Contributors

Chapter 5

Analysis and Prediction of Triadic Closure in Online Social Networks

Hong Huang

University of Goettingen, Germany (Computer Science)

Jie Tang

Tsinghua University, China (Computer Science)

Lu Liu

TangoMe Inc., USA (Computer Science)

Jar-Der Luo

Tsinghua University, China (Sociology)

Xiaoming Fu

University of Goettingen, Germany (Computer Science)

CONTENTS

5.1 Introduction

Online social networks (OSNs) are becoming a bridge that con- nects our physical daily life with the online world. For example, as of July 2014, Facebook has 1.3 billion users, which makes Facebook the second biggest "country" in the world. Twitter has 0.65 billion users, who "tweet" 1 billion times every five days. These connections produce a huge volume of data, including not only the content of their communications, but also user behavioral logs. The popularity of the social web and the availability of social data offer us opportunities to study interaction patterns among users, and to understand the generative mechanisms of different networks, which were previously difficult to explore, due to the unavailability of data. A better understanding of user behavior and underlying network patterns could enable an OSN provider to attract and keep more users, and thus increase its profits.

In social networks, group formation – the process by which people come together, seek new friends, and develop communities – is a central research issue in the social sciences. Examples of interesting groups include political movements and professional organizations [1].

A triad is a group of three people. It is one of the simplest human groups. Roughly speaking, there are two types of triads: *closed triads* and *open triads*. In a closed triad, for any two persons in the triad, there is a relationship between them. In an open triad, there are only two relationships, which means that two of the three people are not connected with each other.

One interesting question is how a closed triad develops from an open triad. The problem is referred to as the *triadic closure process*. It is a fundamental mechanism in

the formation and evolution of dynamic networks [5]. Understanding the mechanism of triadic closure can help in predicting the development of ties within a network, in showing the progression of connectivity, and in gaining insight into decision-making behavior in global organizations [8, 19].

The triadic closure process has been studied in many fields. Sociologists first used the triadic closure process to study human friendship choices – i.e., whether people may choose new acquaintances who are the friends of friends [13] – and found that friends of friends tend to become friends themselves [13, 39]. In computer science, empirical studies have shown that triads tend to aggregate, creating interest groups of widely varying size, but of small diameter. For example, these tightly knit groups indicated a common topic for hyperlinks [9] on the World Wide Web. Literature [10, 19, 34, 44] proposed network generative models based on triadic closure principles. Milo et al. [27] [28] defined the recurring significant patterns of interconnections as "network motifs" and emphasized their importance. But these studies focused only on uses of the triadic closure process, without clarifying the underlying principles of triadic closure.

Romero et al. [32] studied the problem of triadic closure process and developed a methodology based on preferential attachment, for studying how directed "feed-forward" triadic closure occurs. Moreover, Lou et al. [26] investigated how a reciprocal link is developed from a parasocial relationship and how the relationships develop into triadic closure in a Twitter dataset. However, these studies only examined some special cases of the triadic closure process. Many challenges are still open and require further methodological developments. First, how do user demographics, network characteristics, and social properties influence the formation of triadic closure? Moreover, how can we design a unified model for predicting the formation of triadic closure? In particular, how can we quantify correlation (similarity) between triads?

In this paper, employing a dataset from a large microblogging network, Weibo[1], as the basis of our study, we examine patterns in triadic closure process in order to better understand factors that trigger the formation of groups among people. Our contributions are multifold:

- We first investigate the triadic closure patterns in the microblogging network from three aspects: user demographics, network characteristics, and social perspectives. We find some interesting phenomena; for example, men are more willing to form triadic closures than women; celebrities are more likely to form triadic closures (with a probability $421\times$ as high) than ordinary users. Furthermore, we find that interactions like retweeting play an important role in the establishment of friendship and in triadic closure formation.

- Based on our observations, we tackle the issue of triadic closure prediction. We present a probabilistic triad factor-graph model (TriadFG) combined with different kernel functions, which quantify the similarity between triads to predict triadic closure. Compared with alternative methods based on SVM

[1] Weibo.com, the most popular microblogging service in China, with more than 560 million users.

and Logistic Regression, the presented model achieves significant improvement (+7.43%, $p \ll 0.01$) in triadic closure prediction.

- We compare the observations obtained from the Weibo dataset with those from the Twitter dataset. Interestingly, although there are common patterns – e.g., "the rich get richer" – underlying the dynamics of the two networks, some distinct patterns (and corresponding users' motivations) exist, potentially reflecting cultural differences of behaviors between Weibo and Twitter users.
- One straightforward application of our findings is friend recommendation. We apply our proposed triadic closure prediction model to the Weibo dataset to evaluate the effectiveness of friend recommendation. The online A/B test demonstrates that our method can achieve an advantage of +10% over the existing recommendation algorithm. Other potential applications include group formation [1, 32], social search, and user behavior modeling.

5.2 Problem Definition

Let $G = (V, E)$ denote a static network, where $V = \{v_1, \cdots, v_{|V|}\}$ is a set of users and $E \subset V \times V$ is a set of relationships connecting those users. Notation $e_{v_i v_j} \in E$ (or simply e_{ij}) denotes there is a relationship between users v_i and v_j. The network evolves over time. Let us denote the network at time t as G^t. To begin with, we give the definitions of *closed triad* and *open triad* in a static social network based on "following" relationships.

Definition 5.1 [Closed Triad] For three users $\Delta = (A, B, C)$, if there is relationship between any two users – i.e., $e_{AB}, e_{BC}, e_{AC} \in E$ – then we say that Δ is a *closed triad*.

Definition 5.2 [Open Triad] For three users $\Delta = \{A, B, C\}$, if we have only two relationships among them – e.g., $e_{AB}, e_{BC} \in E \wedge e_{AC} \notin E$ – then we call the triad Δ an *open triad*.

The triads are formed in a dynamic process. We use function $t(e_{AB}) \rightarrow 1, 2, \cdots$ to define the timestamp at which the relationship e_{AB} was formed between A and B. For simplicity, we use t to denote the timestamp. In this paper, we try to understand how an *open triad* becomes a *closed triad*. The problem exists in both directed and undirected networks. For example, in a co-author network at time t, if B coauthored with A and C respectively, but A and C did not coauthor, we say (A, B, C) is an open triad. If later, A and C also have a coauthorship, we say A, B, and C form a closed triad. In directed networks, the problem becomes more complicated. In some sense, the problem in undirected networks can be considered a special case of the problem

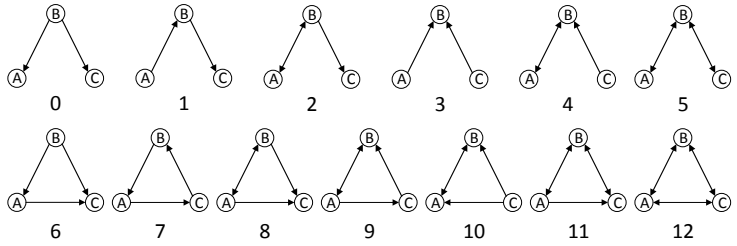


Figure 5.1: Open Triads and Closed Triads. The number below is the index of each triad. Triad 0 **– Triad** 5 **are open triads and Triad** 6 **– Triad** 12 **are closed triads.** *A*, *B* **and** *C* **represent users.**

Table 5.1: How open triad forms triadic closure.

Open$\xrightarrow{A\to C}$Close	Open$\xrightarrow{A\leftarrow C}$Close	Open$\xrightarrow{A\leftrightarrow C}$Close
$0\xrightarrow{A\to C}6$	$0\xrightarrow{A\leftarrow C}6$	$0\xrightarrow{A\leftrightarrow C}10$
$1\xrightarrow{A\to C}6$	$1\xrightarrow{A\leftarrow C}7$	$1\xrightarrow{A\leftrightarrow C}9$
$2\xrightarrow{A\to C}8$	$2\xrightarrow{A\leftarrow C}9$	$2\xrightarrow{A\leftrightarrow C}11$
$3\xrightarrow{A\to C}6$	$3\xrightarrow{A\leftarrow C}6$	$3\xrightarrow{A\leftrightarrow C}8$
$4\xrightarrow{A\to C}9$	$4\xrightarrow{A\leftarrow C}10$	$4\xrightarrow{A\leftrightarrow C}11$
$5\xrightarrow{A\to C}11$	$5\xrightarrow{A\leftarrow C}11$	$5\xrightarrow{A\leftrightarrow C}12$

in directed networks. In this paper we focus on directed networks like Twitter (i.e., follower networks) and Weibo (Chinese Twitter).

Figure 5.1 shows all the possible examples of open and closed triads in a directed network. Table 5.1 shows how these open triads become closed triads when a following action happens between A and C. For each entry in the table, left and right numbers indicate the index of triads in Figure 5.1. The expression above the arrow indicates the action that a new link between A and C is created. For example, $0 \xrightarrow{A\to C} 6$ means if at time t' A follows C, then open triad 0 becomes an isomorphous of closed triad 6.

