5

Learning to Predict Reciprocity and Triadic Closure in Social Networks

TIANCHENG LOU and JIE TANG, Tsinghua University
JOHN HOPCROFT, Cornell University
ZHANPENG FANG and XIAOWEN DING, Tsinghua University

We study how links are formed in social networks. In particular, we focus on investigating how a reciprocal (two-way) link, the basic relationship in social networks, is developed from a parasocial (one-way) relationship and how the relationships further develop into triadic closure, one of the fundamental processes of link formation.

We first investigate how geographic distance and interactions between users influence the formation of link structure among users. Then we study how social theories including homophily, social balance, and social status are satisfied over networks with parasocial and reciprocal relationships. The study unveils several interesting phenomena. For example, "friend's friend is a friend" indeed exists in the reciprocal relationship network, but does not hold in the parasocial relationship network.

We propose a learning framework to formulate the problems of predicting reciprocity and triadic closure into a graphical model. We demonstrate that it is possible to accurately infer 90% of reciprocal relationships in a Twitter network. The proposed model also achieves better performance (+20–30% in terms of F1-measure) than several alternative methods for predicting the triadic closure formation.

Categories and Subject Descriptors: J.4 [**Social and Behavioral Sciences**]: Miscellaneous; H.3.3 [**Information Search and Retrieval**]: Text Mining

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Social network, reciprocal relationship, social influence, predictive model, link prediction, Twitter

ACM Reference Format:
Lou, T., Tang, J., Hopcroft, J., Fang, Z., and Ding. X. 2013. Learning to predict reciprocity and triadic closure in social networks. ACM Trans. Knowl. Discov. Data. 7, 2, Article 5 (July 2013), 25 pages.
DOI: http://dx.doi.org/10.1145/2499907.2499908

1. INTRODUCTION

Online social networks (e.g., Twitter, Facebook, Myspace) significantly enlarge our social circles. The structure of the networks governs the dynamics of the networks

T. Lou and X. Ding are supported in part by the National Basic Research Program of China grant 2011CBA00300, 2011CBA00301, the National Natural Science Foundation of China grant 61033001, 61061130540, 61073174. J. Tang is supported by the Natural Science Foundation of China (no. 61073073, no. 612222212), National Basic Research Program of China (no. 2011CB302302), and Chinese National Key Foundation Research (no. 60933013, no. 61035004). J. Hopcroft was partially supported by the U.S. Air Force Office of Scientific Research under grant FA9550-09-1-0675.
Authors' addresses: T. Lou (corresponding author), Institute for Interdisciplinary Information Sciences, Tsinghua University; email: tiancheng.lou@gmail.com; J. Tang, Department of Computer Science and Technology, Tsinghua University; J. Hofcroft, Department of Computer Science, Cornell University; Z. Fang, Department of Computer Science and Technology, Tsinghua University; X. Ding, Institute for Interdisciplinary Information Sciences, Tsinghua University.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2013 ACM 1556-4681/2013/07-ART5 $15.00
DOI: http://dx.doi.org/10.1145/2499907.2499908

(e.g., information propagation and users' behavior changes). In a social network with directed links such as Twitter, the relationship between users often starts by one user (A) creating a "follow" (parasocial) relationship to another user (B). User B can choose to "follow" A back, which results in a reciprocal relationship between them. On the other hand, if A follows B, and continues to follow B's followee C, then (A, B, C) forms a directed closure triad. This phenomenon is also referred to as "link copying" [Romero and Kleinberg 2010].

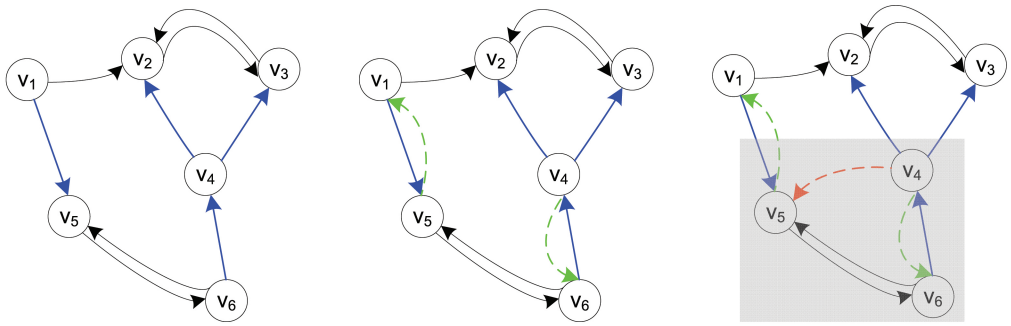
In social science, relationships between individuals are classified into two categories: one-way (called parasocial) relationships and two-way (called reciprocal) relationships [Horton and Wohl 1956]. The most common form of the former are one-way relationships between celebrities and fans, while the most common form of the latter are two-way relationships between close friends. Twitter and Facebook are respectively typical examples of the two types of social relationships. Relationship is the basic object in social network analysis [Weber 1991]. It forms the basis of the social structure. Understanding the formation of social relationships can give us insights into the microlevel dynamics of the social network, such as how an individual user influences her/his friends through different types of social ties [Tang et al. 2009], how friendships have been created across different networks [Tang et al. 2012b], and how a user's opinion spreads in the social network [Tan et al. 2011].

Two interesting questions arise: How is a reciprocal relationship developed from a parasocial relationship and how do pairwise relationships further develop into a triadic closure? Employing Twitter as the basis of our analysis, we try to answer these questions. In particular, when you follow a user on Twitter, how likely is it that the user will follow you back? Some users only follow back those who are real "friends" in their physical world, while some other users (even some top users with tens of thousands of followers) will follow everyone back.[1] This problem also implicitly exists in other social networks such as Facebook and LinkedIn: when you send a friend request to somebody, how likely will she/he confirm your request? How likely will two connected pairwise friendships finally form a closure triad?

Previous research on social relationships can be classified into three categories: link prediction [Liben-Nowell and Kleinberg 2007; Romero and Kleinberg 2010; Leskovec et al. 2010; Backstrom and Leskovec 2011], relationship type inferring [Eagle et al. 2009; Crandall et al. 2010; Wang et al. 2010; Tang et al. 2011], and social behavior prediction [Backstrom et al. 2008; Tan et al. 2010; Yang et al. 2010]. Backstrom and Leskovec [2011] propose an approach called supervised random walk to predict and recommend links in social networks. Crandall et al. [2010] investigate the problem of inferring social ties between people from co-occurrence in time and space. Wang et al. [2010] propose an unsupervised algorithm to infer advisor-advisee relationships from a publication network. However, little research systematically studies how two-way relationships are developed from one-way relationships. Tang et al. [2012a] develop a framework for inferring social ties by learning across heterogeneous networks. Romero and Kleinberg [2010] study the triadic closure process on Twitter. However, they do not give a principled model for predicting the formation of a closure triad. More importantly, what are the fundamental factors that essentially influence the formation of reciprocal relationships and directed triadic closure? And how can existing social theories (e.g., structural balance theory and homophily) be connected to the link formation process?

In this article, we try to conduct a systematic investigation on the problem of predicting reciprocity and triadic closure formation. We precisely define the problem and propose a Triad Factor Graph (TriFG) model. The TriFG model incorporates social theories into a semisupervised learning model, where we have some labeled training data

[1]http://socialnewswatch.com/top-twitter-users/.



(a) Following network at time t (b) Follow back at $(t+1)$ (c) Closure triadic at $(t+2)$

Fig. 1. Motivating example. (a) is the input of our problem: a following network, where the blue arrows indicate new following relationship created at time t. (b) is the output network with follow back relationships, where green dash arrows indicate the follow back relationships developed at time $(t+1)$. (c) is the network with a closure triad, where a new follow relationship denoted as a red dash arrow is created at time $(t+2)$ which forms a directed closure triad.

(reciprocal relationships) but with low reciprocity [Kwak et al. 2010]. For reciprocity prediction, given a historic log of users following actions from time 1 to t, we try to learn a predictive model to infer whether user A will add a follow-back link to user B at time $(t+1)$ if user B creates a new follow link to user A at time t. For triadic closure prediction, we try to infer, when A follow back B at time t, whether user A will add a new follow link to B's followee C at time $(t+1)$. Figure 1 shows an illustrative example of the addressed problem. Figure 1(a) is the input of our problem: a following network, where the blue arrows indicate new following relationship created at time t. Figure 1(b) is the network with follow-back relationships, where green dash arrows indicate the follow-back relationships developed at time $(t+1)$. Figure 1(c) is the network with a closure triad, where a new follow relationship is created at time $(t+2)$ which forms a (directed) closure triad among users v_4, v_5, and v_6. Our goal in this work is to infer the formation of the new links in Figures 1(b) and 1(c) based on the available information at the previous timestamp.

Results. We evaluate the proposed model on a Twitter data consisting of 13,442,659 users and their profiles, tweets, following behaviors (new following or follow-back links) for nearly two months. We show that incorporating social theories into the proposed factor graph model can significantly improve the performance (+22–27% by F1-measure) for predicting reciprocity and (+20–28%) for predicting triadic closure compared with several alternative methods. Our study also reveals several interesting phenomena.

(1) Elite users (opinion leader) tend to follow each other. The likelihood of an elite user following back another elite user is nearly 8 times higher than that of two ordinary users and 30 times that of an elite user and an ordinary user.
(2) Reciprocal relationships on Twitter are balanced, but parasocial relationships are not. More than 88% of social triads (groups of three people) with reciprocal relationships satisfy the social balance theory, while parasocial relationships are unbalanced (only 29% of them satisfy the balance theory).
(3) Social networks are going global, but also stay locally. No matter how far a user is from one by geospatial distance, the likelihood that she/he will follow one back is almost the same, while on the other hand, the number of reciprocal relationships between users within the same time zone is 20 times higher than the number of users from different time zones.

