Triadic Closure and Its Influence in Social Networks

– A Microscopic View of Network Structural Dynamics

Dissertation
for the award of the degree
"Doctor of Philosophy" Ph.D. Division
of Mathematics and Natural Sciences
of the Georg-August-Universität Göttingen

within the doctoral program in Computer Science (PCS)
of the Georg-August University School of Science (GAUSS)

submitted by
Hong Huang

from Jiangxi, China
Göttingen, 2016

Thesis Committee

Prof. Dr. Xiaoming Fu
Institute of Computer Science, University of Göttingen
Dr. Jan Nagler
Max Planck Institute for Dynamics and Self-Organization
Dr. Wenzhong Li
Department of Computer Science, Nanjing University, China

Members of the Examination Board / Reviewer

Prof. Dr. Xiaoming Fu
Institute of Computer Science, University of Göttingen
JProf. Dr.-Ing. Marcus Baum
Institute of Computer Science, University of Göttingen
Prof. Dr. techn. Dipl. Ing. Wolfgang Nejdl
Department for Electrical Engineering and Computer Science, Leibniz Universität Hannover

Further members of the Examination Board

Prof. Dr. Dieter Hogrefe
Institute of Computer Science, University of Göttingen
Prof. Dr. Jens Grabowski
Institute of Computer Science, University of Göttingen
Prof. Dr. Jar-Der Luo
Department of Sociology, Tsinghua University, Beijing, China

Date of the oral examination: 01. 09. 2016

I would like to dedicate this thesis to my parents, my brother and my sister . . .

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

submitted by
Hong Huang
September 2016

Acknowledgements

First and foremost, I would like to thank my supervisor Prof. Dr. Xiaoming Fu. He is open minded and tolerant. He always tried his best to offer us great opportunities to achieve better us. I really appreciate his continuous valuable support, guidance, encouragement throughout my Ph.D study.

I am deeply grateful to Prof. Dr. Jie Tang, who also guided me in the last four years. Without his help, I cannot accomplish my Ph.D study. His deep insight and dedicated attitude helped shape my research and life.

I want to thank my co-supervisors Dr. Jan Nagler and Prof. Dr. Wenzhong Li, who gave me useful advice to conduct better research.

I would also express my gratitude to Yuxiao Dong, Lu Liu, Sen Wu, Jar-Der Luo, I learnt and benefited hugely from our communications and collaboration.

I really appreciate Prof. Ronald Burt, Prof. Jie Tang, Prof. Jar-Der Luo, Prof. Xin Wang and Prof. Shanlu Lu, Prof. Wenzhong Li for their host during my visiting.

My sincere thanks go to Professor Marcus Baum, Professor Wolfgang Nejdl, Professor Jens Grabowski, Professor Dieter Hogrefe and Professor Jar-Der Luo for being my committee members.

I wish to thank all the colleagues and visitors in Computer Networks (NET) Group. I appreciate the wonderful discussions and suggestions from them.

Heartfelt thanks go to China Scholarship Council for its great financial support.

Last but definitely not least, I owe a great deal to my parents, my brother and my sister. Their unconditional and endless love and support is always my motivation to go forward.

Abstract

Social triad—a group of three people—is one of the simplest and most fundamental social groups, which serves as the basis of social network analysis. Triadic closure, a closing process of an open triad, is a useful principle and model to understand and predict network evolution and community growth, which has been widely used in web mining and solving social issues like political movements, professional organizations and religious denominations. Extensive network and social theories have been developed to understand the triadic structure, for example, triadic closure facilitates cooperative behavior and "friend of my friends are my friends". However, over the course of a triadic closure—the transition from open triads to closed triads are much less well understood. Furthermore, the interaction dynamics in networks, particularly in a triad is still unclear.

In order to fill the gap in triadic closure studies, in this thesis, we trace the whole process during and after triadic closure. Starting from open triads, we study the problem of group formation in online social networks and try to understand how closed triads are formed from open triads in dynamic networks. Secondly, we focus on triadic closure's influence on networks, especially its influence on tie strength dynamics of social relations. We investigate whether the new established third link will affect the tie strength dynamics of open triads after triadic closure.