The situation becomes more complex if we further consider the time when each relationship was formed in the (open/closed) triads. To simplify the following explanation, and without loss of generality, we assume that in an open triad $\Delta = (A,B,C)$, the relationship between B and C was established (at time t_2) after the establishment (at time t_1) of a relationship between A and B – i.e., $t_2 > t_1$. Given this, our goal is to predict whether an open triad will become a closed triad at time $t_3 (t_3 > t_2)$. Formally, we have the following problem definition.

Problem 1 Triadic Closure Prediction. *Given a network $G^t = (V,E)$ at time t and*

historical information regarding all existing relationships. To every candidate open triad we associate a hidden variable y^t. Our goal is to use the historical information to train a function f, so that we can predict whether an open triad in G^t will become a closed triad ($y^t = 1$) at some time $t'(t' > t)$ or not ($y^t = 0$) – i.e.,

$$f : (\{G^\alpha, Y^\alpha\}_{\alpha=1,\cdots,t}) \to Y^{t'},$$

where $Y^t = \{y_i^t\}$ denotes the set of all values of the hidden variables at time t.

We also study how interaction between users can help the formation of triadic closure. We consider retweeting behavior in a microblogging network. In particular, for an open triad (A,B,C), if retweeting happens both between A and B, and B and C, suppose the action between B and C happens after the action between A and B (which is called candidate relationship-interaction open triad (R-I open triad)), will this retweeting help A and C to build a relationship?

Please note that the interaction can be in different forms; for example, the above-mentioned retweeting; "mention" ("@" in Twitter or Weibo); or "reply." To simplify the analysis, we focus on retweeting.

We could extend Problem 1 as follows: Given a network $G^t = (V,E)$ at time t. To every candidate R-I open triad, we associate a hidden variable y_{RI}^t. Our goal is to train a function f, so that we can predict whether an open triad in G^t will become a closed triad at time $t'(t' > t)$ – i.e.,

$$f : (\{G^\alpha, Y_{RI}^\alpha\}_{\alpha=1,\cdots,t}) \to Y_{RI}^{t'},$$

where Y_{RI}^t denotes all values of the hidden variables at time t.

5.3 Data and Observation

5.3.1 Data Collection

One objective of the study is to reveal the fundamental factors that influence triadic closure formation in social networks. We use Weibo data as the basis for our study. Triadic closure process is the formation of a directed triad (also referred to as directed closure process [26,32]). To obtain the dynamic information, we crawl a network with dynamic updates from Weibo. The dataset was crawled in the following ways. To begin with, 100 random users were selected; then their followees and followees' followees were collected as seed users. The crawling process produced in total 1,776,950 users and 308,489,739 following links among them, with an average of 200 out-degree per user, 317,555 new links and 745,587 newly formed closed triads per day. We also crawled the profiles of all users, which contains name, gender, location, verified status, and posted microblogs. Finally, the resultant dynamic networks span a period from September 29th, 2012 to October 29th, 2012. Table 5.2 gives statistics of the dataset.

We construct a network based on the following relationships, which is different

Table 5.2: Data statistics of the Weibo dataset.

Item	Number
#Users	1,776,950
#Following-relationships	308,489,739
#Original-microblogs	300,000
#Retweets	23,755,810
#New links per day(average)	317,555
#New open triads per day(average)	6,203,842,388
#New closed triads per day(average)	745,587

from a co-author network or friendship network. The former is a directed network, while the latter is an undirected network. The main difference between the two is the directed nature of a Weibo relationship, which is like a Twitter relationship. In a co-author network or a message network (MSN), a link represents a mutual agreement by users, while on Weibo a user is not obligated to reciprocate followers by following them. Thus a path from one user to another may follow different hops, or not exist in the reverse direction [17].

5.3.2 Observations

We view the network at the first day (September 28th, 2012, denoted as T_0) as the initial network, and then every four days[2] as a timestamp (denoted as $T_1, T_2, ..., T_7$). The number of newly formed links per timestamp period is shown in Figure 5.2(a), and the number of newly formed open triads per timestamp period is shown in Figure 5.2(b). In Figure 5.2(c), we have the cumulative distribution function of newly formed triadic closures per day, from which we can see that within 8 days, about 60% triadic closures are formed. In order to obtain fair and balanced observations among the limited samples, we only consider the triadic closures generated in 8 days[3] after the open triad formed. Figure 5.2(d) shows the triadic closure probability in different timestamp periods, from which we can see that time slightly affects the closure probability of T_1, T_2, T_3 and T_5, (i.e., $P_{T_1} \approx P_{T_2} \approx P_{T_3} \approx P_{T_5}$).

Exceptions occurred in timestamp period T_4 (open triads formed from Oct. 11st to Oct. 14th and triadic closure formed from Oct. 12nd to Oct. 20th) and T_6 (open triads formed from Oct. 22nd to Oct. 25th and triadic closures formed from Oct. 23rd to

[2] We followed the work in [26], where they used four days as a timestamp period to study triadic closure patterns in Twitter. In addition, we also investigated other timestamps in Section **??** to see the effects of timestamps.

[3] As shown in Figure 5.2(c), about 60% open triads closed in eight days, and 80% open triads closed in 13 days. Since we only have one month's worth of observations, eight days seems to be a better choice than 13 days: first, eight days corresponds to two timestamp periods, which is easy for calculating; second, we can get more effective observations with eight days if we choose all samples with the same observed time period. For example, if we select 12 days, triads in the last two timestamp periods can only be observed in two timestamp periods, so their observations are not complete. Thus, eight days yields more observations than 12 days.

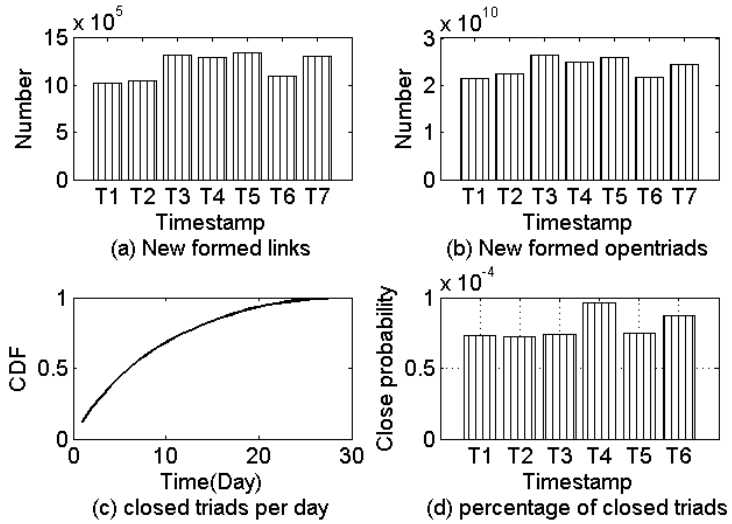


Figure 5.2: Overall observation. (a) Y-axis: the number of new formed links in different timestamp periods. (b) Y-axis: the number of new formed open triads in different timestamp periods. (c) Y-axis: Cumulative distribution function of new formed triadic closures per day. (d) Y-axis: probability that open triads form triadic closures.

Oct. 31st). Coincidentally, on October 11st, the news that Mo Yan (a Chinese writer) won a Nobel prize in literature 2012 began to spread over Weibo. In the following days, an increasing number of people focused on this topic because Mo Yan was the first Chinese citizen to win the Nobel prize in its 111-year history. Maybe it is partly the reason that the closure probability in timestamp period T_4 is much higher than that in other timestamp periods. For simplicity, we only show the overall observations in our later discussion without considering the status of each timestamp period.

Since we are interested in the major factors that contribute to triadic closure formation, we first investigate the impact of different factors from three aspects: user demographics, network characteristics, and social perspectives. For user demographics, we consider location, gender, and user's verified status. For network characteristics, we focus on the network structure before and after the triadic closure. For social perspectives, we focus on the popularity of the people within the triads, people who span "structural holes", the gregariousness of users, and status theory. We also consider the effects of social interaction.

5.3.2.1 User Demographics

Location From user profiles, we can obtain location information (province and city that the user comes from). We test whether a user's location will influence the closure of a triad. We can see from Figure 5.3(a), if three users all come from the same

province, the probability that the open triads will be closed is much larger (about 4 times as large) than the case for which all users are from different province. Even if two of the three users are from the same province, the probability is obviously greater than the NULL case, where all three users are from different provinces. If we consider city scale, the result is more definitive; the probability of closure for three persons from the same city is 8 times as high as that of the NULL case. Although online social networks make distances between people smaller, location is still one important factor that influences the formation of triadic closure.

