Mining Triadic Closure Patterns in Social Networks

Hong Huang[†], Jie Tang[‡], Sen Wu[♯], Lu Liu[⋆], and Xiaoming Fu[†]
[†]University of Göttingen, Germany [‡]Tsinghua University, China
{hhuang,fu}@cs.uni-goettingen.de, jietang@tsinghua.edu.cn,
[♯]Stanford University, USA [⋆]Northwestern University, USA
senwu@stanford.edu, liulu26@gmail.com

ABSTRACT

A closed triad is a group of three people who are connected with each other. It is the most basic unit for studying group phenomena in social networks. In this paper, we study how closed triads are formed in dynamic networks. More specifically, given three persons, what are the fundamental factors that trigger the formation of triadic closure? There are various factors that may influence the formation of a relationship between persons. Can we design a unified model to predict the formation of triadic closure? Employing a large microblogging network as the source in our study, we formally define the problem and conduct a systematic investigation. The study uncovers how user demographics and network topology influence the process of triadic closure. We also present a probabilistic graphical model to predict whether three persons will form a closed triad in dynamic networks. The experimental results on the microblogging data demonstrate the efficiency of our proposed model for the prediction of triadic closure formation.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Miscellaneous; H.4.m [**Information Systems**]: Miscellaneous

Keywords

Social network; Predictive model; Social influence

1. INTRODUCTION

Online social networks are already becoming a bridge to connect our physical daily life with the virtual Web space. The connection produces huge volume of data including not only the spreading information, but also user behaviors. The ubiquity of social web and the prosperity of social data offer us the opportunities to study the interaction patterns among users and to understand the generative mechanisms of different networks, which was difficult to explore before due to the lack of available data.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2576940.

In this paper, employing a large microblogging network from Weibo[1] as the basis in our study, we systematically investigate the problem of mining patterns in triadic closure process. Our major goal is to discover the fundamental factors that trigger the formation of groups among people. We further compare the discoveries from Weibo with that from Twitter. We found many generic patterns underlying the dynamics of the two networks on one hand. And on the other hand, we also identify several important patterns that behave differently, which, from one perspective, reflects the different motivations for users to use these two networks; and, from another perspective, implies the behavioral difference between the Chinese users and users in Twitter.

Based on the interesting discoveries, we further study the problem of triadic closure prediction. We present a probabilistic triad factor graph model (TriadFG), which incorporates both the discovered patterns and the network structure for predicting the formation of triadic closure. Compared with alternative methods based on SVM and Logistic Regression, the presented model can achieve significant improvement (+3.3%, $p \ll 0.01$) for triadic closure prediction.

Our study can benefit many real applications. For example, a straightforward application is to use the discovered closure patterns to help friend recommendation, which is a central application in most social networks. Actually part of the discoveries in this work has been already applied to Weibo for friend recommendation and the online A/B test demonstrates superior (+10%) of our method over the existing recommendation algorithm in the system. Besides, our work provides a basis to study the group formation [1, 16]. Other applications can also be found in social search and user behavior modeling.

Problem Formulation First, we formally define the problem. Let $G = (V, E)$ denote a static network, where V is a set of users and $E \subset V \times V$ is a set of relationships connecting those users. The network evolves over time. Suppose the network at time t is denoted as G^t. From a broad viewpoint, all possible triads construct the basic units in our study. However, for a network of N users, this results in a huge candidate space exponential to N, which is obviously infeasible in the research. To make our goal concrete, we give the definition of *closed triad*: if for any two users in a triad Δ, i.e., $\forall v_i, v_j \in \Delta$, there exists $e_{ij} \in E$, then we say that Δ is a closed triad. Our focus is then to study the last "action" that constitutes the closed triad. Formally for example, suppose at time t, for three users (A, B, C), we have a relationship between A and B, and a relationship between B and C, we call the triad (A, B, C) as a candidate *open triad*. Our goal is to study whether an open

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

triad will gradually (or not) become a *closed triad* at time $t+1$, i.e., predicting the formation of the relationship between A and C. Formally, we have the following problem definition:

Problem 1. **Triadic Closure Prediction.** Suppose given a network $G^t = (V, E)$ at time t. Let Tr^t denote a candidate open triad and we associate a hidden variable y^t to each candidate open triad. If the triad finally becomes closed at time $t+1$, then we have $y^{t+1} = 1$, otherwise $y^{t+1} = 0$. The problem becomes, if we have all the historic information, how to capture the dynamic patterns so that we can accurately predict the value of y, i.e.,

$$f : (\{G^t, Y^t\}_{t=1,\cdots,T}) \rightarrow Y^{T+1}$$

where Y^{T+1} denotes all values of the hidden variables at time $t+1$.