(4) Elite users play an important role for developing triadic closure. The likelihood to form a closure triad when an elite user follows back an ordinary user is 10 times higher than that of an ordinary user following back an elite user.

Organization. Section 2 formulates the problem. Section 3 introduces the dataset and our analyses on the dataset. Section 4 explains the proposed model and describes the algorithm for learning the model. Section 5 presents experimental results that validate the effectiveness of our methodology. Finally, Section 6 reviews the related work and Section 7 concludes this work.

2. PROBLEM DEFINITION

In this section, after presenting several definitions, we formally define the targeted problem in this work. We formulate the problem in the context of Twitter to keep things concrete, though adaptation of this framework to other social network settings is straightforward.

The Twitter network can be modeled as a directed graph $G = \{V, E\}$, where $V = \{v_1, v_2, \ldots, v_n\}$ is the set of users, and $E \subseteq V \times V$ is the set of directed links between users. For easy explanation in the model, we write each edge as e_i with its two end-users as v_i^s and v_i^u. Each directed link $e_i \in E$ indicates that user v_i^s follows user v_i^u. Usually, we also call v_i^s as the follower of v_i^u and v_i^u as the followee.

The Twitter network is dynamic in nature, with links added and removed over time. Our preliminary statistics on a large Twitter dataset show that users tend to add new links much more frequently than to remove existing links (e.g., 95% of changes to links are adding new links). That is to say, adding new links seems to be a more important behavior in forming the structure of the Twitter network. A new link results when a user performs a behavior of following another user in Twitter. Particularly, we define two types of link behaviors.

Definition 2.1. *New Follow and Follow Back.* Suppose at time t, user v_i creates a link to v_j, who has no previous link to v_i, then we say v_i performs a new-follow behavior on v_j. When user v_i creates a link to v_j at time t, who already has a link to v_i before time t, we say v_i performs a follow-back behavior on v_j.

The new-follow and follow-back behaviors respectively correspond to the one-way (parasocial) relationship and the two-way (reciprocal) relationship in sociology. In this work, we focus on investigating the formation of follow-back behaviors. For simplicity, let $y_i^t = 1$ denote that user v_i^s follows back v_i^u at time t and $y_i^t = 0$ denote user v_i^s does not follow back. We are concerned with the following prediction problem.

Problem 1 (*Follow Back Prediction*). Let $< 1, \ldots, t >$ be a sequence of timestamps with a particular time granularity (e.g., day, week, etc.). Given Twitter networks from time 1 to t, $\{G^t = (V^t, E^t, Y^t)\}$, where Y^t is the set of follow-back behaviors at time t, the task is to find a predictive function

$$f : (\{G^1, \ldots, G^t\}) \rightarrow Y^{(t+1)},$$

such that we can infer the follow-back behaviors at time $(t + 1)$.

We further define the triadic closure prediction problem.

Problem 2 (*Triadic Closure Prediction*). Given Twitter networks from time 1 to t, $\{G^t = (V^t, E^t, X^t, Y^t)\}$, where X^t is the set of follow-back behaviors, for example, $v_i \rightarrow v_j$ at time t, the task is to find a predictive function f to infer whether v_i will create a new-follow link $y_{ik} \in Y^{(t+1)}$ to v_j's followee v_k at time $(t + 1)$ such that (v_i, v_j, v_k) forms a closure triad structure.

It bears pointing out that our problem is very different from existing social tie inferring [Diehl et al. 2007; Eagle et al. 2009; Crandall et al. 2010; Leskovec et al. 2010; Tang et al. 2011, 2012a], link prediction [Liben-Nowell and Kleinberg 2007; Romero and Kleinberg 2010; Backstrom and Leskovec 2011], and social action prediction problems [Tan et al. 2010; Yang et al. 2010]. First, as the Twitter network is evolving over time, it is infeasible to collect a complete network at time t. Thus it is important to design a method that could take into consideration the unlabeled data as well. Second, it is unclear what are the fundamental factors that influence the formation of follow-back relationships. Finally, one needs to incorporate the different factors (e.g., social theories, statistics, and our intuitions) into a unified model to better predict the follow-back relationship.

3. DATA AND OBSERVATIONS

3.1. Data Collection

We aim to find a large set of users and a continuously updated network among these users, so that we can use the dataset as the gold standard to evaluate different approaches for our prediction. To begin the collection process, we select the most popular user on Twitter, that is, "Lady Gaga", and randomly collect 10,000 of her followers. We take these users as seed users and use a crawler to collect all followers of these users by traversing following edges. We continue the traversing process, which produces in total 13,442,659 users and 56,893,234 following links, with an average of 728,509 new links per day. The crawler monitors the change of the network structure from 10/12/2010 to 12/23/2010. We also extract all tweets posted by these users and in total there are 35,746,366 tweets.

In our analysis, we consider the geographic location of each user. Specifically, we first extract the location from the profile of each user[2], and then feed the location information to the Google Map API to fetch its corresponding longitude and latitude values. In this way, we obtain the longitude and latitude of about 59% of users in our dataset. More detailed analysis, source-code, and an online demonstration are publicly available. http://reciprocal.aminer.org/

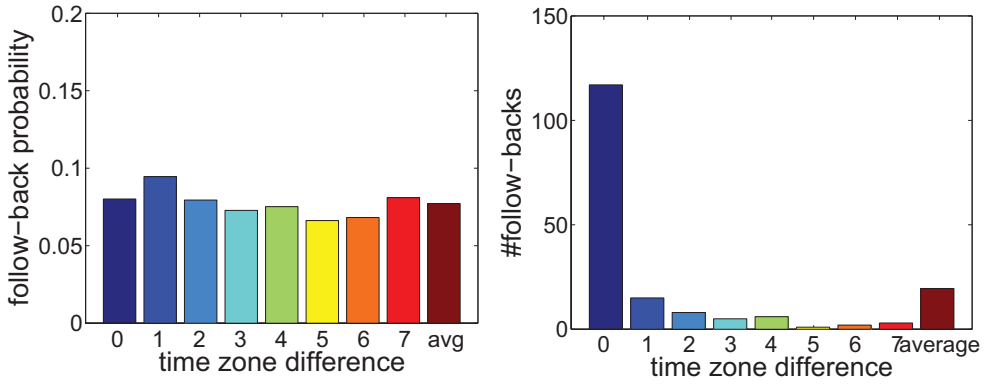
3.2. Observations

We first engage in some high-level investigation of how different factors influence the formation of reciprocity and triadic closure, since one major motivation of our work is to find the underlying factors and their influence to this task. In particular, we study the interplay of the following factors with the formation of follow-backs (or triadic closure).

—*Geographic distance.* Do users have a higher probability to follow each other when they are located in the same region?
—*Homophily.* Do similar users tend to follow each other? We make the analysis for both follow-back and triadic closure predictions.
—*Implicit network.* How does the following network on Twitter correlate with other implicit networks, for example, retweet and reply network?
—*Social balance.* Does the reciprocal relationship network on Twitter satisfy the social balance theory [Easley and Kleinberg 2010]? To which extent?

Geographic distance. Figure 2 shows the correlation between geographic distance and the probability that two users create a reciprocal relationship. Interestingly, it seems that online social networks indeed go global: Figure 2(a) shows the likelihood of a user following another user back when they are from the same time zone or from

[2]For example, Lady Gaga's location information is: "Location: New York, NY".



(a) Global (b) Local

Fig. 2. Geographic distance correlation. x-axis: time zone difference (0 indicates that users are located in the same time zone); y-axis: (a) probability that one user follows back another user, conditioned on the time zone difference of the two users. (b) number of reciprocal relationships among users from the same time zone or different time zones.

different time zones. Clearly, the geographic distance is already not a major issue to stop users from developing a (reciprocal) relationship. Figure 2(b) shows another statistic which indicates a different perspective that the Twitter network (in some sense) still stays local: the average number of reciprocal relationships between users from the same time zone is about 50 times higher than the number between users with a distance of three time zones.

Homophily. The principle of homophily [Lazarsfeld and Merton 1954] suggests that users with similar characteristics (e.g., social status, age) tend to associate with each other. In particular, we study two kinds of homophilies on the Twitter network: link homophily and status homophily. For the link homophily, we test whether two users who share common links (followers or followees) will have a tendency to associate with each other. Figure 3 clearly shows that the probability of two users following back each other when they share common neighbors is much higher than usual. When the number of common neighbors with two-way relationships increases to 3, the likelihood of two users following back each other also triples. The effect is more pronounced when the number increases to 10. But it is worth noting that this only works for reciprocal relationships and does not hold for the parasocial relationship (as indicated in Figure 3).

For the status homophily, we test whether two users with similar social status are more likely to associate with each other. We categorize users into two groups (elite users and ordinary users) by three different algorithms: PageRank [Page et al. 1999][3], #degree, and (α, β) algorithm [He et al. 2011][4]. Specifically, with PageRank, we estimate the importance of each user according to the network structure, and then select top 1% users[5] who have the highest PageRank scores as elite users and the rest as ordinary users; while with #degree, we select top 1% users with the highest number of indegree as elite users and the rest as ordinary users. For (α, β), we input the size of the core community as 200, and after running the algorithm, we use users selected in the core community as elite users and the rest as ordinary users. Then, we examine the difference of follow-back behaviors among the two groups of users. Figure 4 clearly shows

[3]PageRank is an algorithm to estimate the importance of each node in a network.
[4](α, β) algorithm is designed to find core members (elite users) in a social network.
[5]Statistics have shown that less than 1% of the Twitter users produce 50% of its content [Wu et al. 2011].