Gender We test whether or not gender homophily affects triadic closure formation. We use three-bit binary codes to indicate the gender status of a triad – i.e., $(XXX)X = 0$ or 1, where 0 means female and 1 means male. As shown in Figure 5.3(b), we can see that if the three users are all male, triadic closures is about 6 times more likely to form than the case in which all three users are female. We also notice that with more male users in a triad, the triad will have a higher probability to become closed. For example, for any case (such as 001) in Figure 5.3(b), if we replace one female user of "0" with a male user ("1"), the probability that the triad will close will increase to 0.6-1 times higher.

Verified Status In Weibo, users can choose to verify their real status; e.g., organization, company, famous people, media, active users, etc. In some sense, a verified user could be regarded as a celebrity. Among the 1.7 million users in our sample, about 0.7 million users have verified their status. On the other hand, we have 21,622,013 closed triads, among which we have 7,608,598 closed triads with two verified users and 8,995,533 with three verified users.

Here we check whether verified status affects triadic closure formation. We use three-bit binary codes $(XXX)(X = 0$ or 1, where 0 means status is not verified, and 1 means status is verified) to represent triad status. As shown in Figure 5.3(c), we can see that if the middle user (i.e., user B) verified his/her status, it has negative influence on triadic closure ($P(X0X) > P(X1X)$), while if the other users verified their status, an open triad is more likely to become closed($P(XX1) > P(XX0)$, $P(1XX) > P(0XX)$). For example, if users A and C verified their status, the probability that an open triad will close is about 70 times higher than the case in which only user B verified his/her status.

5.3.2.2 Network Characteristics

We then check the correlation between characteristics of the microblogging network and the formation of triadic closure. In a directed network, there are 13 possible three-node subgraphs [28] as shown in Figure 5.1 – if isomorphous subgraphs are only counted once – among which there are 6 open triads and 7 closed triads.

Among all the open triads, open triad 3 is the most frequent, which is around 95% of all open triads. The case corresponds to the tendency of users in Weibo to follow "super stars", such as a famous person or news media, to get information. Figure 5.4(a) shows the distribution of new triadic closures. We can see that triad 6 has the largest number among all the closed triads, while triad 7 has the smallest number.

Figure 5.4(b) shows the probability that each open triad forms triadic closure. We can see that open triad 5 has the highest probability of becoming closed, which means if there exist two two-way (reciprocal) relationships in an open triad, it is likely that the triad becomes closed. Meanwhile, open triad 3 is the least likely to form triadic closure, as there are large numbers of this kind of open triads(94.9%).

Figure 5.4(c) shows the probability for each type of open triad to change from into each type of closed triad. We can see that a one-way relationship is much easier to build than a two-way relationship; e.g., $P_{5\to11} > P_{5\to12}$.

5.3.2.3 Social Perspectives

We turn now to several social metrics, to check how they influence triadic closure formation. These include: popularity, structural hole, gregariousness, status, and interaction.

Popularity For popularity, we test this question: If one of the three users in an open triad is a popular user (e.g., an opinion leader, a celebrity), how likely is the open triad to become closed? Here we employ Pagerank [31] to estimate the users' popularity in the network, based on which the top-1%-ranked users[4] are defined as "popular" users while the rest are viewed as ordinary ones. Among all the 21,622,013 closed triads, we have 5,918,130 with any popular users, and 461,396 with three popular users.

We also test popularity using other metrics, like in-degree, and find similar patterns. We use three-bit binary codes $(XXX)(X = 0 \text{ or } 1)$ to represent a user's status: 0 for an ordinary user and 1 for a popular user. Figure 5.5(a) shows the correlation between users' popularity and the proportion of triadic closures to total open triads. We can see that if the middle user – i.e., user B – is a popular user, the probability to close the open triads is small. We explain this phenomenon thus: User B can be a super star, a politician, or an official account, which has a lot of followers and relatively few followees, and plays a more important role than ordinary users in the network; meanwhile ordinary users, such as A and C, follow them, but are unlikely to interact with each other, so the probability to close the open triads is small in these cases. But if the three users are all popular users, the probability that the open triads will close is high.

Social Structural Hole The theory of structural holes [4] suggests that individuals would benefit from filling the holes (called "structural hole spanners") between people or groups that are otherwise disconnected [25]. We further test whether users who span structural holes will have different influences on the formation of closed triads. Again, we use three-bit binary codes $(XXX)(X = 0 \text{ or } 1)$ to represent triad status: 0 indicates an ordinary user and 1, a structural hole spanner. Figure 5.5(b) shows the correlation between users' social structural hole properties and the proportion of triadic closures to total open triads. We can see from this figure that if only user B is a structural hole spanner, the open triad is not likely to become closed. In another case, if A or C is a structural hole spanner, A and C are more willing to connect with

[4]We follow the work [40] which has shown that less than 1% of Twitter users produce 50% of its content, and [26], which also uses the top-1%-ranked users to study triadic closure in Twitter.

each other to get more resource for themselves [30, 33, 35], so the open triads are more likely to become closed.

Gregariousness Gregariousness represents the degree that a user is social and enjoys being in crowds. In sociology, gregariousness is often simply represented by out-degree; i.e., a high out-degree reflects a strong desire to be socially active and accepted. Here we examine whether gregariousness will play some role in triadic closure formation. Similarly, we view the top-1%-ranked out-degree users as gregarious. Among all the 21,622,013 closed triads, we have 1,105,892 closed triads with two gregarious users and 109,030 with three gregarious users.

We still use three-bit binary codes $(XXX)(X = 0 \text{ or } 1)$ to represent the triad status: 0 refers to a common user and 1 refers to a gregarious user. Figure 5.5(c) shows the correlation between users' gregariousness and the ratio of triadic closures to the total open triads. We can see from this figure that if three users are all common users (000), open triads are less likely to become closed. On the other hand, if the three users are all gregarious (111), the open triads have a high probability of becoming closed – almost 39 times as high as that of case 000. We also notice that with more gregarious users in a triad, the triad will have a higher probability to become closed. For example, for any case (such as 001) in Figure 5.5(c), if we replace one user of "0" with a gregarious user ("1"), the probability that the triad becomes closed will double or triple.

Transitivity Transitivity [21, 39] is an important concept that attaches many social theories to triadic structures. One social relation among three users A, B, and C, is transitive if the relations $A \to B$, $B \to C$, and $A \to C$ are present. Extending this definition, a triad is said to be transitive if all the relations it contains are transitive. For example, where A's friends' friends are A's friends as well. In Weibo, it is more likely (98.8%) for users to be connected in a transitive way.

Social Interaction We next consider the effects of interaction information upon the triads – say, retweet information. For each user, the crawler collected the $1,000$ most recent microblogs (including tweets and retweets). Since we focus on retweet behaviors in the microblogging network, we select $300,000$ popular microblog diffusion episodes from the dataset. Each diffusion episode contains the original microblog and all its retweets. On average, each microblog has been retweeted about 80 times. The sampled dataset ensures that for each diffusion episode, the active (retweet) statuses of followees in one τ-ego network[5] is completed. The dataset was previously used for studying social influence in the diffusion process [43]. With this retweeting data, we study how triadic closure formation has been influenced by the retweeting behaviors.

First, let us define some notations: $t_{R_{BC}}$ denotes the time that a retweeting behavior happens between B and C; $t_{R_{AB}}$ denotes the time that a retweet happens between A and B. If there are several actions, $t_{R_{BC}}$, $t_{R_{AB}}$ denotes the time that the first action happens; $t_{L_{AC}}$ denotes the time that link AC is established. For retweeting behaviors,

[5] A τ-ego network means a subnetwork formed by the user's τ-degree friends in the network; $\tau \geq 1$ is a tunable integer parameter that controls the scale of the ego network.

according to the time ordering of retweeting behaviors, we have the following four cases:

I) User B posted one tweet, then users A and C retweeted it respectively. Given that A retweeted it earlier than C, we have $t_{R_{BC}} > t_{R_{AB}}$;

II) Assume that A has retweeted some tweets posted by B and C has retweeted some tweets posted by B. Suppose A did it earlier than C; then we have $t_{R_{BC}} > t_{R_{AB}}$;

III) User A posted one tweet, then user B and C retweeted it respectively. Given that B retweeted it earlier than C, we have $t_{R_{BC}} > t_{R_{AB}}$;

IV) Assume that B has retweeted some tweets posted by A and C has retweeted some tweets posted by B. Suppose A did it earlier than C; then we have $t_{R_{BC}} > t_{R_{AB}}$.

Our intent is to study whether one kind of retweeting will influence triadic closure formation. Figure 5.6 shows the probability of triadic closure in different cases. We see that if the connecting node B is the first to post a tweet (case I and II), regardless of whether others retweet the tweet or once retweeted his tweets, the retweeting behavior has little influence on triadic closure formation. However, if user A is the initial user who posts a tweet (case III and IV), the open triads are more likely (about 3 times as probable) to become closed.