Employing a large microblogging network as the source in our study, we first focus on open triads closing process. By investigating the impact of different factors from three aspects: user demographics, network characteristics, and social perspectives, we find some interesting phenomena including: male, celebrity and gregarious users are more inclined to closing triads; structural hole spanners are eager to close open triads for more social resources, but they are also reluctant to have two disconnected friends to be linked together.

Then, we examine triadic closure and its influence – tie strength dynamics of triads after closure, especially whether and how the formation of the third tie among three users in a triad affects the strengths of the existing two ties

using two dynamic networks from Weibo and mobile communication. We find that the closure of 80% social triads weakens the strength of the first two ties. Surprisingly, we discover that although males are easier to get closed, the decrease in tie strength among three males is more sharp than that among females, and celebrities are more willing to form triadic closure. However, the tie strengths between celebrities are more likely to be weakened as the closure of a triad than those between ordinary people. We also demonstrate that while strong ties result in weakened relationships in open triads, they can promote the stronger ties in closed social triads.

Further, we formalize a prediction problem to predict triadic closure. We propose a probabilistic graphical method to solve the triadic closure prediction problem by incorporating user demographics, network topology, and social information. With better instantiating attribute factors, we also extended our model with kernel density estimates. Unlike triadic closure prediction, the prediction for triadic tie strength dynamics is far more complicated when time dynamics is took into account. We further propose a dynamic probabilistic graphical to solve the problem of triadic tie strength dynamics prediction with the consideration of user demographics and temporal as well as structural correlations.

Extensive experiments demonstrate that our proposed model offers a greater predictability for both prediction tasks. We demonstrate that our methodology offers a better-than-82% potential predictability for inferring the dynamics status of social triads in both networks, and the leveraging of the kernel density estimate together with structural correlations enables our models to outperform baselines by up to 30% in terms of F1-score.

The triadic closure and its influence studied in this thesis will be a good guide to practical applications, like friend recommendation and new friend invitation for online microblogging services.

Table of contents

Chapter 1

Introduction

1.1 Motivation

The increasing popularity of social networks, especially microblogging service encourages more and more users to participate in various online activities, which are becoming a bridge that connects 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, containing 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 unprecedented opportunities to study interaction patterns among users, and to understand the generative mechanisms of versatile networks, which was previously difficult to explore, due to the unavailability of data. A better understanding of user behavior and underlying network patterns can enable an OSN provider to attract and maintain more users, and thus increase its profit. While for individuals, a better understanding of their networks can help them share and collect reliable information in a more effective and efficient manner.

The interactions between individuals form the structural backbone of human societies, which manifest as networks. From network perspective, individuals matter in the ways that their interactions and groupings activate the emergence of new phenomena at larger and societal levels. 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 [4].

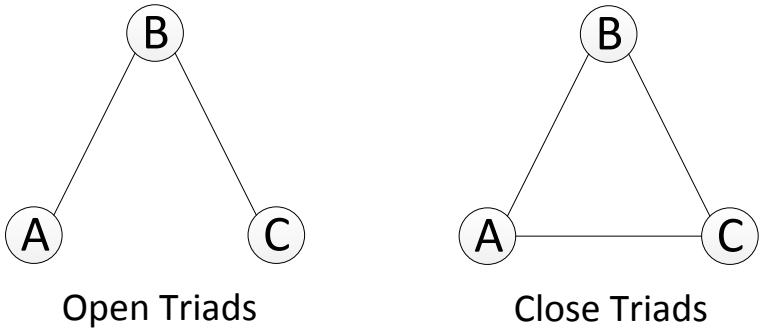


Fig. 1.1 Undirected open triad and close triad.

A triad is a group of three people, which is one of the simplest human groups. Roughly speaking, there exist two types of triads: *open triads* and *closed 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. In undirected networks, the structure for triad is very simple. There are only one open triad and one close triad, as shown in Figure 1.1. While in directed networks, the situations are much more complicated. Figure 1.2 shows all the possible examples of open and closed triads in directed networks when each isomorphous triad is only considered once.

One interesting question is how a closed triad develops from an open triad. The problem is referred to as the *triadic closure process*, which is a fundamental mechanism in the formation and evolution of dynamic networks [21]. 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 [29, 75].