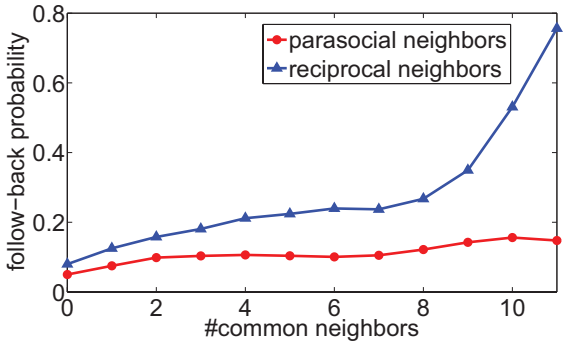


Fig. 3. Link homophily. y-axis: probability that two users follow back each other, conditioned on the number of common neighbors of two-way (reciprocal) relationships or one-way (parasocial) relationships.

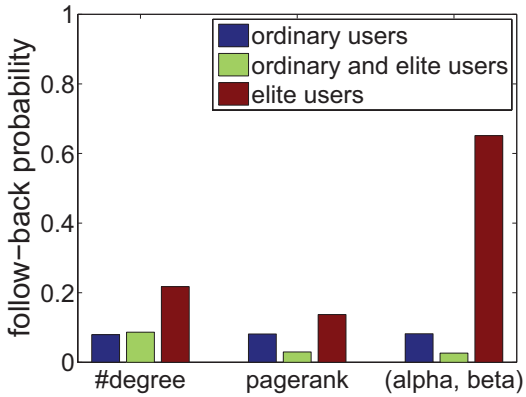


Fig. 4. Status homophily by different algorithms. y-axis: probability that two users follow back each other, conditioned on whether the two users are from the same group of elite/ordinary users or from different groups. #Degree, PageRank, and (α, β) are three algorithms to distinguish elite users from ordinary users.

that, though the three algorithms present different statistics, "elite" users have a much stronger tendency to follow each other: the likelihood of two elite users following back each other is nearly 8 times higher than that of ordinary users (by the (α, β) algorithm). The (α, β) algorithm seems able to better distinguish elite users from ordinary users in our problem setting. This is because besides the global network structure, the (α, β) algorithm also considers the community structure among elite users.

Implicit structure. On Twitter, besides the explicit network with following links, there are also some implicit network structures that can be induced from the textural information. For example, user A may mention user B in her tweet, that is, "@B", which is called a reply link; user A may forward user B's tweet, which results in a retweet link. We study how the implicit links correlate with the formation of the follow-back relationship on Twitter. Figure 5 clearly shows that when users A and B retweet or reply to each other's tweets, the likelihood of their following back each other is higher (3 times higher than chance). Another interesting phenomenon is that compared with replying to someone's tweet, retweeting (forwarding) her tweet seems to be more helpful (15% versus. 9%) to win her follow-back.

Structural balance. Now, we connect our work to a basic social psychological theory: structural balance theory [Easley and Kleinberg 2010]. Let us first explain the

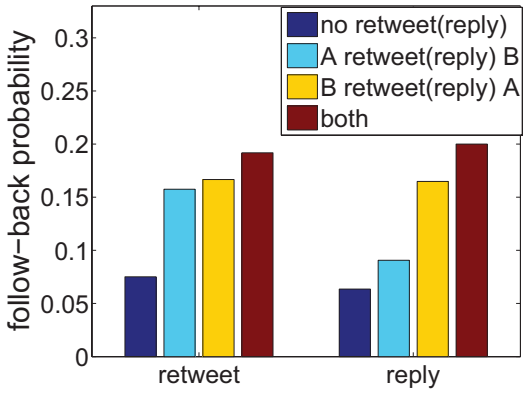


Fig. 5. Implicit network correlation. y-axis: probability that user B follows user A back, conditioned on one user (A or B) retweets or replies the other user's tweet.

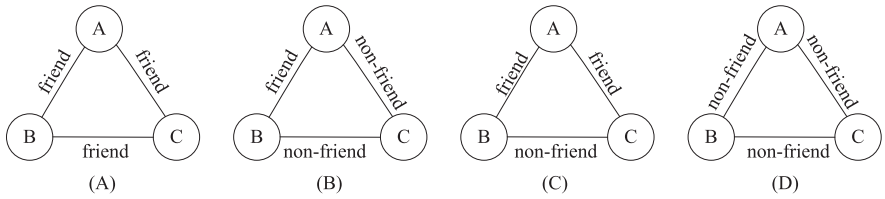


Fig. 6. Illustration of structural balance theory. (A) and (B) are balanced, while (C) and (D) are not balanced.

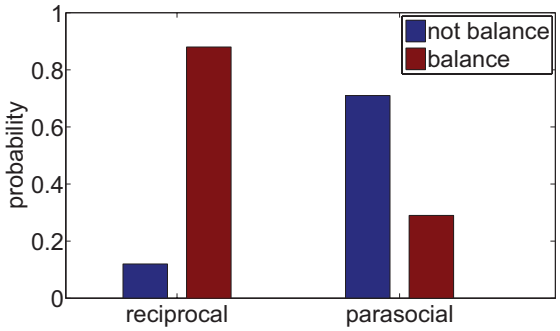
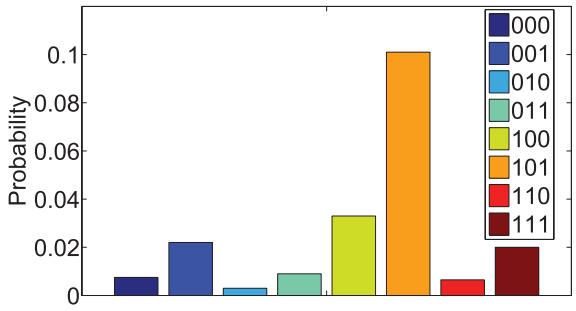
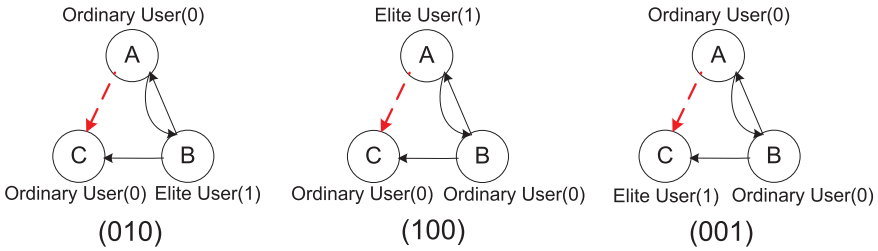


Fig. 7. Structural balance correlation. y-axis: probability that a triad creates two-way (reciprocal) relationships, conditioned on whether the resultant structure is balanced or not.

structural balance property. For every group of three users (called triad), the balance property implies that either all three of these users are friends or only one pair of them are friends. Figure 6 shows such an example. To adapt the theory to our problem, we can map either the reciprocal relationship or the parasocial relationship on the friendship. Then we examine how the Twitter network (only reciprocal relationships or parasocial relationships) satisfies the structural balance property. More precisely, we compare the probabilities of the resultant triads that satisfy the balance theory based on reciprocal relationships and parasocial relationships on Twitter. Figure 7 clearly shows that it is much more likely (88%) for users to be connected with a balanced structure of reciprocal relationships, while with parasocial relationships, the resultant structure is very unbalanced. This is because two users are very likely to follow a same movie star, but they do not know each other, which results in a unbalanced triad (Figure 6(c)).



(a) triadic status correlation



(b) illustration of triadic code (010, 100, 001)

Fig. 8. Triadic status correlation. y-axis: probability of triadic closure, conditioned on the social status of the three users (A, B, and C). The three digits on x-axis represent the status of the three users (A, B, and C, with 1 indicating elite user and 0 indicating ordinary user) and y-axis represents the probabilities of different categories of users who formed triadic closure.

We now present some observations of the formation of triadic closure. We focus on studying how users' status and activity influence the formation of the triadic closure.

Triadic status. We examine the correlation between users' social status and triads associated with them. We divide users into two categories (elite users and ordinary users). For simplicity, we select the top 200 users with the highest indegree as elite users, and the others as ordinary users. Then we study the probability of A creating a new follow link to B's followee C, when A follows back B, conditioned on the status of A, B, and C. Figure 8 shows the analysis result. The three digits on the x-axis represent the status of the three users (A, B, and C, with 1 indicating elite user and 0 indicating ordinary user) and y-axis represents the probabilities of different categories of users who formed triadic closure. We find a striking pattern that the highest probability is resulted by 101 (high status, low status, high status), which means that it is very likely a high-status user spends time investigating whom a low-status user follows, when she/he follows back the low-status user, and finally follows some high-status followees of the low-status user. The likelihood is almost ten times higher than chance. Another interesting phenomenon is that when a low-status user A follows back another low-status user B, the likelihood of A following a low-status followee of B is very low (about 0.005%), while the likelihood of A following a high-status followee of B is much higher (4 times higher). Some other interesting patterns can be summarized as follows.

—$P(1XX) > P(0XX)$. Elite users play a more important role to form the triadic closure. The average probability of $1XX$ is three times higher than that of $0XX$. Here X indicates any status (ether high status or low status).

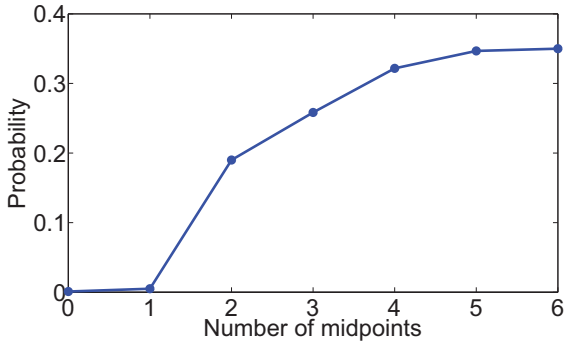


Fig. 9. Number of midpoints correlation. y-axis: probability that relationship can be established, conditioned on number of midpoints.

—$P(X0X) > P(X1X)$. Low-status users act as a bridge to connect users so as to form a closure triad. The likelihood of $X0X$ is 2.8 times higher than $X1X$.
—$P(XX1) > P(XX0)$. The rich get richer. This result validates the mechanism of preferential attachment [Newman 2001].