5.3.2.4 Summary

We summarize our observations as below:

- Male users trigger triadic closure formation. The probability that three male users form a closed triad is $6\times$ as high as that of three female users.
- Gregarious users help form closed triads. The probability that three gregarious users form a closed triad is $39\times$ as high as that of three ordinary users.
- Celebrity users are more likely to form closed triads. Three users with high Pagerank scores are $421\times$ as likely to form closed triads as three ordinary users. We also find similar patterns in the study for verified status users.
- Structural hole spanner is eager to close an open triad for more social resources ($> 10\times$ higher than that of three ordinary users). On the other hand, they are also reluctant to have two disconnected friends to be linked together.
- Interaction among users plays an important role in forming closed triads. An open triad is $3\times$ as likely to become closed if there is interaction among the users in certain cases, than if there is none.
- In general, the closing action is often done by the third user (Figure 5.3(b), Figure 5.5(c)); since the third user is the last "active" user, he or she is more

willing than the other users to connect the link. However, if the user has some social position, like "celebrity" or "resource holder," then ordinary users are more likely to connect with them (Figures 5.3(c), 5.5(a) and 5.5(b)) and close the triad.

5.4 Triadic Closure Prediction

Based on the observations in section 5.3, we see that the closure of an open triad not only depends on the demographics of the users involved in the triad, but is also influenced by the structural position and social position of the users within the triad in the network. Technically, the challenge in triadic closure prediction is how to integrate all relevant information in a unified model. In this paper, we present a Triad Factor Graph (TriadFG) model and its variations (TriadFG-BF, TriadFG-KF, TriadFG-EKF) for triadic closure prediction. A similar model has been studied in [26] for reciprocal relationship prediction.

5.4.1 Modeling

For a given network $G^t = \{V, E, X, Y\}$ at time t, we first extract all candidate open triads and define features for each triad. Here we use Tr to denote candidate open triads; X to denote features defined for candidate open triads – e.g., the demographics of users as analyzed in Section 5.3; Y indicates whether open triads become closed or not. With this information, we can construct a TriadFG model.

For simplicity, we remove the superscript t if there is no ambiguity. Therefore, according to the Bayes theorem, we can get the posterior probability of $P(Y|\mathbf{X}, G)$ as below:

$$P(Y|\mathbf{X}, G) = \frac{P(\mathbf{X}, G|Y)P(Y)}{P(\mathbf{X}, G)} \propto P(\mathbf{X}|Y) \cdot P(Y|G) \tag{5.1}$$

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 triad Tr, given their label Y. Assuming that the generative probability of attributes, given the label of each triad, is conditionally independent, then

$$P(Y|\mathbf{X}, G) \propto P(Y|G) \prod_i P(\mathbf{x}_i|y_i) \tag{5.2}$$

$$P(\mathbf{x}_i|y_i) = \prod_j F_j(x_{ij}, y_i), \tag{5.3}$$

where $P(\mathbf{x}_i|y_i)$ is the probability of generating attributes $\mathbf{x}_i$ given the label y_i, $F_j(x_{ij}, y_i)$ is j^{th} factor function defined for attribute x_i.

The problem is how to instantiate the probabilities $P(Y|G)$ and $F_j(x_{ij}, y_i)$. In principle, they can be instantiated in different ways. In this work, we instantiate them in the following three ways.

5.4.1.1 TriadFG-BF

Straightforwardly, we model these factor functions in a Markov random field, and by the Hammersley-Clifford theorem [12], we have

$$F_j^{BF}(x_{ij}, y_i) = \frac{1}{Z_1} \exp\{\alpha_j f_j(x_{ij}, y_i)\} \tag{5.4}$$

$$P(Y|G) = \frac{1}{Z_2} \exp\{\sum_c \sum_d \mu_d h_d(Y_{Tr_c})\}, \tag{5.5}$$

where Z_1 and Z_2 are normalization factors. Eq. 5.4 indicates that we define a feature function $f_j(x_{ij}, y_i)$ for each attribute x_{ij} associated with each triad, where α_j is the weight of the j^{th} attribute. Eq. 5.5 represents that we define a set of correlation feature functions $\{h_d(Y_{Tr_c})\}_d$ over each triad Tr_c in the network, where μ_d is the weight of the d^{th} correlation feature function, and Y_{Tr_c} is correlation attribute associated with triad Tr_c.

For factor functions $f_j(x_{ij}, y_i)$, and $h_d(Y_{Tr_c})$, it can be defined as a binary function. For example, if three users in one triad come from the same city, then a feature $f_j(x_{ij}, y_i)$ is specified as 1; otherwise it is 0. Note that such a feature definition is often used in graphical models such as Conditional Random Fields [18].

We call this approach, Triad Factor Graph with Binary Function (TriadFG-BF).

5.4.1.2 TriadFG-KF

Generally speaking, the binary feature function can discriminate closed triads and open triads. However, it cannot accurately capture correlation between features. To this end, we propose a variant of the TriadFG model: TriadFG with Kernel Function (TriadFG-KF). Given some attribute samples $\mathbf{X}$, we want to choose feature function F so that $(\mathbf{X}, F)$ is as similar as possible to the training samples. In this sense, we can use a kernel function as a similarity measure/weighting function to estimate variable density. Kernel methods like SVM have led to generalizations of algorithms in the machine learning field, and to successful real-world applications [3, 37, 41]. In this paper, we use kernel-density estimate (KDE) [38] to estimate the density functions of samples $\mathbf{X}$.

To form a kernel-density estimate, we need to place a kernel – a smooth, strongly peaked function – at the position of each data point, then add up the contributions from all kernels to obtain a smooth curve, which can be evaluated at any point along the x axis. For instance, for a network structure feature, we have six open triads, and we want to obtain some functions to see which kind of open triads are more likely to become closed. In order to use kernel-density estimates, we need to know the distance between the incoming samples. To this end, we define the distance metric based on the similarity of open triads.

We set a 3×3 matrix with rows and columns labeled by vertices for every open triad, with a 1 or a 0 in position (m_i, m_j), according to whether there is a link from m_i to m_j. So we have the matrix representations of open triads in Figure 5.7. Hence, we can define the similarity of triads using a Pearson's correlation coefficient as follows:

Definition 5.3 **[Triad Similarity]** Suppose triad i has matrix representation I and triad j 's matrix representation is J; then the similarity $Sim(i,j)$ of triad i and triad j is

$$Sim(i,j) = \frac{\sum_n (I_n - \bar{I})(J_n - \bar{J})}{\sqrt{\sum_n (I_n - \bar{I})^2}\sqrt{\sum_n (J_n - \bar{J})^2}}, \tag{5.6}$$

where n is the number of entries in the matrix, $\bar{I} = \frac{1}{n}\sum_n I_n$, $\bar{J} = \frac{1}{n}\sum_n J_n$.

Since the distance function is required to satisfy the four conditions [36]: non-negativity, identity of indiscernibles, symmetry, and triangle inequality, we define the triad similarity-based distance function as follows:

Definition 5.4 **[Triad Distance]** Suppose the similarity between triad i and triad j is $Sim(i,j)$; we define the distance $Dis(i,j)$ between these two triads as

$$Dis(i,j) = \sqrt{1 - Sim(i,j)} \tag{5.7}$$

Suppose that the region that encloses the N examples is a hypercube with sides of length β centered at the estimation point x; then its volume is given by $V = \beta^D$, where D is the number of dimensions. We can use kernel function $k(\cdot)$ to find the number of examples that fall within this region. The total number of points inside the hypercube is then

$$Q = \sum_{n=1}^{N} k(\frac{x - x_n}{\beta}) \tag{5.8}$$

So the structure feature function can be rewritten as

$$F_j^{KF}(x_{ij}, y_i) = \sum_{i=1}^{N} \frac{1}{\beta} k(\frac{x - x_{ij}}{\beta}), \quad j = s; \tag{5.9}$$

where $k(\cdot)$ is the kernel function – e.g., Gaussian kernel $k(x) = \frac{1}{\sqrt{(2\pi)}} \exp(-\frac{1}{2}x^2)$, β is the kernel bandwidth, and s represents the structure feature. The kernel-density estimates of structure information using the Gaussian kernel is shown in Figure 5.8 (green curve), and the histogram of the distance to open triad 3 is shown in Figure 5.8 (blue part).

For other factors, we model them similarly in TriadFG-BF. Thus, we have

$$F_j^{KF}(x_{ij}, y_i) = \begin{cases} \sum_{i=1}^{N} \frac{1}{\beta} k(\frac{x - x_{ij}}{\beta}), & j = s, \\ \exp\{\alpha_j f_j(x_{ij}, y_i)\}, & j \neq s \end{cases} \tag{5.10}$$

We name this approach, Triad Factor Graph with Kernel Function (TriadFG-KF).