The problem exists in both directed or undirected networks. For example, in the undirected co-author network, suppose B coauthored with A and C separately till time t, then we want to infer whether A and C will also collaborate at the following time $t+1$ or not. In the directed networks, the problem becomes much more complicated. Specifying in this paper, we focus on the directed networks like twitter(e.g., follower network), Weibo (Chinese twitter). Figure 1(a) gives several examples of open and closed triads in a directed network.

Related Work There are a few works on triadic closure analysis. For example, Milo et al. [13] defined the recurring significant patterns of interconnections as "network motif" and emphasized the importance of these patterns (refer to Figure 1(a)). Romero and Kleinberg [16] studied the problem of triadic closure process and developed a methodology based on preferential attachment for studying the directed triadic closure process. Lou et al. [11] investigated how a reciprocal link was developed from a parasocial relationship and how the relationships further developed into triadic closure on twitter dataset. There are also several works on social network analysis based on triadic closure. E.g., [2, 8] focused on network evolution, [14] used triadic closure to define the global and local clustering coefficients. However, none of these works systematically study the triadic closure prediction in real large-scale networks.

Another line of related work is the research on microblogging data. Twitter is a very popular microblogging service worldwide. Weibo is another popular microblogging service in China. Both have attracted more than 500 millions users. Existing Microblogging study mainly centers around the following three aspects: 1) *Network topology*. Java et al. [7] studied the topological and geographical properties of the Twitter network. Kwak et al. [9] conducted a similar study on the entire Twittersphere and they observed some notable properties of Twitter, such as a non-power-law follower distribution, a short effective diameter, and low reciprocity, marking a deviation from known characteristics of human social networks. 2) *Tweet content*. Sakaki et al. [17] proposed to utilize the real-time nature of Twitter to detect a target event; while Mathioudakis and Koudas [12] presented a system, TwitterMonitor, to detect emerging topics from the Twitter content. 3) *User behavior*. Work in this category mainly focuses on identifying influential users in the microblogging service [20, 9] or examining users' tweeting behavior [6, 18]. However, to the best of our knowledge, the problem of triadic closure prediction has not been systematically studied.

Our work is also related to link prediction problem, which is one of the core tasks in social networks. However, unlike link prediction problem, we only focus on triadic closures, which means we only focus on the last "link" that constitutes the closed triad.

Organization The rest of this paper is organized as follows. Section 2 introduces the data sets used in our study and our observations in the Weibo network. Section 3 presents the proposed model and describes the algorithm for solving the model; Section 4 presents the results. Finally, Section 5 concludes.

2. DATA AND OBSERVATION

2.1 Data Collection

One objective of the study is to unveil what are the fundamental factors that influence the triadic closure formation in social networks. We chose Weibo data as the basis in our study. Thus the triadic closure process is the formation of a directed triad (also referred to as directed closure process [16]). To obtain the dynamic information, we try to crawl a sub network but with dynamic updates every time stamp from Weibo. The data set was crawled in the following ways. To begin with, 100 random users were selected, and then their followees and followees' followees were collected as seed users. The crawling process produced in total $695,842$ users and $423,347,905$ following links among them, with an average of 200 out-degree per user, $364,600$ new links and $44,320$ newly formed closed triads per day. We took every day as a time stamp and updated every user's followers and followees information for each time stamp. We also crawled the profile of all users which contains name, gender, location, and posted miscroblogs. Finally, the resultant dynamic networks span from August 29th, 2010 to September 29th, 2010.

2.2 Observations

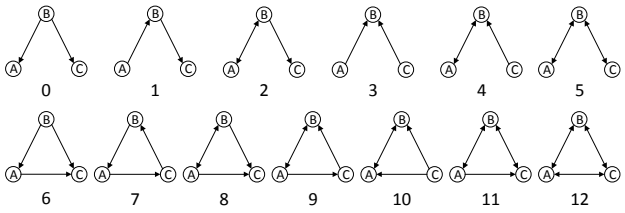
Since we are interested in the major factors that contribute to the triadic closure process, we first investigate the impact of different factors: network topology, demography and social role. For network topology, we focus on the network structure before and after the triadic closure forms. For demography, we focus on location and gender while for social role we focus on the popularity of the people within the triad, and people who are spanning structural holes. We will discuss them in details as follows.