Link homophily. Similar to the analysis to follow-back, we test whether users who share common links (followers or followees) will have a tendency to form a closure triad. Figure 9 shows the probability of user A following user C, conditioned on the number of common neighbors. It clearly shows that when the number is one or zero, the probability is very low, while there is a sharp increase when the number becomes two. After that, the sublinear behavior takes over. The deviation at $0, 1, 2$ can be seen as a slight "S-shaped" effect: the plots mainly show sublinear increase, while we observe a superlinear between 1 and 2.

In summary, according to the preceding statistics, we have the following observations.

(1) Geographic distance has a pronounced effect on the number of reciprocal relationships created between users, but little effect on the likelihood of users following back each other.
(2) Users with common friends (reciprocal relationships) tend to follow each other.
(3) Elite users have a much stronger tendency (status homophily) to follow each other than ordinary users.
(4) The implicit networks of retweet or reply links have a strong correlation with the formation of two-way (reciprocal) relationships.
(5) The network of reciprocal relationships on Twitter is balanced (88% of triads satisfying the structural balance property), while the network of parasocial relationships is unbalanced (71% are unbalanced).
(6) Elite users play an important role for developing triadic closure. The probability of an elite user developing a closure triad is almost ten times higher than chance.

4. MODEL FRAMEWORK

In this section, we propose a novel Triad Factor Graph (TriFG) model to incorporate all the information within a single entity for better modeling and predicting the formation of reciprocal relationships and triadic closure.

For an edge $e_i \in E$, if user v_i^s follows v_i^u at time t, our task is to predict whether user v_i^u will follow v_i^s back, that is, $y_i = 1$ or 0. For the follow-back prediction task, we assume that v_i^s follows v_i^u at time t, and our task is to predict whether v_i^u will follow v_i^s back at time $(t + 1)$. Based on the observations in Section 3, we define a number of

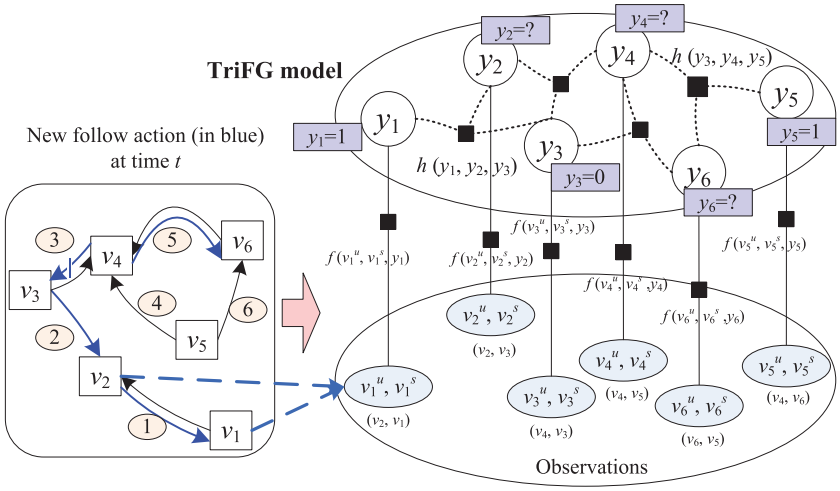


Fig. 10. Graphical representation of the TriFG model. The left figure shows the follow network at time t. Blue arrows indicate new-follow actions, black arrows indicate previously existing follow links, and blue $\nrightarrow$ indicates user v_i^u does not follow user v_i^s back. The right figure is the TriFG model derived from the following graph. Each gray eclipse indicates relationship (v_i^u, v_i^s) between users and each white circle indicates the hidden variable y_i. $f(v_i^s, v_i^u, y_i)$ represents an attribute factor function and $h(.)$ represents a triad factor function.

attributes for each edge, denoted as $\mathbf{x}_i$. The $|E| \times d$ attribute matrix $\mathbf{X}$ describes edge-specific characteristics, where d is the number of attributes. For example, on Twitter, an attribute can be defined as whether two end-users are from the same time zone. An element x_{ij} in the matrix $\mathbf{X}$ indicates the j^{th} attribute value of edge e_i. For the triadic closure prediction task, we assume that v_i^s follows back v_i^u at time t, and our task is to predict whether v_i^u will follow v_i^s's followees at time $(t + 1)$. For easy explanation, we will mainly use the follow-back prediction in our explanation and the extension to the triadic closure prediction is straightforward.

4.1. The Proposed Model

The name of the Triad Factor Graph (TriFG) model is derived from the idea that we incorporate social theories (structural balance and homophily) over triads into the factor graph model.

Figure 10 shows the graphical structure of the TriFG model. The left figure shows the following network of six users at time t. Blue arrows indicate new-follow actions, black arrows indicate follow actions performed before time t, and blue $\nrightarrow$ indicates user v_i^u does not follow user v_i^s back at time t. The right figure is the factor graph model derived from the left input network. Each gray ellipse indicates a relationship (v_i^u, v_i^s) between users and each white circle indicates the hidden variable y_i, with $y_i = 1$ representing v_i^u performs a follow-back action, $y_i = 0$ not, and $y_i =?$ unknown, which actually is the variable we need to predict. Factor $h(.)$ represents a balance factor function defined on a triad; and $f(v_i^s, v_i^u, y_i)$ (or $f(\mathbf{x}_i, y_i)$) represents a factor to capture the information associated with edge e_i.

Given a network at time t, that is, $G^t = (V^t, E^t, X^t)$ with some known variables $y = 1$ or 0 and some unknown variables $y =?$, our goal is to infer values of those unknown variables. For simplicity, we remove the superscript t if there is no ambiguity. We begin with the posterior probability of $P(Y|\mathbf{X}, G)$, according to Bayes' theorem, we have

$$P(Y|\mathbf{X}, G) = \frac{P(\mathbf{X}, G|Y)P(Y)}{P(\mathbf{X}, G)} \propto P(\mathbf{X}|Y) \cdot P(Y|G), \tag{1}$$

where $P(Y|G)$ denotes the probability of labels given the structure of the network and $P(\mathbf{X}|Y)$ denotes the probability of generating the attributes $\mathbf{X}$ associated with each edge given their label Y. Assuming that the generative probability of attributes given the label of each edge is conditionally independent, we get

$$P(Y|\mathbf{X}, G) \propto P(Y|G) \prod_i P(\mathbf{x}_i|y_i), \tag{2}$$

where $P(\mathbf{x}_i|y_i)$ is the probability of generating attributes $\mathbf{x}_i$ given the label y_i. Now, the problem is how to instantiate the probabilities $P(Y|G)$ and $P(\mathbf{x}_i|y_i)$. In principle, they can be instantiated in different ways. In this work, we model them in a Markov random field, and thus by the Hammersley-Clifford theorem [Hammersley and Clifford 1971], the two probabilities can be instantiated as

$$P(\mathbf{x}_i|y_i) = \frac{1}{Z_1} \exp\left\{\sum_{j=1}^{d} \alpha_j f_j(x_{ij}, y_i)\right\}, \tag{3}$$

$$P(Y|G) = \frac{1}{Z_2} \exp\left\{\sum_c \sum_k \mu_k h_k(Y_c)\right\}, \tag{4}$$

where Z_1 and Z_2 are normalization factors. Eq. (3) indicates that we define a feature function $f_j(x_{ij}, y_i)$ for each attribute x_{ij} associated with edge e_i and α_j is the weight of the j^{th} attribute; while Eq. (4) represents that we define a set of correlation feature functions $\{h_k(Y_c)\}_k$ over each triad Y_c in the network. Here μ_k is the weight of the k^{th} correlation feature function.

Based on Eqs. (2)–(4), we define the following log-likelihood objective function $\mathcal{O}(\theta) = \log P_\theta(Y|\mathbf{X}, G)$.

$$\mathcal{O}(\theta) = \sum_{i=1}^{|E|} \sum_{j=1}^{d} \alpha_j f_j(x_{ij}, y_i) + \sum_c \sum_k \mu_k h_k(Y_c) - \log Z \tag{5}$$

Here Y_c is a triad derived from the input network, $Z = Z_1 Z_2$ is a normalization factor, and $\theta = (\{\alpha\}, \{\mu\})$ indicates a parameter configuration. One example of factor decomposition is shown in Figure 10. There are six edges, three with known variables (two $y = 1$ and one $y = 0$) and three with unknown values ($y =$?). We have four triads (e.g., $Y_c = (y_1, y_2, y_3)$) based on the structure of the input network. For each edge, we define a set of factor functions $f(v_i^s, v_i^u, y_i)$ (also written as $f(\mathbf{x}_i, y_i)$).

We now briefly introduce possible ways to define the factor functions $f_j(x_{ij}, y_i)$ and $h_k(Y_c)$. $f_j(x_{ij}, y_i)$ is an attribute factor function. It can be defined as either a binary function or a real-valued function. For example, for the implicit network feature, we simply define it as a binary feature, that is if user v_i^s forwarded (retweeted) v_i^u's tweet before time t and user v_i^u follows user v_i^s back, then a feature $f_j(x_{ij} = 1, y_i = 1)$ is defined and its value is 1; otherwise 0. (Such a feature definition is often used in graphical models such as conditional random fields [Lafferty et al. 2001]. For the triad factor function $h(Y_c)$, we define four features, two balanced and two unbalanced factor functions, as depicted in Figure 6.) The triad function is defined as a binary function, that is, if a triad satisfies the structural balance property, then the value of a corresponding triad factor function is 1, otherwise 0. More details of the factor function definition are given in the Appendix.