5.4.1.3 TriadFG-EKF

With the discoveries regarding network structure, and taking TriadFG-BF into account, we can use the kernel function together with an exponential function to rewrite $F_j(x_{ij}, y_i)$ as follows:

$$F_j^{EKF}(x_{ij}, y_i) = \begin{cases} \exp\{\alpha_j \sum_{i=1}^{N} \frac{1}{\beta} k(\frac{x - x_{ij}}{\beta})\}, & j = s \\ \exp\{\alpha_j f_j(x_{ij}, y_i)\}, & j \neq s \end{cases} \tag{5.11}$$

We call this approach, Triad Factor Graph with Exponential Kernel Function (TriadFG-EKF).

Objective Function Based on the above equations, we can define the following log-likelihood objective function $\mathcal{O}(\theta) = \log P_\theta(Y|\mathbf{X}, G)$

$$\mathcal{O} = \sum_{i}^{|Tr|} \{ \sum_{j}^{|fe|} \alpha_j F_j(x_{ij}, y_i) + \sum_{c} \sum_{d} \mu_d h_d(Y_{Tr_c}) \}, \\ - \log Z \tag{5.12}$$

where Z is a normalization factor to guarantee that the result is a valid probability; $|Tr|$ denotes the number of candidate (open) triads in the network; $|fe|$ is the number of features defined for the triads (more details for feature definition are given in Section 5.4.2); x_{ij} is the j^{th} feature value of the i^{th} triad; c corresponds to a correlation function; and Tr_c indicates a set of all related triads in the correlation function.

Example To provide a concrete understanding of the proposed model, we give a simple example of TriadFG in Figure 5.9. The left part is the input network, where we have five users and four kinds of following links among them. From the input network we can derive six open triads – e.g., (v_1, v_2, v_3) and (v_1, v_3, v_4). In the prediction task, we view each open triad as a candidate; thus we have six candidates, which are illustrated as blue ellipses in the right-hand model. All features defined over open triads are denoted as such – i.e., $f(v_1, v_2, v_3)$. In addition, we also consider social correlation. For example, the closure of (v_1, v_2, v_3) may imply a higher probability that (v_1, v_3, v_4) will also be closed at time $t + 1$. Given this, we build a correlation function $h(\cdot)$ among related triads. Based on all the considerations, we construct the TriadFG (as shown in Figure 5.9).

5.4.2 Feature Definitions

We now depict how we define the factor functions in our models. According to the observations in the previous section, we define 11 features of five categories: Network Structure(N), Demographics(D), Verified Status(V), Social Information(S), and Social Interaction(I).

Network Structure According to Figure 5.4(b), we notice open triads 2, 4, 5 are more likely to be closed than others, so for TriadFG-BF, we define one feature:

whether the open triad is of open triad 2, 4, or 5. For TriadFG-KF and TriadFG-EKF, we use a kernel-density estimate to get the feature value.

Demographics Here we consider location and gender features. For location, we define one feature: whether the three users come from the same place; for gender, we define two features: whether all three users in one triad are female or male.

Verified Status We define two features for verified status: whether the connecting user verified her status or not; other users have the opposite status (cases 010 and 101).

Social Information We consider popularity, structural hole spanning, and gregariousness here. For popularity, we define one feature: whether all the three users in the triad are popular users. For structural hole spanning, we define one feature: whether user A and user B are structural hole spanners. For gregariousness, we define two features: whether all three users are gregarious users, and whether the three users follow the pattern: A and C are gregarious users while user B isn't.

Social Interaction For the problem of triadic closure prediction with interaction information, we define one feature for social interaction: whether a retweeting action happens among the three users in one triad.

5.4.3 Learning and Prediction

We then want to estimate a parameter configuration of the TriadFG model $\theta = (\{\alpha_j\}, \{\mu_d\})$ that maximizes the log-likelihood objective function, $\theta = \arg\max \mathcal{O}(\theta)$. We employ a gradient descent method for model learning. The basic idea is that each parameter – e.g., μ_d – is assigned an initial value, and then the gradient of each μ_d with regard to the objective function is derived. Finally, the parameter with learning rate η is updated. The details of the learning algorithm can be found in [26].

With the estimated parameters θ, we can predict the labels of unknown variables $y_i =?$ by finding a label configuration that maximizes the objective function – i.e., $Y^\star = \arg\max \mathcal{O}(Y|X, G, \theta)$. To do this, we use the learned model to calculate the marginal distribution of each open triad with unknown variable $P(y_i|\mathbf{x}_i, G)$, and assign each open triad a label of the maximal probability.

5.5 Experiments and Discussions

5.5.1 Experiment Setup

We use the dataset described in Section 5.3 in our experiments. To quantitatively evaluate the effectiveness of the proposed model and the methods for comparison, we divide the network into seven timestamp periods, by viewing every four days as a timestamp period. For each timestamp period, we divide the network into two subsets by using the first two-thirds of the data as a training set and the rest as a test set. Our goal is to predict whether an open triad will become closed in the test set.

Comparison Methods and Evaluation Measures We compare the proposed three approaches with two alternative baselines.

SVM Uses the same attributes associated with each triad as features to train a classification model, and then uses the classification model to predict triadic closure in the test data.

Logistic Similar to the SVM method. The only difference is that it uses a logistic regression model as the classification model.

TriadFG-BF Represents the proposed TriadFG model with binary feature functions (Cf. § 5.4.1.1).

TriadFG-KF Represents the proposed TriadFG model with kernel feature functions (Cf. § 5.4.1.2).

TriadFG-EKF Represents the proposed TriadFG model with exponential kernel functions (Cf. § 5.4.1.3).

For SVM and Logistic, we use Weka [11]. All the TriadFG models are implemented in C++, and all experiments are performed on a PC running Windows 7 with an AMD Opteron(TM) Processor 6276$(2.3GHz)$ and 4GB memory. We evaluate the performance of different approaches in terms of accuracy, precision, recall, and F1-Measure.

5.5.2 Triadic Closure Prediction

Prediction Performance We now list the performance results for different methods in Table 5.3. It can be seen that our proposed TriadFG-BF outperforms the other two comparison methods (SVM and Logistic), and TriadFG-EKF performs the best among all the methods. In terms of F1-Measure, TriadFG-BF achieves a $+7.43\%$ improvement over SVM, and $+7.85\%$ over Logistic. TriadFG-KF achieves a $+6.93\%$ improvement over TriadFG-BF, $+14.88\%$ over SVM, and $+15.32\%$ over Logistic. TriadFG-EKF achieves a $+1.24\%$ improvement over TriadFG-KF, $+8.26\%$ over TriadFG-BF, $+16.31\%$ over SVM, and $+16.76\%$ over Logistic. Our proposed algorithm is much better than SVM and Logistic in terms of F1-Measure. TriadFG-BF perform slightly better than they do because it uses binary feature functions that do not capture the similarities/correlations between different features. That is why we propose TriadFG-KF and TriadFG-EKF, which incorporate kernels to quantity the similarities. Meanwhile, the new proposed methods also do better on recall, which is partly because TriadFG can detect some cases by leveraging transitive correlation and homophily correlation.

Factor Contribution Analysis For triadic closure prediction, we examine the contribution of four different factor functions: Network Structure(N), Demographics(D), Verified Status(V), and Social Information(S). We first rank the individual factors by respectively each factor from our model and evaluating the decrease in prediction performance. Thus, a larger decrease means a higher predictive power for the removed factor. We thus rank these factors according to predictive power as follows: Network Structure(N)$>$ Verified Status(V)$>$ Demographics(D)$>$ Social Information(S).

We then remove them one by one in reverse order of their prediction power.

Table 5.3: Triadic closure prediction performance

Algorithm	Accuracy	Precision	Recall	F1-score
Logistic	0.7394	0.7657	0.7393	0.7316
SVM	0.7422	0.7683	0.742	0.7344
TriadFG-BF	0.7523	0.6989	0.9068	0.7890
TriadFG-KF	0.8426	0.8102	0.8613	0.8482
TriadFG-EKF	**0.8444**	**0.8360**	**0.9084**	**0.8564**

Table 5.4: Triadic closure prediction performance of each open triads

Triads	Accuracy	Precision	Recall	F1-score
0	0.5479	0.5533	0.5478	0.5335
1	0.5320	0.5472	0.5322	0.4695
2	0.5894	0.6085	0.5895	0.5797
3	**0.6420**	**0.7058**	**0.6420**	**0.6097**
4	0.5988	0.6145	0.5990	0.5823
5	0.5551	0.5562	0.5552	0.5503

We denote TriadFG-S as removing social information and TriadFG-SD as removing demographics, finally removing verified status, denoted as TriadFG-SDV. As shown in Figure 5.10, we can observe a slight performance decrease when ignoring social information and demographics, which means these factors contribute significantly to predicting triadic closure.

Prediction Performance on Triads We now consider the prediction performance for each of the triads shown in Table 5.4. We can see that for triad 3, the prediction performance is much better than others, while for triad 1, the performance is the worst. This may be because triad 3, which corresponds to the case in which two fans follow one popular user, can be trained with a large number of features in our model, such as social information, which gives better prediction results than for other kinds of triads. However, the closure of triad 1, which has some transitive cases, can not be easily predicted using our features, and shows worse prediction performance than triad 3.