2.2.1 Network Topology

In the directed network, there are 13 possible different three-node subgraphs [13], shown in Figure 1(a), which includes 6 open triads and 7 closed triads. Figure 1(b) shows how these open triads become triadic closures when one of the following actions happens. Furthermore, Figure 2(a) shows the probability that each open triad forms triadic closures, while 2(b) shows the probability for each open triad to form each concrete triadic closure. From Figure 2(a), we can see that open triad 5 has the highest probability to become closed, that is to say, if the existing two links between three users are both two-way relationship, the open triads are more likely to become closed, and further, one way relationship is much easier to build than two-way relationship (Seen in Figure 2(b), e.g., $P(5 \rightarrow 11) > P(5 \rightarrow 12)$).

2.2.2 Demography

Location From user profile, we can obtain the location information (province and city that the user comes from). We test whether user's location will influence the closure of a triad. Though, intuitively people from the same place may tend to follow with each other, surprisingly, we found that the probability that three persons from the same province is merely 0.0053% higher than random cases. We further consider the city level(as shown in Figure 3(a)). The result is similar, the probability that three persons from the



(a) Network motifs. The number below each motif is the index of the triad. Triad 0 – Triad 5 are 6 open triads and Triad 6 – Triad 12 are 7 closed triads. A, B and C represent users.

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

(b) How open triad forms triadic closure. For each entry in the table, left and right numbers indicate the index of triads, the expression above the arrow indicates the new formed link in time $t+1$, e.g, $0\xrightarrow{A\to C}6$ means if at time $t+1$, A follows C, then open triad 0 becomes an isomorphous of closed triad 6.

Figure 1: Open triads, triadic closures and how open triads form triadic closures.

same city is only 0.0032% higher than random cases. It seems that thanks to the online social networks, the geographic location is no longer a limitation factor for people to know each other any more.

Gender We test whether the gender homophily will play some role on the triadic closure formation. As shown in Figure 3(b), we can see the possibility are nearly the same based on different gender combinations. we use three bit binary codes $(XXX)(X = 0 \text{ or } 1$ (0 means female and 1 means male) to represent the triad status. We notice that $P(XX0) > P(XX1)$, which means if the third user in the triad is female, it is more likely to form the closed triad, and it is always accomplished by the third person $(P(A \leftarrow C) > P(A \rightarrow C)$, the green part is larger than the blue part).

2.2.3 *Social Role*

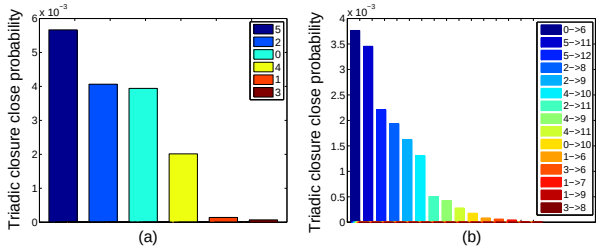


Figure 2: Network topology correlation. Y-axis: probability that each open triad forms triadic closures. (a) The number to the color bars means the index of open triads. (b) Expressions attached to color bars represent that one open triad to form one concrete triadic closure, e.g., $0 \rightarrow 6$ represents triad 0 forms triad 6.

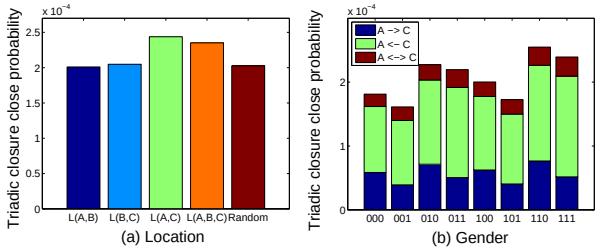


Figure 3: Demography correlation. Y-axis: probability that triadic closures form. (a) Location distribution. Expressions on X-axis means whether certain users are from the same city, e.g., $L(A,B)$ represents only A, B are in the same city. Random means users in a triad all come from different cities. (b) Gender distribution. Number on the X-axis means the gender status of the triad, 0 means female and 1 means male, e.g., 001 means A and B are female while C is male. The status of the new formed link is presented in different color, e.g., blue represents the third link is accomplished by user A, who follows user C.