4.2. Model Learning and Prediction

We now address the problem of estimating the free parameters and inferring users' follow-back behaviors. Learning the TriFG model is to estimate a parameter configuration $\theta = (\{\alpha\}, \{\mu\})$ to maximize the log-likelihood objective function $\mathcal{O}(\theta) = \log P_\theta(Y|\mathbf{X}, G)$, that is,

$$\theta^\star = \arg\max \mathcal{O}(\theta). \tag{6}$$

To solve the objective function, we adopt a gradient descent method (or a Newton-Raphson method). We use μ as the example to explain how we learn the parameters. Specifically, we first write the gradient of each μ_k with regard to the objective function (Eq. (5))

$$\frac{\mathcal{O}(\theta)}{\mu_k} = \mathbb{E}[h_k(Y_c)] - \mathbb{E}_{P_{\mu_k}(Y_c|\mathbf{X},G)}[h_k(Y_c)], \tag{7}$$

where $\mathbb{E}[h_k(Y_c)]$ is the expectation of factor function $h_k(Y_c)$ given the data distribution (essentially it can be considered as the average value of the factor function $h_k(Y_c)$ over all triads in the training data); and $\mathbb{E}_{P_{\mu_k}(Y_c|\mathbf{X},G)}[h_k(Y_c)]$ is the expectation of factor function $h_k(Y_c)$ under the distribution $P_{\mu_k}(Y_c|\mathbf{X}, G)$ given by the estimated model. A similar gradient can be derived for parameter α_j.

One challenge here is that the graphical structure in the TriFG model can be arbitrary and may contain cycles, which makes it intractable to directly calculate the marginal distribution $P_{\mu_k}(Y_c|\mathbf{X}, G)$. A number of approximate algorithms can be considered, such as Loopy Belief Propagation (LBP) [Murphy et al. 1999] and mean-field [Xing et al. 2003]. We chose Loopy Belief Propagation due to its ease of implementation and effectiveness. Specifically, we approximate the marginal distribution $P_{\mu_k}(Y_c|\mathbf{X}, G)$ using LBP. With the marginal probabilities, the gradient can be obtained by summing over all triads. It is worth noting that we need to perform the LBP process twice in each iteration, one time for estimating the marginal distribution of unknown variables $y_i =?$ and the other time for estimating the marginal distribution over all triads. In this way, the algorithm essentially performs a semisupervised learning over the complete network. This idea was first proposed in Tang et al. [2011] for learning to categorize social relationships. In this work, we extend it for learning the TriFG model. Finally with the obtained gradient, we update each parameter with a learning rate η. The learning algorithm is summarized in Algorithm 1.

ALGORITHM 1: Learning algorithm for the TriFG model.

Input: network G^t, learning rate η
Output: estimated parameters θ

Initialize $\theta \leftarrow 0$;
repeat
  Perform LBP to calculate marginal distribution of unknown variables $P(y_i|x_i, G)$;
  Perform LBP to calculate the marginal distribution of triad c, i.e., $P(y_c|\mathbf{X}_c, G)$;
  Calculate the gradient of μ_k according to Eq. 7 (for α_j with a similar formula):

$$\frac{\mathcal{O}(\theta)}{\mu_k} = \mathbb{E}[h_k(Y_c)] - \mathbb{E}_{P_{\mu_k}(Y_c|\mathbf{X},G)}[h_k(Y_c)]$$

  Update parameter θ with the learning rate η:

$$\theta_{\text{new}} = \theta_{\text{old}} + \eta \cdot \frac{\mathcal{O}(\theta)}{\theta}$$

until *Convergence*;

Predicting follow-back. With the estimated parameters θ, we can predict the label of unknown variables $\{y_i =?\}$ by finding a label configuration which maximizes the objective function, that is, $Y^* = \text{argmax}\mathcal{O}(Y|\mathbf{X}, G, \theta)$. It is still intractable to obtain the exact solution. Again, we utilize the loopy belief propagation to approximate the solution, that is, to calculate the marginal distribution of each relationship with unknown variable $P(y_i|\mathbf{x}_i, G)$ and finally assign each relationship with a label of the maximal probability.

Predicting triadic closure. The proposed TriFG model is flexible and can be easily extended to predict triadic closure. The main difference from reciprocity prediction is the feature definition. Section 7 gives the feature definition for triadic closure prediction. Based on the defined features, we can learn a factor graph model using the proposed TriFG model. In the prediction, we first select candidate triads, that is, those triads where A follows back B at time t, then A, B, and B's followee C form a candidate triad. Then analogous to the follow-back prediction, with the learned parameters, we can predict the label of unknown variables $\{y_i =?\}$ by finding a label configuration which maximizes the objective function. We again utilize the loopy belief propagation to calculate the marginal distribution of each relationship with unknown variable $P(y_i|\mathbf{x}_i, G)$ and finally assign the label (1-follow or 0-not follow) with the maximal probability to those candidate triads.

5. EXPERIMENTS

In this section, we first describe our experimental setup. We then present the performance results for different approaches in different settings. Next, we present several analyses and discussions. Finally, we use a case study further to demonstrate the advantage of the proposed model.

5.1. Experimental Setup

Prediction setting. We use the dataset described in Section 3 in our experiments. To quantitatively evaluate the effectiveness of the proposed model and compare with other alternative methods, we carefully select a dynamic network which consists of a completely historic log of link formation information among users, that is, each user is associated with a complete list of followers and followees at each timestamp. The network is comprised of 112,044 users, 468,238 following links among them, and 2,409,768 tweets. On average, there are 40,943 new-follow links and 3,337 new-follow-back links per day. We divide the subnetwork into 13 timestamps by viewing every four days as a timestamp.

Our general task is to predict whether a user will follow another user back (or follow another user's followee so as to form a closure triad) at the next timestamp when she follows back the user. By a more careful study, however, we find that it is very challenging if we restrict the prediction just for the next timestamp. Figure 11 shows the distribution of time span in which a user performs the follow-back action, which indicates that 60% of follow-backs are performed in the next timestamp though 37% of the follow-backs would be still performed in the following three timestamps. For the triadic closure formation, it is the similar case, that is, 59% of formed triadic closure happens in the next timestamp and 37% in the following three timestamps. A further data analysis shows that active users often either perform an immediate follow-back (at the next timestamp) or reject to follow back; while some other (inactive) users may not frequently login into Twitter, thus the time span of follow-backs varies a lot. According to this observation, in our first experiment, we use a network of the first 8 timestamps for training and predicate follow-back actions in the following 4 (9th–12th) timestamps (Test Case 1). Then we incrementally add the network of the 9th timestamp into the

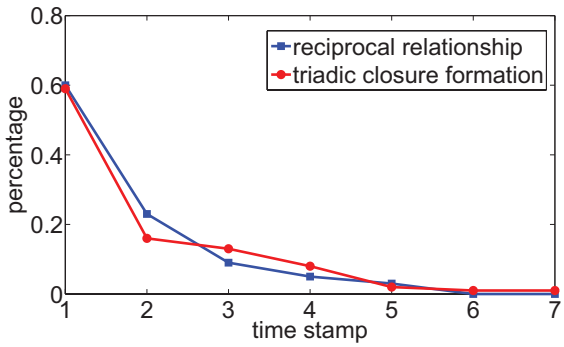


Fig. 11. Follow-back probability for different timestamps.

training data and again use the following 4 (10th–13th) timestamps for prediction (Test Case 2). We respectively report the prediction performance of different approaches for the two test cases.

Comparison methods. We compare the proposed TriFG model with the following methods:

SVM. It uses the same attributes associated with each edge as features to train a classification model and then employs the classification model to predict edges' label in the test data. For SVM, we employ SVM-light.

LRC. It uses the same attributes associated with each edge as features to train a logistic regression classification model [Leskovec et al. 2010] and then predicts edges' labels in the test data.

CRF-balance. It trains a conditional random field [Lafferty et al. 2001] model with attributes associated with each edge. The difference of this method from our model is that it does not consider structural balance factors.

CRF. It trains a conditional random field model with all factors (including attributes and structural balance factors) and predicts edges' labels in the test data.

TriFG. The proposed model trains a factor graph model with unlabeled data and all factors we defined in Section 4.

Weak TriFG (wTriFG). The difference of wTriFG from TriFG is that we do not consider status homophily and structural balance here. We use this method to evaluate how social theories can help this task.

In the six methods, SVM and CRF-balance only consider attribute factors; wTriFG further considers unlabeled data. CRF considers all factors we defined, but does not consider unlabeled data. Our proposed TriFG model considers all factors as well as the unlabeled data.

Evaluation measures. We evaluate the performance of different approaches in terms of precision (Prec.), recall (Rec.), and F1-measure (F1).

All algorithms are implemented in C++, and all experiments are performed on a PC running Windows 7 with Intel(R) Core(TM) 2 CPU 6600 (2.4 GHz) and 4GB memory. The proposed algorithm has the tractable running times on networks of 112,044 size/order of magnitude. Our reciprocity predictions required 2 to 5 minutes, and triadic closure predictions required 2 to 18 minutes.

Table I. Follow-Back Prediction Performance of Different Methods in the Two Test Cases

Data	Algorithm	Prec.	Rec.	F1
Test Case 1	SVM	0.6908	0.6129	0.6495
	LRC	0.6957	0.2581	0.3765
	CRF-balance	0.9968	0.5161	0.6801
	CRF	**1.0000**	0.6290	0.7723
	wTriFG	0.9691	0.5483	0.7004
	TriFG	**1.0000**	**0.8548**	**0.9217**
Test Case 2	SVM	0.7323	0.6212	0.6722
	LRC	0.8333	0.3030	0.4444
	CRF-balance	0.9444	0.5151	0.6667
	CRF	**1.0000**	0.6333	0.7755
	wTriFG	0.9697	0.5697	0.7177
	TriFG	**1.0000**	**0.8788**	**0.9355**

Test Case 1: predicting follow-back actions in the 9th–12th timestamps; and Test Case 2 for the 10th–13th timestamps.

5.2. Reciprocity Prediction Performance

We now describe the performance results for the different methods we considered. Table I shows the results in the two test cases (prediction performance for the 9th–12th timestamps and that for the 10th–13th timestamps).

