EM based algorithm for model learning and fitting. We show that JAR exploits bidirectional mutual interactions and benefits, by taking advantage of the learned social influence and users' personal preferences, for boosted user activity prediction and user-user relation discovery. Empirical study on real world dataset demonstrates the promise of our approach. Several interesting issues are also analyzed and discussed. For the future work, we play to (1) apply and test our approach on other large-scale social network datasets, and (2) further investigate and extend our framework to more general semi-supervised and unsupervised learning scenarios.

References

[Bakshy *et al.*, 2012] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of WWW-12*, pages 519–528, Lyon, France, 2012.

[Bartunov *et al.*, 2012] Sergey Bartunov, Anton Korshunov, Seung-Taek Park, Wonho Ryu, and Hyungdong Lee. Joint link-attribute user identity resolution in online social networks. In *Proceedings of The 6th SNA-KDD Workshop*, Beijing, China, 2012.

[Bond *et al.*, 2012] Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489:295–298, 2012.

[Crandall *et al.*, 2008] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. Feedback effects between similarity and social influence in online communities. In *Proceedings of KDD-08*, pages 160–168, Las Vegas, NV, USA, 2008.

[Friedkin, 1982] Noah E. Friedkin. Information flow through strong and weak ties in intraorganizational social networks. *Social networks*, 3(4):273–285, 1982.

[Gong *et al.*, 2014] Neil Zhenqiang Gong, Ameet Talwalkar, Lester Mackey, Ling Huang, Eui Chul Richard Shin, Emil Stefanov, Elaine Shi, and Dawn Song. Joint link prediction and attribute inference using a social-attribute network. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(2), 2014.

[Goyal *et al.*, 2010] Amit Goyal, Francesco Bonchi, and Laks V.S. Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of WSDM-10*, pages 241–250, New York City, NY, USA, 2010.

[Hofmann *et al.*, 1999] Thomas Hofmann, Jan Puzicha, and Michael I. Jordan. Learning from dyadic data. In *Proceedings of NIPS-99*, pages 466–472, 1999.

[Leskovec *et al.*, 2010] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of WWW-10*, pages 641–650, Raleigh, North Carolina, USA, 2010.

[Ma *et al.*, 2009] Hao Ma, Irwin King, and Michael R. Lyu. Learning to recommend with social trust ensemble. In *Proceedings of SIGIR-09*, pages 203–210, Boston, MA, USA, 2009.

[Ma *et al.*, 2011] Hao Ma, Irwin King, and Michael R. Lyu. Learning to recommend with explicit and implicit social relations. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–19, 2011.

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

[Neal and Hinton, 1998] Radford M. Neal and Geoffrey E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in Graphical Models*, pages 355–368, 1998.

[Tang *et al.*, 2009] Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. Social influence analysis in large-scale networks. In *Proceedings of KDD-09*, pages 807–816, Paris, France, 2009.

[Tang *et al.*, 2013] Jie Tang, Sen Wu, and Jimeng Sun. Confluence: Conformity influence in large social networks. In *Proceedings of KDD-13*, pages 347–355, Chicago, Illinois, USA, 2013.

[Wang *et al.*, 2011] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-László Barabási. Human mobility, social ties, and link prediction. In *Proceedings of KDD-11*, pages 1100–1108, San Diego, California, USA, 2011.

[Wasserman *et al.*, 1994] Stanley Wasserman, Katherine Faust, Dawn Iacobucci, and Mark Granovetter. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[Yang *et al.*, 2011] Shuang-Hong Yang, Bo Long, Alex Smola, Narayanan Sadagopan, Zhaohui Zheng, and Hongyuan Zha. Like like alike: joint friendship and interest propagation in social networks. In *Proceedings of WWW-11*, pages 537–546, Hyderabad, India, 2011.

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

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

[Yu and Xie, 2014a] Xiaofeng Yu and Junqing Xie. Learning interactions for social prediction in large-scale networks. In *Proceedings of CIKM-14*, pages 161–170, Shanghai, China, 2014.

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

[Yu *et al.*, 2009] Xiaofeng Yu, Wai Lam, and Bo Chen. An integrated discriminative probabilistic approach to information extraction. In *Proceedings of CIKM-09*, pages 325–334, Hong Kong, China, 2009.

[Yu *et al.*, 2011] Xiaofeng Yu, Irwin King, and Michael 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.

[Yu, 2010] Xiaofeng Yu. *Probabilistic models for information extraction: from cascaded approach to joint approach*. PhD thesis, The Chinese University of Hong Kong, December 2010.