5.5.3 Triadic Closure Prediction With Interaction Information

Prediction Performance Now we consider the triadic closure prediction problem with interaction information. Here, we consider retweeting behavior as interaction information.

Since TriadFG-EKF performs the best on problem 1, we use TriadFG-EKF here to study this extended problem. The performance of TriadFG-EKF and TriadFG-EKF-I (with interaction information) is shown in Table 5.5. We can see that our proposed TriadFG-EKF-I outperforms TriadFG-EKF. In terms of F1-Measure, TriadFG-EKF-I achieves a $+7.55\%$ improvement over TriadFG-EKF, which indicates that in-

Table 5.5: Triadic Closure Prediction Performance with Interaction Information

	Accuracy	Precision	Recall	F1-score
TriadFG-EKF	0.6805	0.6834	0.7075	0.6953
TriadFG-EKF-I	**0.7276**	**0.7149**	**0.7838**	**0.7478**

teraction information, such as retweeting behavior, plays an important role. We will further discuss how much it contributes to triadic closure prediction.

Factor Contribution Analysis In this section, we again examine the contribution of five different factor functions, especially the retweeting function: Network Structure (N), Demographics (D), Verified Status (V), Social Information (S) and Interaction (I). According to predictive power of each factor, we rank these factors as follows: Interaction (I) > Network Structure (N)> Verified Status (V)> Social Information (S)> Demographics (D). We then remove them one by one in reverse order of their prediction power. TriadFG-D denotes removing Demographics; TriadFG-SD denotes removing Social Information from that set; TriadFG-SDV signifies removing Verified Status from that; and TriadFG-SDVN denotes removing Network Structure.

As shown in Figure 5.11, we observe a slight performance decrease when ignoring Social Information and Demographics, but a large performance decrease when ignoring Network Structure – which means Network Structure information also contributes a lot to the prediction of triadic closure. However, Interaction information has the strongest predictive power here, which indicates that Interaction information is a good feature in this microblogging service, and plays an important role in the establishment of friendship.

5.5.4 Comparison with Twitter Observations

We compare the results with a similar study about popularity within triads on Twitter [14] and find:

- Both results demonstrate the phenomenon of "the rich get richer" – i.e., $P(1XX) > P(0XX)$, which validates the mechanism of preferential attachment in both networks (Twitter and Weibo).
- In Twitter, popular users play an important role in forming closed triads – i.e., $P(X1X)$ is about three times as high as $P(X0X)$, while in Weibo, the result is opposite. Possibly it is because Weibo provides more features to help users interact with each other, and ordinary users have more chances to connect with others. In China, Weibo is a combination of Twitter and Facebook, and integrates the features of both. For example, for the #IceBucketChallenge, between July 29 and August 13, about 135 thousand tweets were posted in

Twitter[6]; however, in Weibo, more than 1.55 million tweets were posted between July 29 and August 20[7].

- The probability $P(111)$ for popular users in Weibo is much higher than that in Twitter. In Twitter, $P(111)$ is twice as high as $P(000)$; while in Weibo, $P(111)$ is eight times as high, which implies that popular users in China have more closeness connections.

5.6 Related work

In terms of related work, we identify two areas: triadic closure and link prediction in social networks. We will discuss them in detail as follows:

Triadic Closure Study There are many studies on triadic closure study. They mainly focus on the following three aspects:

1) *Network evolution/formation*. One of the fundamental issues of social networks is to reveal the possible generic laws governing the formation/evolution of networks. Since it is unrealistic to get global information for preferential attachment processes to establish new social ties, the triadic closure principle, whose assumption is that a node's linking dynamics only rely on its neighbors or next neighbors is relevant to social network formation. Klimek et al. [16] and Li et al. [22] both declared that triadic closure could be identified as one of the fundamental dynamic principles in social multiplex network formation/evolution. [6, 7, 19] also provided some triadic-closure-based network generation models.

2) *Network structure*. Milo et al. [27] [28] defined recurring significant patterns of interconnections as "network motifs" and emphasized the importance of these patterns, which included 6 open triads and 7 closed triads, which we use in this paper. Romero et al. [32] studied the problem of triadic closure and developed a methodology based on preferential attachment for studying the directed triadic closure process. Zhang et al. [42] use triadic structures to study link diffusion process.

3) *Triadic closure formation*. Lou et al. [26] investigated how a reciprocal link is developed from a parasocial relationship, and how the relationships further develop into triadic closure, in a Twitter dataset. Zignani et al. [45] studied the triadic closure problem on undirected networks like Facebook and Renren.

However, none of these works systematically studied triadic closure formation and prediction in real large-scale directed networks.

Link Prediction Our work is also related to the link prediction problem, 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 prediction usually assigns scores to potential links based on intuition – the more similar the pair

[6] http://www.bostonglobe.com/business/2014/08/15/facebook-million-icebucketchallenge-videos-posted/ 24D8bnxFlrMce5BRTixAEM/story.html

[7] http://media.people.com.cn/n/2014/0821/c120837-25512105.html

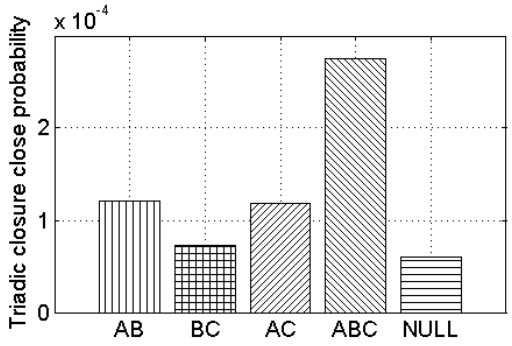
of users are, the more likely they are to be linked. Various similarity measures of users are considered, such as preferential attachment [29], and the Katz measure [15]. [24] presented a flow-based method for link prediction. A survey of unsupervised link prediction research can be found in [23].

There are also a number of works that employ supervised approaches to predict links in social networks, such as [2, 20, 24]. [2] proposed a supervised random walk algorithm to estimate the strength of social links. [20] employed a logistic regression model to predict positive and negative links in online social networks.

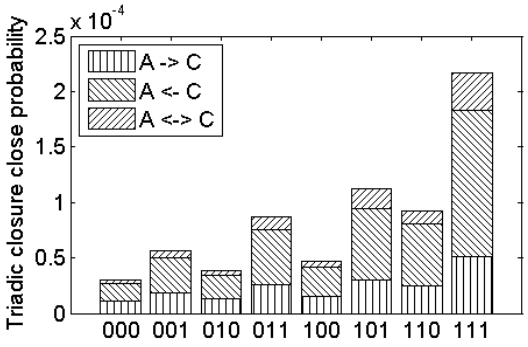
However, unlike link prediction studies, we focus only on triadic closure, which means we only focus on the last "link" that constitutes the closed triad. Moreover, our model is dynamic and can learn from the evolution of the Weibo network. We also combine social theories into the semi-supervised learning model.

5.7 Conclusion

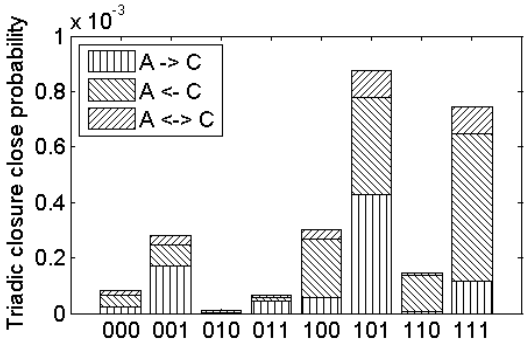
In this paper, we study an important phenomenon of triadic closure formation in dynamic social networks. Employing a large microblogging network (Weibo) as the source in our study, we formally define the problem and systematically study it. We propose a probabilistic factor model for modeling and predicting whether three persons in a social network will finally form a triad. Our experimental results on Weibo show that the proposed model can more effectively predict triadic closure than alternative methods, in terms of F1 measurement.