Moreover, as networks evolving, the strength of social ties is not static over time. Some ties may become "strong ties" at first and then weaken over time, while other social ties appear as "weak ties" and become stronger later. The strength dynamics become even more complicated when we consider the interpersonal interactions. For example, after triadic closure, the degree to which the formation of the third tie in a triad affects the strength of the existing two ties.

A significant amount of work has been devoted to investigating triadic relationships in social networks for decades. Simmel pioneered the study of “triad" and suggested that a social triad is fundamentally different from a dyad as interaction between members decreases, intimacy declines, strength

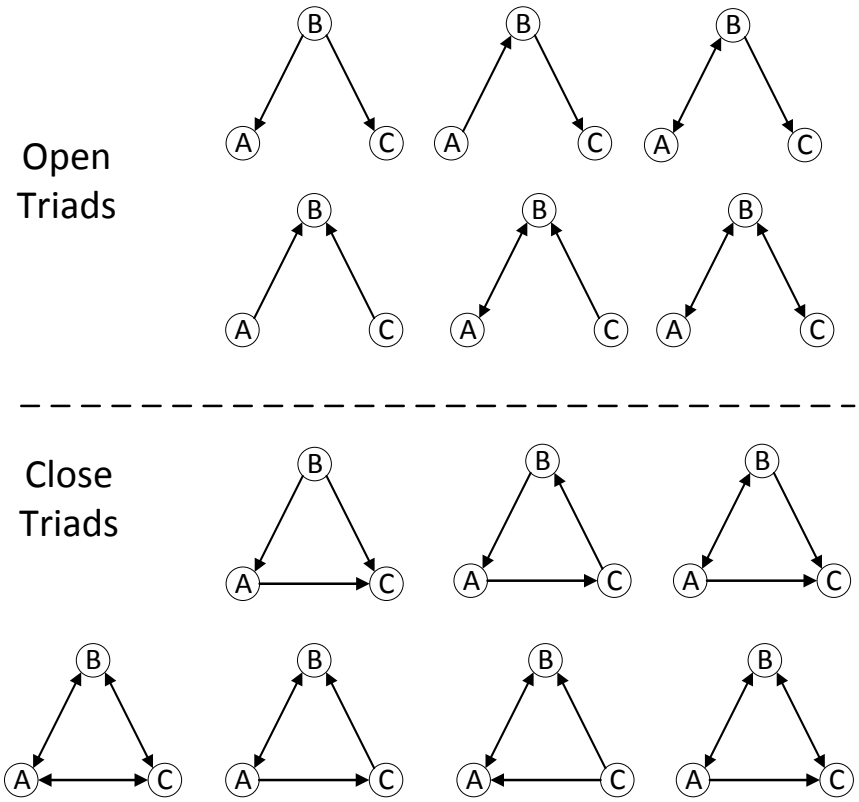


Fig. 1.2 Directed open triad and close triad.

and stability increase [115]. 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 [51] – and found that friends of friends tend to become friends themselves [51, 124]. 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 [30] on the World Wide Web. Existing work [75, 137, 108, 42] proposed network generative models based on triadic closure principles. Milo et al. [92][93] defined the recurring significant patterns of interconnections as "network motifs" and emphasized their importance in uncovering the basic building blocks of most networks. But these studies focused only on uses of the triadic closure process, without clarifying the underlying principles of triadic closure. Romero et al. [105] 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. [86] investigated how a reciprocal link was developed from a parasocial relationship and how the relationships developed into triadic closure in a Twitter dataset. However, these studies

only examined some special cases of the triadic closure process. A commonly observed behavior in a triad is that two of the members will tend to unite against the other one, which is known as *two-against-one* phenomenon [18]. Heider developed the balance theory [50] in social triads that explains the famous proverbs – "A friend of my friend is my friend" and "The enemy of my enemy is my friend." Davis et al. proposed a status theory [22] that provides an organizing principle for directed networks of signed links. They addressed how the interplay between signed (i.e., positive and negative) relationships affected the structure of networks. However, nowadays, in most of our online social networks (e.g., Twitter, Facebook, Weibo), there is no evidence to tell the sign of relationships in the networks where such theories, like balance theory and status theory would not be applied any more. Furthermore, the process of triadic closure has been empirically demonstrated to be relevant for characterizing both social ties at a micro level [116], and scaling laws at a macro level in social and information networks [75, 67].