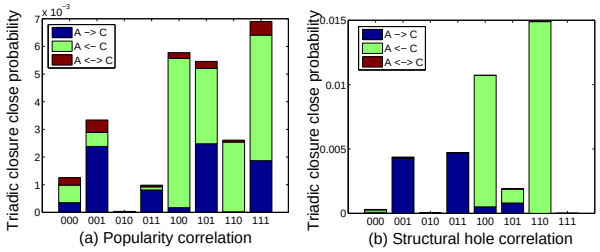


Figure 4: Social role correlation. Y-axis: probability that triadic closures form. The status of the new formed link is presented in different color, e.g., blue represents the third link is accomplished by user A, who follows user C. (a) Popularity correlation. Number on the X-axis means the popularity of the triad. 0 represents ordinary user and 1 represents popular user, e.g., 001 represent A and B are ordinary user while C is popular user. (b) Structural hole correlation. Number on the X-axis means the structural hole spanners' status of the triad. 0 means ordinary user and 1 means structural hole spanner.

Popularity For the popularity, we test if one of the three users is popular user, whether the open triad will be closed? Here we employ Pagerank [15] to estimate the user "Popular" status in the network, based on which the top 1% ranked users are defined as the "Popular" users while the rest are the ordinary ones. We use three bit binary codes $(XXX)(X = 0$ or 1) to represent the triad status: 0 means ordinary user and 1 means popular user. Figure 4(a) shows the correlation between users' popularity and the proportions of triadic closures to the total open triads. We can see from this figure, if the common neighbor of the three users is a popular user, they are the least to close the open triads. We can explain this phenomenon as below: B can be a super star or a politician figure or an official account, which have a lot of followers and much fewer followees, and play a more important role than ordinary users in the network; meanwhile ordinary users, e.g., A and C, follow them, but they are unlikely to interact with each other, so the probability to close the open triads is small under this type of cases.

Social Structural Hole We further test whether a structural hole spanner will play a role on the open triad closure process. We get the structural hole spanners from [10]. We also use three bit binary codes $(XXX)(X = 0$ or 1) to represent the triad status: 0 means ordinary user and 1 means structural hole spanner. Fig-

ure 4(b) shows the correlation between users' social structural hole properties and the proportions of triadic closures to the total open triads ($P(111) = 0$, there are no triads that contain three structural hole spanners). we can see from this figure, if B is a structural hole spanner, the open triad is less likely to become closed. This is because B hold the resources in hand, he/she is unwilling to let A and C share information, otherwise B is no longer the structural hole spanner. However, if A or C is structural hole spanner besides B, A and C are more willing to connect with each other to get more resource himself, so the open triads are more likely to be closed.

Transitivity Transitivity [19] 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 $A \rightarrow B$, $B \rightarrow C$, and $A \rightarrow C$ are present . Extending this definition, a triad is said to be transitive if all the relations it contains are transitive. For example, for A's friends's friends is A's friends as well. In Weibo, it is more likely (72%) for users to be connected in a transitive way. While some irtransitive triads also exists this is partly because two users are very likely to follow a super star, but they may not know each other.

2.2.4 *Summary*

We summarize our observations as below: (1) Location is not so important for the triadic closure process since the location is no longer a limitation factor for people to know each other any more. (2) Women are more willing to close the open triads than men do. (3) Popular users play little role on the formation of triadic closure, but popular users themselves are more willing to get closed with other popular users. (4) If structural hole spanner is the connected node which connects the other two users, he is unwilling to make the open triad closed, however, if he is not the connected node, he is happy to close the open triad to get more resources from others.

We compare the results to a similar study on Twitter about the popularity in triads [5] and find: (1) Both results demonstrate the phenomenon of "the richer gets richer", i.e., $P(1XX) > P(0XX)$, which validates the mechanism of preferential attachment in both networks (Twitter and Weibo). (2) In Twitter, popular users play an important role to form a closed triad, i.e., $P(X1X) > P(X0X)$, while in Weibo, the result is reverse. Possibly it is because Weibo provides more features to help users interact with each other, and ordinary users have more chances to connect other users. (3) The probability $P(111)$ for popular users in Weibo is much higher than that in Twitter, which implies that popular users in China have more closeness connections.