(a) Location correlation

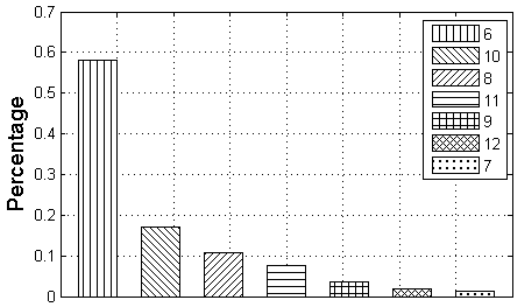


(b) Gender correlation

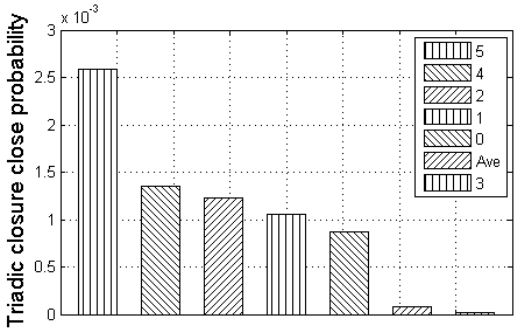


(c) Verified status correlation

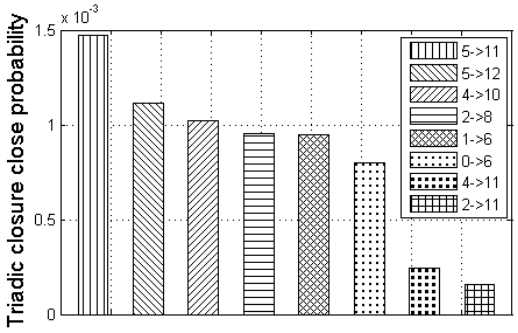
Figure 5.3: User Demographics. Y-axis: probability of triadic closures. The status of the third link – the new formed link is presented in a different color; e.g., blue means the third link is accomplished by user A, who follows user C. (a) X-axis: represents whether certain users are from the same province; e.g., AB means that only A, B are in the same province. $NULL$ means users in a triad all come from different provinces. (b) X-axis: represents genders in the triad; 0 means female and 1 means male. (c) X-axis: represents the verified status of the triad; 0 means the user hasn't been verified and 1 means the user is verified.



(a) Distribution of close triads

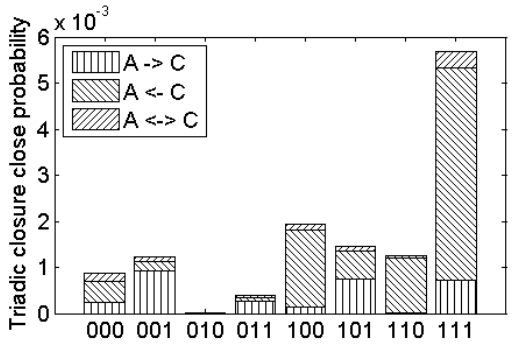


(b) Open triads that form triadic closure

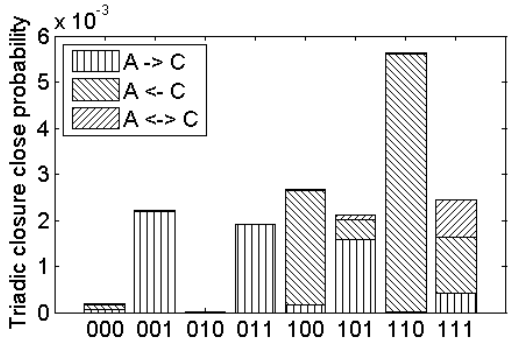


(c) Triads evolution

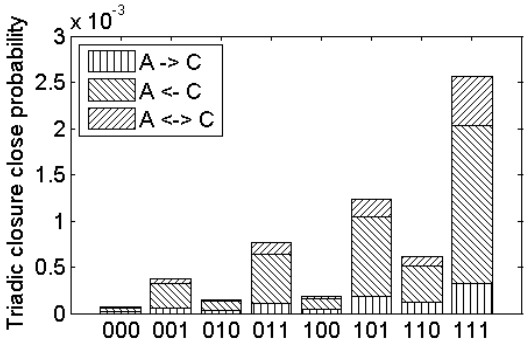
Figure 5.4: Network Characteristics. (a) Y-axis: Percentage of newly formed closed triads. (b) Y-axis: probability that each open triad becomes closed. The number by the color bars means the index of open triads. (c) Y-axis: probability for each type of open triad (i.e., triad 0) to change from into each type of closed triad (i.e., triad 6). Expressions attached to color bars represent the probability that an open triad becomes a specific triadic closure; e.g., $0 \rightarrow 6$ represents the probability that triad 0 forms triad 6.



(a) Popularity correlation



(b) Structural hole correlation



(c) Gregariousness correlation

Figure 5.5: Social Perspectives. Y-axis: probability that triadic closures form. The status of a newly formed link is presented in a different color; e.g., blue represents the fact that a third link is accomplished by user A, who follows user C. (a) X-axis: represents the popularity of the triad. 0 **represents an ordinary user and** 1 **represents a popular user. (b) X-axis: represents the structural hole spanner status of the triad.** 0 **means an ordinary user and** 1 **means a structural hole spanner. (c) X-axis: represents the gregariousness of the triad.** 0 **indicates an ordinary user and** 1 **is used for a gregarious user.**

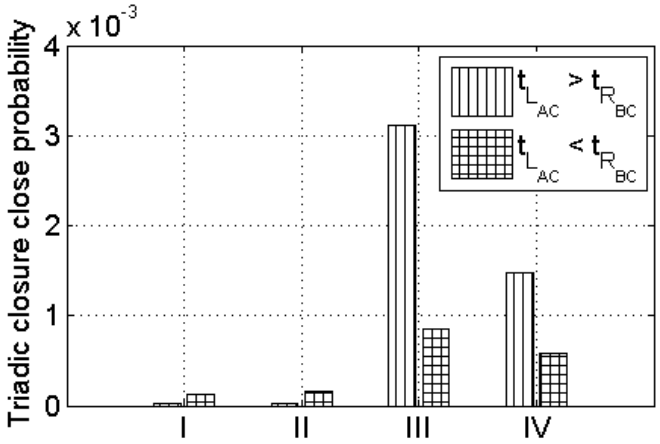


Figure 5.6: Open triads that form triadic closures with social interaction information in different cases. X-axis: Cases. Y-axis: probability that open triads form triadic closures. $t_{L_{AC}}$ means the time that link AC is established, and $t_{R_{BC}}$ means the time that a retweet happens between user B and C.

$$\begin{array}{c} \\ A \\ B \\ C \end{array}\!\!\begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \left[\begin{array}{ccc} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{array}\right] \end{array}$$

Triad 0

$$\begin{array}{c} \\ A \\ B \\ C \end{array}\!\!\begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \left[\begin{array}{ccc} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{array}\right] \end{array}$$

Triad 1

$$\begin{array}{c} \\ A \\ B \\ C \end{array}\!\!\begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \left[\begin{array}{ccc} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{array}\right] \end{array}$$

Triad 2

$$\begin{array}{c} \\ A \\ B \\ C \end{array}\!\!\begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \left[\begin{array}{ccc} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{array}\right] \end{array}$$

Triad 3

$$\begin{array}{c} \\ A \\ B \\ C \end{array}\!\!\begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \left[\begin{array}{ccc} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{array}\right] \end{array}$$

Triad 4

$$\begin{array}{c} \\ A \\ B \\ C \end{array}\!\!\begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \left[\begin{array}{ccc} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{array}\right] \end{array}$$

Triad 5

Figure 5.7: Matrix representation of open triads.

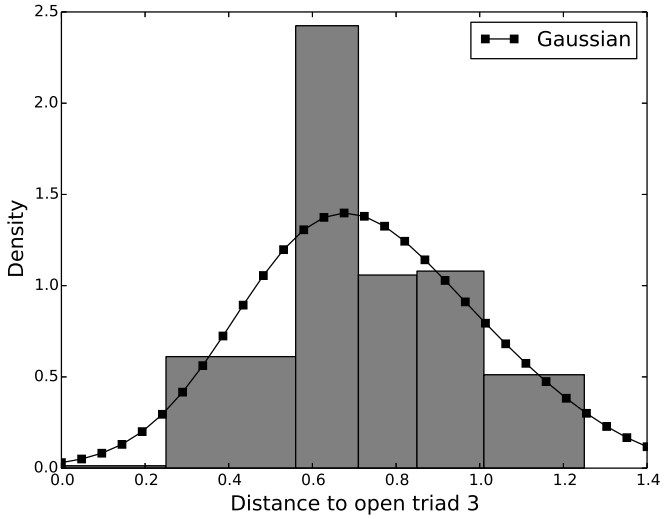


Figure 5.8: Kernel-density estimation for structure information

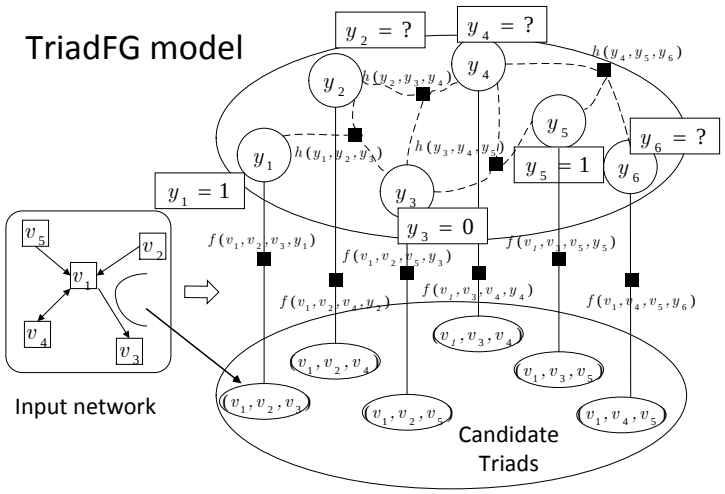


Figure 5.9: Graphical representation of the TriadFG model. There are five users in the input network. Candidate open triads are illustrated as blue ellipses in the bottom right. White circles indicate hidden variables y_i. $f(v_1, v_2, v_3)$ represents the attribute factor function, and $h(.)$, the correlation function among triads.