It can be clearly seen that our proposed TriFG model significantly outperforms the four comparison methods. In terms of F1-measure, TriFG achieves a +27% improvement compared with the (SVM). Comparing with the other three graph-based methods, TriFG also results in an improvement of 22–25%. The advantage of TriFG mainly comes from the improvement on recall. One important reason here is that TriFG can detect some difficult cases by leveraging the structural balance correlation and homophily correlation. For example, without considering the two kinds of social correlations, the performance of wTriFG decreases to 70–72% in terms of F1-measure in the two test cases. Another advantage of TriFG is that it makes use of the unlabeled data. Essentially, it further considers some latent correlations in the dataset, which cannot be leveraged with only the labeled training data.

Now, we perform several analyses to examine the following aspects of the TriFG model: (1) contribution of different factors in the TriFG model; (2) convergence property of the learning algorithm; (3) effect of different settings for the time span; and (4) effect of different algorithms for elite user finding.

Factor contribution analysis. In TriFG, we consider five different factor functions: Geographic distance (G), link homophily (L), status homophily (S), implicit network correlation (I), and structural balance correlation (B). Here we examine the contribution of the different factors. We first rank the individual factors by their predictive power[6], and then remove them one by one in reversing order of their prediction power. In particular, we first remove structural balance correlation denoted as TriFG-B, followed by further removing the implicit network correlation denoted as TriFG-BI, status homophily denoted as TriFG-BIS, and finally removing link homophily denoted as TriFG-BISL. We train and evaluate the prediction performance of the different versions of TriFG. Figure 12 shows the average F1-measure score of the different versions of the TriFG model. We can observe a clear drop on the performance when ignoring

[6]We did this by respectively removing each particular factor from our model and evaluated the decrease of the prediction performance by the TriFG model. A larger decrease means a higher predictive power.

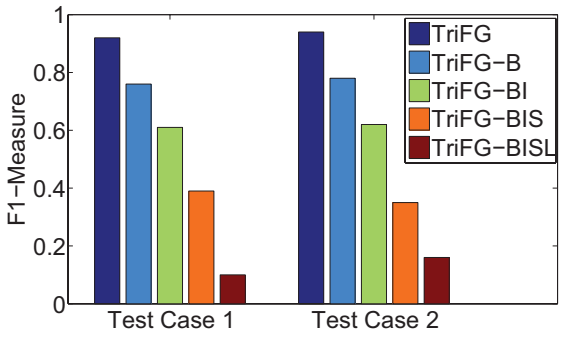


Fig. 12. Factor contribution analysis. TriFG-B stands for ignoring structural balance correlation. TriFG-BI stands for ignoring both structural balance correlation and implicit network correlation. TriFG-BIS stands for further ignoring status homophily and TriFG-BISL stands for further ignoring link homophily.

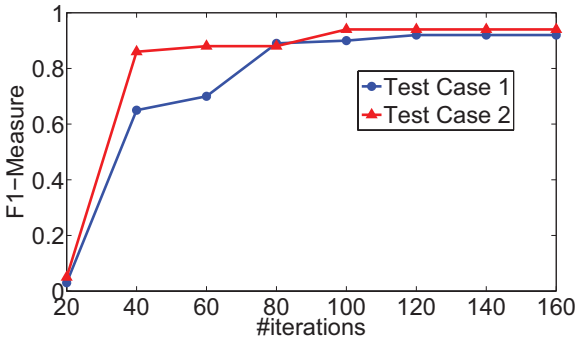


Fig. 13. Convergence analysis of the learning algorithm.

each of the factors. This indicates that our method works well by combining the different factor functions and each factor in our method contributes improvement in the performance.

Convergence property. We conduct an experiment to see the effect of the number of the loopy belief propagation iterations. Figure 13 illustrates the convergence analysis results of the learning algorithm. We see on both test cases, the learning algorithm can converge in less than 200 iterations. After 120 learning iterations, the prediction performance of TriFG on both test cases becomes stable. This suggests that the learning algorithm is very efficient and has a good convergence property.

Effect of time span. Figure 11 already shows the distribution of follow-backs in different time stamps. Now, we quantitatively examine how different settings for the time span will affect the prediction performance. Figure 14 lists the average prediction performance of TriFG in the two test cases with different settings of the time span. It shows that when setting the time span as two or less timestamps, the prediction performance of TriFG drops sharply; while when setting it as three timestamps, the performance is acceptable. The results are consistent with the statistics in Figure 11: more than 90% of follow-back actions are performed in the first three timestamps, and only about 80% of the follow-back actions are in the first two timestamps.

Effect of different algorithms for elite user finding. The status homophily factor depends on results of elite user finding. We use three different algorithms, that is,

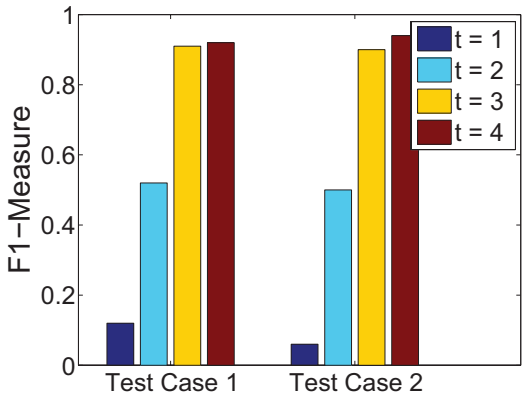


Fig. 14. Follow-back prediction for different timestamps.

Table II. Follow-Back Prediction Performance of TriFG with Three Different Algorithms (#degree, PageRank and (α, β)) for Finding Elite Users from Ordinary Users

Data	Algorithm	Prec.	Rec.	F1
Test Case 1	(α, β)	**1.0000**	**0.8548**	**0.9217**
	#degree	**1.0000**	0.7903	0.8829
	pagerank	**1.0000**	0.7581	0.8624
Test Case 2	(α, β)	**1.0000**	**0.8788**	**0.9355**
	#degree	**1.0000**	0.8363	0.9109
	pagerank	**1.0000**	0.8181	0.9000

PageRank, #degree, and (α, β) algorithm, to find elite users. Now we examine how the different algorithms would affect the prediction performance. Table II shows the prediction performance of TriFG with different elite user finding algorithms in the two test cases. Interestingly, though TriFG with the (α, β) algorithm achieves the best performance, the difference of performance among the three algorithms, especially in the second test case, is not that pronounced (with a difference of 1%–4% in terms of F1-measure score). This confirms the effectiveness and generalization of incorporating the status homophily factor into our TriFG model.

5.3. Triadic Closure Prediction Performance

We now turn to discuss the performance of triadic closure prediction by the different methods we considered. Table III shows the results in the two test cases (prediction performance for the 9th–12th timestamps and that for the 10th–13th timestamps). In the task of triadic closure prediction, the labeled data is very unbalanced (a large portion of instances are negative instances, i.e., A follows B back, but does not follow B's followees), thus it is more difficult than the reciprocity prediction task. Even the best performance of the baseline methods on the first test case is only 10% by F1 and 22% on the second test case. Our proposed TriFG significantly improves the performance (+18.6% in terms of F1-score). The situation is similar on the second test case. Comparing with the other three graph-based methods, TriFG also results in an improvement of 23–34%. The advantage of TriFG mainly comes from the improvement on precision.

Factor contribution analysis. For the triadic closure prediction, we mainly consider three factor functions: structural balance correlation (B), status homophily (S), and link homophily (L). Here we examine the contribution of the different factors defined

Table III. Triadic Closure Prediction Performance of Different Methods in the Two Test Cases

Data	Algorithm	Prec.	Rec.	F1
Test Case 1	SVM	0.0870	0.1429	0.1081
	LRC	0.0536	0.1304	0.0759
	CRF-balance	0.0208	0.0436	0.0282
	CRF	0.1111	0.0870	0.0976
	wTriFG	0.3333	0.0373	0.0671
	TriFG	**0.4545**	**0.2174**	**0.2941**
Test Case 2	SVM	0.2000	0.2222	0.2105
	LRC	0.1071	0.1667	0.1304
	CRF-balance	0.0909	0.0556	0.0690
	CRF	0.2222	0.2222	0.2222
	wTriFG	0.5000	0.0556	0.1000
	TriFG	**0.8571**	**0.3333**	**0.4800**

Test Case 1: predicting triadic closure in the 9th–12th timestamps; and Test Case 2 for the 10th–13th timestamps.