3. TRIADIC CLOSURE PREDICTION

Based on the observations in section 2, we see that the closure of an open triad not only is depending on the demography of the users involved in the triad, but also influenced by the structural position of the triad in the network. Technically, for triadic closure prediction, the challenge is how to design a unified model to combine the two pieces of information together. In this paper, we present a Triad Factor Graph (TriadFG) model for triadic closure prediction. A similar model has been also studied in [11] for reciprocal relationship prediction. However, [11] mainly focuses on investigating when a user follows back another follower, whether she will continue to follow back another follower so as to form a triad, while in this work we try to generalize the problem as a prediction task on how a closed triad is developed from an open triad.

Therefore, for a given network $G^t = \{V, E, X, Y\}$ at time t, where V is a set of nodes, $E \subset V \times V$ is a set of edges connecting those nodes, we first extract all open triads and define features for each triad. Here we use X to denote features defined for each open triad. The features can be defined based on the observations in Section 2. It can be also defined based on other statistics. Finally, we use Y to denote the set of status whether the open triads become close or not. Given this, we could construct the TriadFG model.

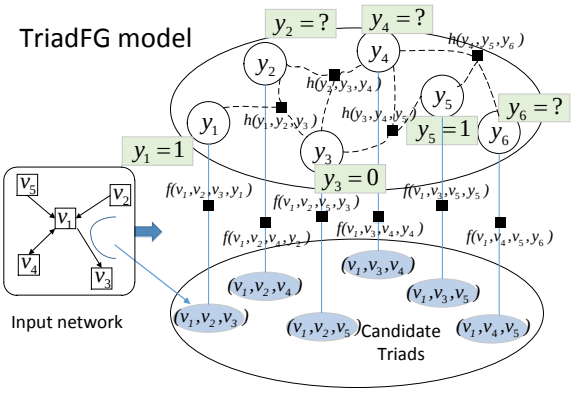


Figure 5: Graphical representation of the TriadFG model with five users in the input network. Candidate open triads are illustrated as blue ellipses in the bottom right. White circles indicate hidden variables y_i. $f(v_1, v_2, v_3)$ represents attribute factor function and $h(.)$ the correlation function among triads.

For example, Figure 5 shows a simple example of TriadFG. 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 candidate, thus we have six candidates which are illustrated as blue ellipses in the right model. All features defined over open triads are denoted as, for example, $f(v_1, v_2, v_3)$. In addition, we also consider the 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 $h(.)$ among related triads. Based on all the considerations, we construct the TriadFG (as shown in Figure 5).

To instantiate the TriadFG model, we still need to give the formal definition of the objective function and instantiate the feature definitions. Given a network at time t, i.e., $G^t = (V^t, E^t, X^t)$ 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 for all variables if there is no ambiguity. We begin with the posterior probability of $P(Y|\mathbf{X}, G)$. Directly solving the posterior probability is obviously intractable. Here, we instantiate the probabilities $P(Y|G)$ and $P(x_i|y_i)$ within Markov random field and Hammersley-Clifford theorem [4]:

$$P(Y|\mathbf{X}, G) = \frac{1}{Z} \exp\{\sum_{i=1}^{|Tr|} \sum_{j=1}^{d} \alpha_j f_j(x_{ij}, y_i) + \sum_c \sum_k \mu_k h_k(Y_{Tr_c})\} \quad (1)$$

where $|Tr|$ denotes the number of candidate (open) triads in the network, d is the number of features defined for the triads (more details for feature definition are given in Appendix), 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. For example in Figure 5, the correlation function $h(y_1, y_2, y_3)$ is related to the three triads; α_j and μ_k are parameters corresponding to the two kinds of functions $f(.)$ and $h(.)$. Finally Z is a

Table 1: Triadic closure prediction performance

Algorithm	Prec.	Rec.	F1	Accu.
SVM	0.890	0.844	0.866	0.882
Logistic	0.882	0.913	0.897	0.885
TriadFG	**0.901**	**0.953**	**0.926**	**0.931**

normalization factor to guarantee that the resultant is a valid probability.