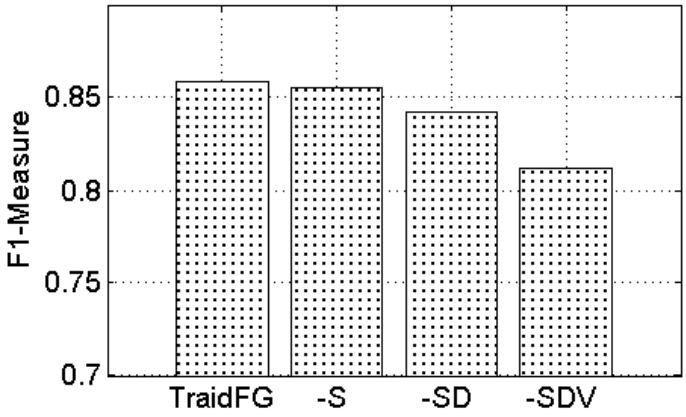


Figure 5.10: Factor contribution analysis. -S denotes ignoring social information when we use TriadFG model, -SD denotes ignoring social information and demographics while -SDV denotes further ignoring verified status information.

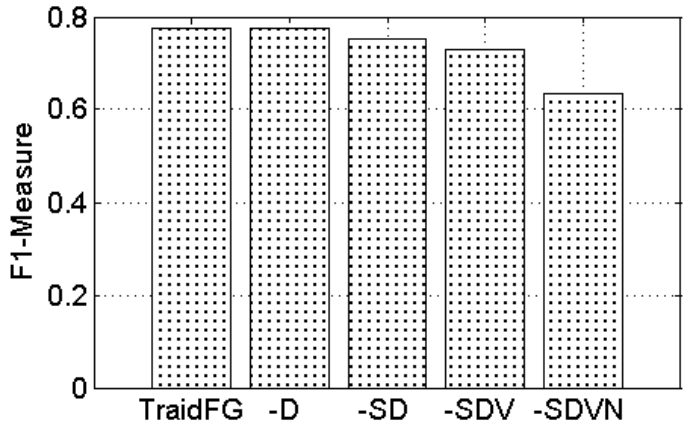


Figure 5.11: Factor contribution analysis. -D denotes ignoring Demographics when we use the TriadFG model; -SD denotes ignoring Social Information and Demographics; while -SDV denotes also ignoring Verified Status information; and -SDVN denotes further ignoring Network sSructure information.

References

[1] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD'06*, pages 44–54, 2006.

[2] Lars Backstrom and Jure Leskovec. Supervised random walks: predicting and recommending links in social networks. In *WSDM'11*, pages 635–644, 2011.

[3] Asa Ben-Hur and William Stafford Noble. Kernel methods for predicting protein–protein interactions. *Bioinformatics*, 21(suppl 1):i38–i46, 2005.

[4] Ronald S Burt. The social structure of competition. *Explorations in economic sociology*, 65:103, 1993.

[5] James Coleman. *Foundations of Social Theory*. Harvard, 1990.

[6] Yuxiao Dong, Jie Tang, Sen Wu, Jilei Tian, Nitesh V. Chawla, Jinghai Rao, and Huanhuan Cao. Link prediction and recommendation across heterogeneous social networks. In *ICDM '12*, pages 181–190, 2012.

[7] Yuxiao Dong, Yang Yang, Jie Tang, Yang Yang, and Nitesh V. Chawla. Inferring user demographics and social strategies in mobile social networks. In *KDD '14*, pages 15–24. ACM, 2014.

[8] David Easley and Jon Kleinberg. Networks, crowds, and markets. *Cambridge Univ Press*, 6(1):6–1, 2010.

[9] Jean-Pierre Eckmann and Elisha Moses. Curvature of co-links uncovers hidden thematic layers in the world wide web. *PNAS*, 99(9):5825–5829, 2002.

[10] Neil Zhenqiang Gong, Wenchang Xu, Ling Huang, Prateek Mittal, Emil Stefanov, Vyas Sekar, and Dawn Song. Evolution of social-attribute networks: measurements, modeling, and implications using google+. In *IMC'12*, pages 131–144, 2012.

[11] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[12] J. M. Hammersley and P. Clifford. Markov field on finite graphs and lattices. *Unpublished manuscript*, 1971.

[13] Paul W Holland and Samuel Leinhardt. Transitivity in structural models of small groups. *Comparative Group Studies*, 1971.

[14] John Hopcroft, Tiancheng Lou, and Jie Tang. Who will follow you back? reciprocal relationship prediction. In *CIKM'11*, pages 1137–1146, 2011.

[15] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[16] Peter Klimek and Stefan Thurner. Triadic closure dynamics drives scaling laws in social multiplex networks. *New Journal of Physics*, 15(6):063008, 2013.

[17] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *WWW'10*, pages 591–600, 2010.

[18] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML'01*, pages 282–289, 2001.

[19] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In *KDD'08*, pages 462–470, 2008.

[20] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Predicting positive and negative links in online social networks. In *WWW'10*, pages 641–650, 2010.

[21] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed networks in social media. In *CHI'10*, pages 1361–1370, 2010.

[22] Menghui Li, Hailin Zou, Shuguang Guan, Xiaofeng Gong, Kun Li, Zengru Di, and Choy-Heng Lai. A coevolving model based on preferential triadic closure for social media networks. *Scientific reports*, 3, 2013.

[23] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *JASIST*, 58(7):1019–1031, 2007.

[24] Ryan N Lichtenwalter, Jake T Lussier, and Nitesh V Chawla. New perspectives and methods in link prediction. In *KDD'10*, pages 243–252, 2010.

[25] Tiancheng Lou and Jie Tang. Mining structural hole spanners through information diffusion in social networks. In *WWW'13*, pages 837–848, 2013.

[26] Tiancheng Lou, Jie Tang, John Hopcroft, Zhanpeng Fang, and Xiaowen Ding. Learning to predict reciprocity and triadic closure in social networks. *TKDD*, 7(2):5, 2013.

[27] Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer, and Uri Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542, 2004.

[28] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, pages 824–827, 2002.

[29] Mark EJ Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):025102, 2001.

[30] David Obstfeld. Social networks, the tertius iungens orientation, and involvement in innovation. *Administrative science quarterly*, 50(1):100–130, 2005.

[31] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. *Stanford InfoLab*, 1999.

[32] Daniel M Romero and Jon Kleinberg. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. *stat*, 1050:12, 2010.

[33] K. Cook Ronald S. Burt, in N. Lin and R. S. *Structural Holes Versus Network Closure as Social Capital*. Social capital: theory and research, 2001.

[34] Alessandra Sala, Lili Cao, Christo Wilson, Robert Zablit, Haitao Zheng, and Ben Y Zhao. Measurement-calibrated graph models for social network experiments. In *WWW'10*, pages 861–870, 2010.

[35] Zuzana Sasovova, Ajay Mehra, Stephen P Borgatti, and Michaéla C Schippers. Network churn: The effects of self-monitoring personality on brokerage dynamics. *Administrative Science Quarterly*, 55(4):639–670, 2010.

[36] Berthold Schweizer and Abe Sklar. *Probabilistic metric spaces*. Courier Dover Publications, 2011.

[37] John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[38] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer, 2004.

[39] Stanley Wasserman. *Social network analysis: Methods and applications*, volume 8. Cambridge University Press, 1994.

[40] Shaomei Wu, Jake M Hofman, Winter A Mason, and Duncan J Watts. Who says what to whom on twitter. In *WWW'11*, pages 705–714, 2011.

[41] Ming-Hsuan Yang. Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods. In *FG'02*, pages 0215–0215, 2002.

[42] Jing Zhang, Zhanpeng Fang, Wei Chen, and Jie Tang. Diffusion of following links in microblogging networks. *TKDE*, 2015.

[43] Jing Zhang, Biao Liu, Jie Tang, Ting Chen, and Juanzi Li. Social influence locality for modeling retweeting behaviors. In *IJCAI'13*, pages 2761–2767, 2013.

[44] Elena Zheleva, Hossam Sharara, and Lise Getoor. Co-evolution of social and affiliation networks. In *KDD'09*, pages 1007–1016, 2009.

[45] Matteo Zignani, Sabrina Gaito, Gian Paolo Rossi, Xiaohan Zhao, Haitao Zheng, and Ben Y Zhao. Link and triadic closure delay: Temporal metrics for social network dynamics. In *ICWSM'14*, pages 564–573, 2014.