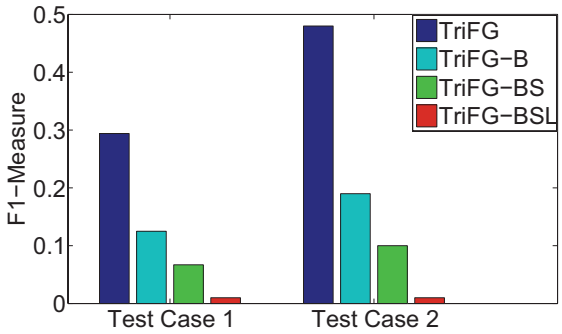
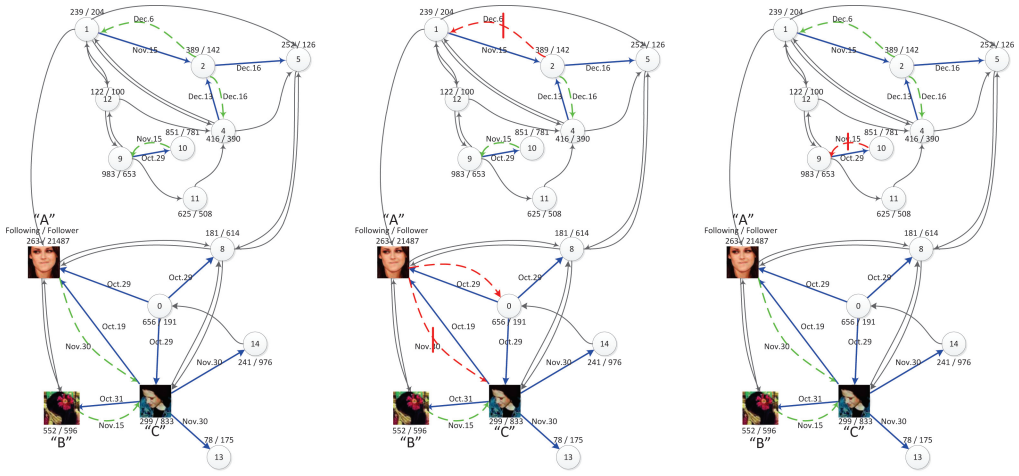


Fig. 15. Factor contribution analysis. TriFG-B stands for ignoring structural balance correlation. TriFG-BS stands for ignoring both structural balance correlation and status homophily, and TriFG-BSL stands for further ignoring link homophily.

in our model. Again, we first rank the individual factors by their predictive power, and then remove them one by one in reversing order of their prediction power. In particular, we remove structural balance correlation denoted as TriFG-B, followed by further removing the status homophily denoted as TriFG-BS, and finally removing link homophily denoted as TriFG-BSL. We train and evaluate the prediction performance of the different versions of TriFG. Figure 15 shows the average F1-measure score of the different versions of the TriFG model. We can observe a clear drop on the performance when ignoring each of the factors.

5.4. Qualitative Case Study

Now we present a case study to demonstrate the effectiveness of the proposed model. Figure 16 shows an example generated from our experiments. It represents a portion of the Twitter network from the 10th–13th timestamps. Black arrows indicate following links created 4 timestamps (we use 4 timestamps as the time span for prediction) before. Blue arrows indicate new-following link in the past 4 timestamps. Dash arrows indicate follow-back links in our dataset (a), predicted by SVM (b), and predicted by our model TriFG (c), with green color denoting a correct one and red color denoting



(a) Ground Truth (b) SVM (c) Our approach (TriFG)

Fig. 16. Case study. Portion of the Twitter network during the 10th–13th timestamps. The two numbers associated with each user are respectively the number of followees and that of followers. Black arrows indicate following links created 4 timestamps (we use 4 timestamps as the time span for prediction) before. Blue arrows indicate new-following link in the past 4 timestamps. Dash arrows indicate follow-back links in our dataset (a), predicted by SVM (b), and predicted by our model TriFG (c), with green color denoting a correct one and red color denoting a mistake one. Red colored $\nrightarrow$ indicates there should be a follow-back link, which the approach did not predict.

a mistake one. Red colored $\nrightarrow$ indicates there should be a follow-back link, but the approach does not detectit.

We look at specific examples to study why the proposed model can outperform the comparison methods. "A", "B", and "C" are three elite users identified using the (α, β) algorithm [He et al. 2011]. SVM correctly predicts that there is a follow-back link from "C" to "B", but misses predicting the follow-back link from "C" to "A". Our model TriFG correctly predicts both the follow-back links. This is because TriFG leverages the structural balance factor. The resultant structure among the three users by SVM is unbalanced. TriFG leverages the structural balance factor and tends to result in a balanced structure.

It is also worth looking at the situation of users 9 and 10. TriFG made a mistake here: it does not predict the follow-back link, while the link is correctly predicted by SVM. Users 9 and 10 have a similar social status (similar indegree) and also they are from the same time zone, thus SVM successfully predicts the follow-back link. However, as the resulting structure is unbalanced, TriFG make a compromise and finally results in a mistaken prediction.

5.5. Prototype System

We have developed and deployed a Web application for reciprocal prediction based on the proposed TriFG model[7]. Figure 17 shows a screenshot of the reciprocity prediction system. The system trains a TriFG model offline using all the follow and follow-back relationships in our dataset. When a user wants to know how likely another user will follow him back, he first inputs his Twitter ID and the other user's ID. Then the system analyzes his social circle and the other user's social circle, and extracts features defined in our approach. Next, it makes the prediction based on the trained TriFG model (refer

[7]http://reciprocal.aminer.org.

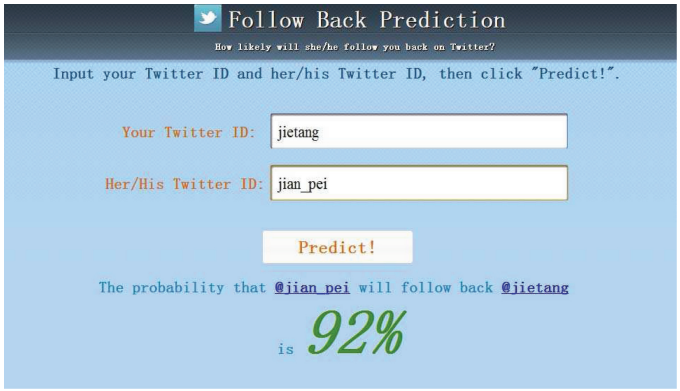


Fig. 17. A screenshot of the reciprocity prediction system.

to Section 4.2). As the example in Figure 17 shows, the user "jian_pei" has a probability of 92% to follow back user "jietang".

6. RELATED WORK

In this section, we review related work on link prediction and Twitter study in social networks.

Social tie analysis. There are several works on inferring the meanings of social relationships. Diehl et al. [2007] try to identify the manager-subordinate relationships by learning a ranking function. Wang et al. [2010] propose an unsupervised probabilistic model for mining the advisor-advisee relationships from the publication network. Crandall et al. [2010] investigate the problem of inferring friendship between people from co-occurrence in time and space. Tang et al. [2011] propose a learning framework based on partially labeled factor graphs for inferring the types of social relationships in different networks. Zhuang et al. [2012] further propose using active learning to enhance the inferring performance. Eagle et al. [2009] present several patterns discovered in mobile phone data, and try to use these patterns to infer the friendship network. Tang et al. [2012a] study the problem of inferring social ties across heterogeneous networks. However, these algorithms cannot be directly applied to infer the follow-back relationships and they do not consider the problem of triadic closure prediction.

Another type of related work is social behavior analysis. Tang et al. [2009] study the difference of the social influence on different topics and propose Topical Affinity Propagation (TAP) to model the topic-level social influence in social networks and develop a parallel model learning algorithm based on the map-reduce programming model. Tan et al. [2010] investigate how social actions evolve in a dynamic social network and propose a time-varying factor graph model for modeling and predicting users' social behaviors. The proposed methods in these works can be utilized in the problem defined in this work, but the problem is fundamentally different.

In our previous work [Hopcroft et al. 2011], we study the problem of reciprocal relationship prediction. In this work, we extend this work from the following aspects. First, we further investigate how closure triads are formed from pairwise relationships, and how the formation is correlated with factors such as link homophily and social status. Second, we extend the factor graph model to infer the triadic closure formation. Last, we evaluate the proposed model on the dataset Twitter.

Link prediction. Our work is related with link prediction, which is one of the core tasks in social networks. Existing work on link prediction can be broadly grouped into

two categories based on the learning methods employed: unsupervised link prediction and supervised link prediction. Unsupervised link predictions usually assign scores to potential links based on the intuition that the more similar the pair of users are, the more likely they are linked. Various similarity measures of users are considered, such as *preferential attachment* [Newman 2001], and the Katz measure [Katz 1953]. Lichtenwalter et al. [2010] present a flow-based method for link prediction. A survey of unsupervised link prediction can be found in Liben-Nowell and Kleinberg [2007].

There are also a number of works which employ supervised approaches to predict links in social networks, such as Wang et al. [2007], Lichtenwalter et al. [2010], Backstrom and Leskovec [2011], and Leskovec et al. [2010]. Backstrom and Leskovec [2011] propose a supervised random walk algorithm to estimate the strength of social links. Leskovec et al. [2010] employ a logistic regression model to predict positive and negative links in online social networks. The main differences between existing work on link prediction and our work are about two aspects. First, existing work handles undirected social networks, while we address the directed nature of the Twitter network and predict a directed link between a pair of users given an existing link in another direction. Second, most existing models for link prediction are static. In contrast, our model is dynamic and learned from the evolution of the Twitter network. Moreover, we combine social theories (such as homophily and structural balance theory) into a semisupervised learning model.

Twitter study. There is little doubt that Twitter has intrigued worldwide netizens, and research communities alike. Existing Twitter study is mainly centered around the following three aspects: (1) *the Twitter network*. Java et al. [2007] study the topological and geographical properties of the Twitter network. Their findings verify the *homophily* phenomenon that users with similar intentions connect with each other. Kwak et al. [2010] conduct a similar study on the entire Twittersphere and they observe some notable properties of Twitter, such as a nonpower-law follower distribution, a short effective diameter, and low reciprocity, marking a deviation from known characteristics of human social networks. (2) *the Twitter users*. Work of this category mainly focus on identifying influential users in Twitter [Weng et al. 2010; Cha et al. 2010; Kwak et al. 2010] or examining and predicting tweeting behaviors of users [Huberman et al. 2009; Tan et al. 2010]. (3) *the Tweets*. Sakaki et al. [2010] propose to utilize the real-time nature of Twitter to detect a target event, while Mathioudakis and Koudas [2010] present a system, TwitterMonitor, to detect emerging topics from the Twitter content.

Triadic closure formation. There are a few works on triadic closure analysis. Romero and Kleinberg [2010] study the problem of the triadic closure process and develop a methodology based on preferential attachment, for studying the directed triadic closure process. Backstrom et al. [2008] propose a partitioning on the data that selects for active communities of engaged individuals.