We then define a log-likelihood objective function $\mathcal{O}_{\alpha,\mu} = \log P_{\alpha,\mu}(Y|\mathbf{X}, G)$. Learning the TriadFG model is to estimate a parameter configuration $\theta = (\{\alpha_j\}, \{\mu_k\})$ from a given historical data, that maximizes the log-likelihood objective function, i.e., $\theta = \arg\max \mathcal{O}(\theta)$. We employ a gradient descent method for model learning. Specifically, for each parameter, for example μ, we randomly assign an initial value, and then derive the gradient of each μ_k with regard to the objective function, finally update the parameter with a learning rate η. Interested readers can refer to [11] for details of the learning algorithm.

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, $Y^\star = \arg\max \mathcal{O}(Y|X, G, \theta)$. 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 finally assign each open triad with a label of the maximal probability.

4. EXPERIMENTS

We use the data set described in Section 2 in our experiments and briefly summarize the major results here.

Experiment Setup To quantitatively evaluate the effectiveness of the proposed model and the methods for comparison, we divide the network into two subsets by using the first two-third of the data as training and the rest as test set. Our goal is to predict whether an open triad in the training set will become close in the test set.

We compare TriadFG with two alternative baselines.

SVM It 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 It is similar to the SVM method. The only difference is that it uses logistic regression model as the classification model.

For SVM and Logistic, we use Weka[3]. We evaluate the performance of different approaches in terms of Precision(Prec.), Recall(Rec.), F1-Measure(F1) and Accuracy(Accu.). All algorithms are implemented in C++, and all experiments are performed on PC running Windows 7 with AMD Opteron(TM) Processor $6276(2.3GHz)$ and $4GB$ memory.

Prediction Performance We now demonstrate the performance results for the different methods in Table 1. It can be seen that our proposed TriadFG model outperforms the other two comparison methods. In terms of F1-Measure, TriadFG achieves a $+6.99\%$ improvement compared with the SVM, and $+3.3\%$ with Logistic. Meanwhile, TriadFG also makes some progress on recall, it is partly bacause TriadFG can detect some cases by lerveraging the transitive correlation and homophily correlation.

Factor Contribution Analysis In this section, we examine the contribution of four different factor functions: Demography(D), Popularity(S), Network topology(N) and Structural hole spanner(H). We first rank the individual factors by respectively removing each particular factor from our model and evaluate the decrease of the prediction performance. Thus, a larger decrease means a higher predictive power. And then remove them one by one in reversing order of their prediction power. We denote TriadFG-D as removing Demography and TriadFG-DS as further removing social role, finally removing the network structure denoted as TriadFG-DSN. As shown in Table 2, we can observe significant performance decrease when ignoring social role information while slightly drop when ignoring Demography information.

Table 2: Factor contribution analysis.

Method	TriadFG	TriadFG-D	TriadFG-DS	TriadFG-DSN
F1-Measure	0.927	0.835	0.769	0.683

Qualitative Case Study Now we present a case study to demonstrate the effectiveness of the proposed model. Figure 6 shows an example generated from our experiments. It is a portion of the Weibo network among our dataset. User A and B are popular users, and A is also a structural hole spanner. The numbers associated with each user are respectively the number of followers and that of followings. If the label is red, then it means the user is female; if blue, then male. Black arrows indicate following links created before. Red arrows indicate the link will form in the next time stamp. In Figure 6(b), green dash lines indicate the links are predicted by SVM however will not connect in the next time stamp. The red dash lines indicate the link will form in the next time stamp, but not predicted by SVM. In Figure 6(c), the green dash lines indicate the links are predicted by our approach however will not connect in the next time stamp.

We look into specific example to study why the proposed model can outperform the comparison methods. SVM misses predicting the formation of triad $(5, 1, 6)$ and wrongly predicts the closure $(3, 1, 4)$. However, our approach correctly predicts these two triadic closures, partly because we use social correlation such as transitive correlation and homophily correlation among features in our model.

5. CONCLUSION

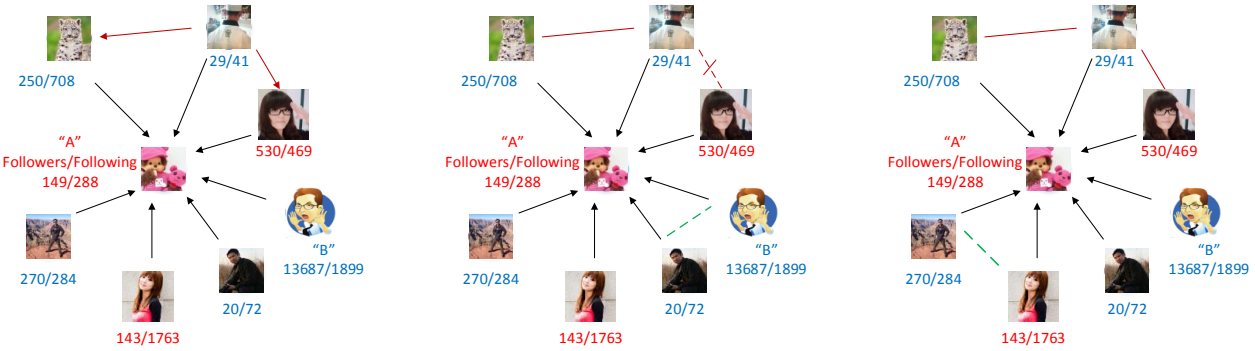
In this paper, we study an important phenomenon of triadic closure formation in dynamic social networks. Employing a large microblogging network (Weibo) as the source in our study, we formally define the problem and systematically study it. We propose a probabilistic factor model for modeling and predicting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

Acknowledgements Hong Huang is supported by the China Scholarship Council. Jie Tang is supported by the Natural Science Foundation of China (No. 61222212, No. 61073073, No. 61170061), National Basic Research Program of China (No. 2011CB302302), and a research fund supported by Huawei Inc. Xiaoming Fu is partly supported by Lindemann Foundation.

6. REFERENCES

[1] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD'06*, pages 44–54, 2006.

[2] P. Grindrod, D. J. Higham, and M. C. Parsons. Bistability through triadic closure. *Internet Mathematics*, pages 402–423, 2012.



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

Figure 6: Case study. Portion of the Weibo network among our dataset. User A and B are popular users, and A is also a structural hole spanner. The number associated with each user are respectively the number of followers and that of followings. If the label is red, then it means the user is female; if blue, then male. Black arrows indicate following links created before. Red arrows indicate the link will form in the next time stamp. In (b), green dash lines indicate the links are predicted by SVM however will not connect in the next time stamp. The red dash line indicates the link will form in the next time stamp, but not predicted by SVM. (c), the green dash lines indicate the links are predicted by our approach however will not connect in the next time stamp.

[3] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

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

[5] J. Hopcroft, T. Lou, and J. Tang. Who will follow you back? reciprocal relationship prediction. In *CIKM'11*, pages 1137–1146, 2011.

[6] B. Huberman, D. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. 2008.

[7] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *WebKDD'07*, pages 56–65, 2007.

[8] G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311(5757):88–90, 2006.

[9] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW'10*, pages 591–600, 2010.

[10] T. Lou and J. Tang. Mining structural hole spanners through information diffusion in social networks. In *WWW'13*, pages 837–848, 2013.

[11] T. Lou, J. Tang, J. Hopcroft, Z. Fang, and X. Ding. Learning to predict reciprocity and triadic closure in social networks. *TKDD*, 2013.

[12] M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *SIGMOD'10*, pages 1155–1158, 2010.

[13] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, pages 824–827, 2002.

[14] T. Opsahl. Triadic closure in two-mode networks: Redefining the global and local clustering coefficients. *Social Networks*, 2011.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: bringing order to the web. 1999.

[16] D. M. Romero and J. M. Kleinberg. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In *ICWSM'10*, 2010.

[17] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW'10*, pages 851–860, 2010.

[18] J. Tang, S. Wu, and J. Sun. Confluence: Conformity influence in large social networks. In *KDD'13*, pages 347–355, 2013.

[19] S. Wasserman. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[20] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM'10*, pages 261–270, 2010.

7. APPENDIX

This section depicts how we define factor functions in our experiments. In total, we define 20 features of three categories: Network topology, Demography and Social role.

Network topology For the open triads, we have six different types. Then we define six features based on these types to see which types they belong to.

Demography We have Gender and Location user profile. Based on these two values, we define two features: whether the three users in one triad are from the same location and whether they are of the same gender.

Social role Here we consider popularity and social structural hole spanners. We define six different features for both of them to see how many typical users (popular user or structural hole spanner) are in one triad and whether one is the typical user.