7. CONCLUSION

In this article, we study the novel problem of predicting reciprocity and triadic closure in social networks. We formally define the two subproblems and propose a Triad Factor Graph (TriFG) model, which incorporates social theories into a semisupervised learning model. We evaluate the proposed model on a large Twitter network. We show that with the proposed factor graph model it is possible to accurately infer 90% of reciprocal relationships in a dynamic network. We also demonstrate that the proposed model significantly improves the performance (+22%–27% by F1-measure) for triadic closure prediction comparing with several alternative methods. Our study also reveals several interesting phenomena.

The general problem of understanding the structure of networks represents a novel research direction in social network analysis. There are many potential future directions of this work. First, what are the fundamental differences of the structure between different networks? Can we classify the networks into different categories? Second, some other social theories can be further explored and validated for reciprocal relationship prediction. Looking farther ahead, it is also interesting to develop a real friend suggestion system based on the proposed method. We can further study theoretical methodologies for improving the predictive performance by incorporating user interactions. Finally, building a theory of why and how users create relationships with each other in different kinds of networks is an intriguing direction for further research.

APPENDIX: FACTOR FUNCTION DEFINITION

A.1. Feature Definition for Reciprocity Prediction

This section depicts how we define the factor functions in our experiments of reciprocal relationship prediction. In total, we define 26 features of five categories: geographic distance, link homophily, status homophily, structural balance, and implicit network correlation.

Geographic distance. We use Google Map API to get the exact locations (longitude and latitude) of some users. Based on the two values, we define the following three features: the absolute distance and the time zone difference between two users, and whether or not the two users are from the same country.

Link homophily. First, we treat each link as an undirected link, and define the following four features: the number of common neighbors, percentage of common neighbors of the two users (respectively), and the average percentage.

Then we consider directed links and define another three features: the number of common reciprocal links, number of common followers, and number of common followees.

Status homophily. We also test whether two users have similar social status, and define the following four features: whether or not the two users are both elite users (two features), an ordinary and an elite, and both ordinary users.

Implicit network correlation. We consider the interaction between user A and user B, and define the following four features that respectively represent the number of retweets (replies) from A to B and from B to A.

Structural balance. Based on the structural balance theory, as in Figure 6, we define eight features capturing all situations of structural balance theory for each triad.

A.2. Feature Definition for Triadic Closure Prediction

This section depicts how we define the factor functions in our experiments of triadic closure prediction. In total, we define 46 features of four categories: geographic distance, link homophily, status homophily, and structural balance.

Geographic distance. We use Google Map API to get the exact locations (longitude and latitude) of some users. Based on the two values, we define the following 9 features (three features for each pair among the three users): the absolute distance and the time zone difference between two users, and whether or not the two users are from the same country.

Link homophily. First, we treat each link as an undirected link, and define the following 12 features (four features for each pair among the three users): the number

of common neighbors, percentage of common neighbors of the two users (respectively), and the average percentage.

Then we consider directed links and define another 9 features (three features for each pair among the three users): the number of common reciprocal links, number of common followers, and number of common followees.

Status homophily. We also test whether each pair of users have similar social status, and define the following 12 features (four features for each pair among the three users): whether or not the two users are both elite users, an ordinary and an elite (two features), and both ordinary users.

Structural balance. Based on structural balance theory, as in Figure 6, we define four features capturing all situations of structural balance theory for each triad.

REFERENCES

Backstrom, L., Kumar, R., Marlow, C., Novak, J., and Tomkins, A. 2008. Preferential behavior in online groups. In *Proceedings of the International Conference on Web Search and Data Mining* (*WSDM'08).* 117–128.

Backstrom, L. and Leskovec, J. 2011. Supervised random walks: Predicting and recommending links in social networks. In *Proceedings of the International Conference on Web Search and Data Mining* (*WSDM'11).* 635–644.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P. K. 2010. Measuring user influence in twitter: The million follower fallacy. In *Proceedings of the AAAI International Conference on Weblogs and Social Media* (*ICWSM'10).*

Crandall, D. J., Backstrom, L., Cosley, D., Suri, S., Huttenlocher, D., and Kleinberg, J. 2010. Inferring social ties from geographic coincidences. *Proc. Nat. Acad. Sci. United States 107*, 22436–22441.

Diehl, C. P., Namata, G., and Getoor, L. 2007. Relationship identification for social network discovery. In *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI'07).* 546–552.

Eagle, N., Pentland, A. S., and Lazer, D. 2009. Inferring social network structure using mobile phone data. *Proc. Nat. Acad. Sci. United States 106,* 36.

Easley, D. and Kleinberg, J. 2010. *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press.

Hammersley, J. M. and Clifford, P. 1971. Markov field on finite graphs and lattices.

He, J., Hopcroft, J., Liang, H., Suwajanakorn, S., and Wang, L. 2011. Detecting the structure of social networks using (α, β)-communities. In *Proceedings of the 8th International Conference on Algorithms and Models for the Web Graph* (*WAW'11).*

Hopcroft, J., Lou, T., and Tang, J. 2011. Who will follow you back? Reciprocal relationship prediction. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (*CIKM'11).* 1137–1146.

Horton, D. and Wohl, R. R. 1956. Mass communication and para-social interaction: Observations on intimacy at a distance. *Psychiatry 3,* 1, 215–229.

Huberman, B., Romero, D. M., and Wu, F. 2009. Social networks that matter: Twitter under microscope. *First Monday 14,* 118–138.

Java, A., Song, X., Finin, T., and Tseng, B. L. 2007. Why we twitter: An analysis of a microblogging community. In *Proceedings of the 9th International Workshop on Knowledge Discovery on the Web and the 1st International Workshop on Social Networks Analysis (WebKDD/SNA-KDD'07).* 118–138.

Katz, L. 1953. A new status index derived from sociometric analysis. *Psychometrika 18,* 1, 39–43.

Kwak, H., Lee, C., Park, H., and Moon, S. B. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web* (*WWW'10).* 591–600.

Lafferty, J. D., Mccallum, A., and Pereira, F. C. N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML'01).* 282–289.

Lazarsfeld, P. F. and Merton, R. K. 1954. Friendship as a social process: A substantive and methodological analysis. In *Freedom and Control in Modern Society,* M. Berger, T. Abel, and C. H. Page, Eds., Van Nostrand, New York, 8–66.

Leskovec, J., Huttenlocher, D., and Kleinberg, J. 2010. Predicting positive and negative links in online social networks. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10).* 641–650.

LIBEN-NOWELL, D. AND KLEINBERG, J. M. 2007. The link-prediction problem for social networks. *JA-SIST 58,* 7, 1019–1031.

LICHTENWALTER, R., LUSSIER, J. T., AND CHAWLA, N. V. 2010. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD'10).* 243–252.

MATHIOUDAKIS, M. AND KOUDAS, N. 2010. Twittermonitor: Trend detection over the twitter stream. In *Proceedings of the International Conference on Management of Data (SIGMOD'10).* 1155–1158.

MURPHY, K. P., WEISS, Y., AND JORDAN, M. I. 1999. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI'99).* 467–475.

NEWMAN, M. E. J. 2001. Clustering and preferential attachment in growing networks. *Phys. Rev. E64,* 2, 25–102.

PAGE, L., BRIN, S., MOTWANI, R., AND WINOGRAD, T. 1999. The pagerank citation ranking: Bringing order to the web. Tech. rep. SIDL-WP-1999-0120, Stanford University.

ROMERO, D. M. AND KLEINBERG, J. M. 2010. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In *Proceedings of the AAAI International Conference on Weblogs and Social Media* (*ICWSM'10).*

SAKAKI, T., OKAZAKI, M., AND MATSUO, Y. 2010. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web (WWW'10).* 851–860.

TAN, C., LEE, L., TANG, J., JIANG, L., ZHOU, M., AND LI, P. 2011. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD'11).* 1397–1405.

TAN, C., TANG, J., SUN, J., LIN, Q., AND WANG, F. 2010. Social action tracking via noise tolerant timevarying factor graphs. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD'10).* 1049–1058.

TANG, J., LOU, T., AND KLEINBERG, J. 2012a. Inferring social ties across heterogeneous networks. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining* (*WSDM'12).* 743–752.

TANG, J., SUN, J., WANG, C., AND YANG, Z. 2009. Social influence analysis in large-scale networks. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD'09).* 807–816.

TANG, J., WU, S., SUN, J., AND SU, H. 2012b. Cross-domain collaboration recommendation. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD'12).*

TANG, W., ZHUANG, H., AND TANG, J. 2011. Learning to infer social ties in large networks. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD'11).* 381–397.

WANG, C., HAN, J., JIA, Y., TANG, J., ZHANG, D., YU, Y., AND GUO, J. 2010. Mining advisor-advisee relationships from research publication networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD'10).* 203–212.

WANG, C., SATULURI, V., AND PARTHASARATHY, S. 2007. Local probabilistic models for link prediction. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM'07).* 322–331.

WEBER, M. 1991. *The Nature of Social Action in Runciman, W.G. 'Weber: Selections in Translation'.* Cambridge University Press.

WENG, J., LIM, E.-P., JIANG, J., AND HE, Q. 2010. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM'10).* 261–270.

WU, S., HOFMAN, J. M., MASON, W. A., AND WATTS, D. J. 2011. Who says what to whom on twitter. In *Proceedings of the 20th International Conference on World Wide Web* (*WWW'11).* 705–714.

XING, E. P., JORDAN, M. I., AND RUSSELL, S. 2003. A generalized mean field algorithm for variational inference in exponential families. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence* (*UAI'03).* 583–591.

YANG, Z., GUO, J., CAI, K., TANG, J., LI, J., ZHANG, L., AND SU, Z. 2010. Understanding retweeting behaviors in social networks. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10).* 1633–1636.

ZHUANG, H., TANG, J., TANG, W., LOU, T., CHIN, A., AND WANG, X. 2012. Graph-based ranking algorithms for e-mail expertise analysis. *Data Mining Knowl. Discov. 25,* 2, 270–297.

Received March 2012; revised May 2012; accepted October 2012