Although the interesting and promising discoveries that have been made in the field of social triads, many open challenges require further methodological developments. The underlying mechanism of triad closure is still unclear and little has been studied concerning what happens after triadic closure, especially the dynamics of tie strength within a triad. How people are embedded and interact within a closed social triad over time? Essentially, previous attempts are limited by merely focusing on the triadic applications, and ignoring the underlying mechanism that govern triadic closure formation and dynamics of triadic relationships over time in unsigned social networks. In other words, after triadic closure, the interaction dynamics of original relationships is still unclear. Further complications arise due to the complexity of scrutinizing various factors that drive the interaction dynamics of social triads, such as the demographics of a triad's three users, their tie strengths, and the formation order of the links in a triad. Moreover, there lacks of a basic understanding of the predictability of triadic closure and triadic tie strength dynamics in social networks.

In light of these limitations, we put forward two fundamental problems aiming to uncover underlying mechanism that governs triad's evolution. Figure 1.3 shows an illustrative example of the evolution of triadic relationships in social networks. We observe that an open triad $\mathscr{O}_{ABC}$ becomes a closed one $\mathscr{T}_{ABC}$ after the formation of a new link e_{AC} at time t. Our first goal is to understand how user demographics, network characteristics, and social properties influence the formation of triadic closure. Our second goal is to

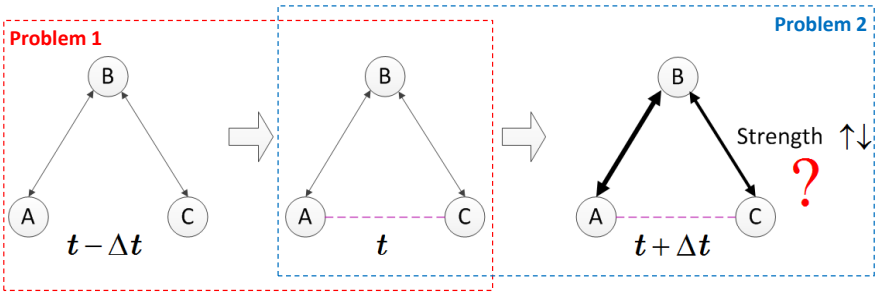


Fig. 1.3 An illustrative example of the evolution of triadic relationships in social networks.

trace the dynamics of triadic interactions within these three users for a short period Δt before and after t.

Moreover, how can we design a unified model for predicting the formation of triadic closure and triadic tie strength dynamics? In particular, how can we quantify correlation (similarity) between triads? Specifically, we aim to understand 1) triadic closure formation 2) triadic tie strength dynamics after the formation of the third link 3) the predictability of triadic closure and the extent to which 4) the dynamics of triadic relationships can be predicted from social networks.

One straightforward application of our work is friend recommendation and new friend invitation in online social networks. Examples are shown in Figure 1.4.Other potential applications include group formation [4, 105], social search, and user behavior modeling.

Fig. 1.4 Friend recommendation in social networks.

1.2 Contributions

In this thesis, we have done a comprehensive study on triadic closure. By applying social theories to social network data, we aim to solve two basic tasks in social networks: influential user mining (Chapter 3, Chapter 4) and community detection (Chapter 3, Chapter 4) under user-based tasks, link prediction (Chapter 5.2) and tie strength prediction (Chapter 5.3) under relation-based tasks. The general graphical structural in this thesis shown in Figure 1.5.

By employing datasets from Weibo[1], one of the large microblogging networks and a mobile communication service, as the basis of our study, we first examine patterns in triadic closure process in order to better understand factors that trigger the formation of groups among people. Then we trace the dynamics of social interactions within social triads and systematically investigate the dynamics of tie strength in social triads over time. Our contributions are multifold:

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

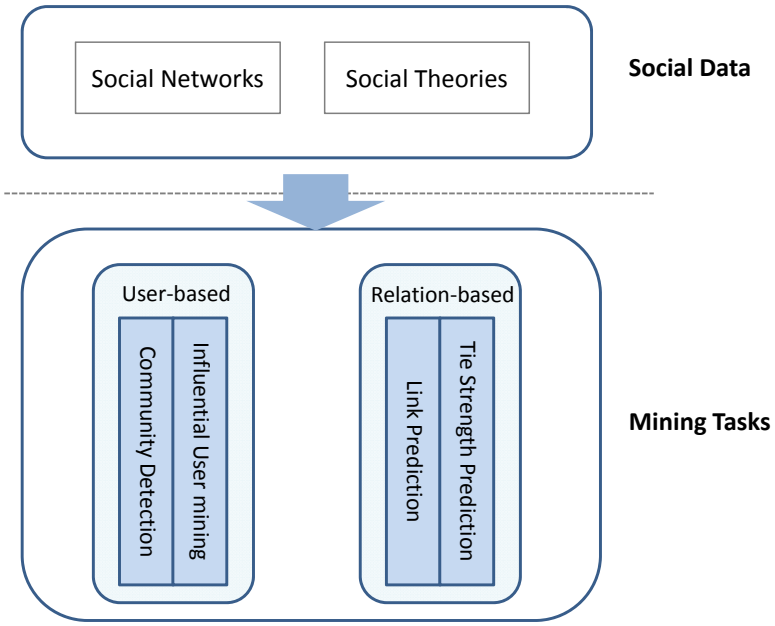


Fig. 1.5 Main mining tasks in the thesis.

- 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.
- We find that in around 80% of closed social triads, the strengths of the first two ties become weakened, as measured by the interaction dynamics in both online social media and mobile communication networks. We also discover that the stronger(as measured by interaction frequency and reciprocity) the third tie is, the less likely the first two ties are weakened; while the stronger the first two ties are, the more likely they are weakened. Surprisingly, we find that the decrease in tie strength among three males is more sharply than that among females. Finally, we observe that in social media, tie strengths between celebrities are more likely to be weakened as the closure of a triad than those between ordinary people.
- 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 tri-

ads to predict triadic closure. We then formalize the question of whether tie strength of a triad after closure will become weakened as a triadic tie strength dynamics prediction problem. The prediction task is to infer whether the formation of the third link in a given triad will, within a predefined timeframe, make the interactions of the other two links infrequent. To solve this problem, we propose a triadic tie strength dynamics (TRIST) model — a kernel density estimation (KDE)-based factor graph. As a graphical model, TRIST incorporates not only attribute features but also structural features into a unified framework. Another advantage of the model comes from kernel density estimation, which smoothly models discrete attribute features. The TRIST-ST model is a reduced version of the TRIST model, which utilizes only attribute features by kernel density estimation, and ignores structural features.

- Compared with alternative methods based on SVM and Logistic Regression, the presented model TriadFGachieves significant improvement (+7.43%, $p \ll 0.01$) in triadic closure prediction. As for triadic tie strength tie prediction problem, our experimental results on both types of networks demonstrate that by using the same set of attribute features with logistic regression, SVM, decision trees, and naïve Bayes, the TRIST-ST model improves the prediction performance by up to 10% over the benchmarks due to the leverage of kernel density estimation. By also leveraging structural features, the proposed TRIST offers a greater-than-82% potential predictability for triadic tie strength dynamics, and outperforms alternative methods by up to 30%, in terms of F1-score.
- 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.

1.3 Thesis Structure

This thesis is organized as follows.

Chapter 1 provides an overview of the thesis: introducing the triadic closure prediction problem and triadic tie strength problem, stating our research methodology and contributions.

Chapter 2 presents the literature review and related social theories.

Chapter 3 demonstrates our observation for triadic closure from three dimensions: user demographics, network characteristics and social perspective with data from Weibo.

Chapter 4 shows our observation for triadic tie strength dynamics from three dimensions: social ties, user demographics and temporal effects with data from microblogging services and mobile networking.

Chapter 5 proposes two effective models to predict the triadic closure and the dynamics of triadic tie strength.

Chapter 6 shows experiment settings and results for triadic closure and triadic tie strength dynamics prediction.

Chapter 7 concludes the thesis and discusses our future work.

Chapter 2

Background

2.1 Literature Review

In this Chapter, we would first describe the existing work in multiple categories, and then revisit some social theories that will used in this thesis.

2.1.1 Social Network Study

In general, there are three types of objects in social network data – users, social relations and user generated content, which allows us to roughly classify microblogging studies into three groups based on the studied objects – user-based studies, relation-based studies and content-based studies [119].

For individuals, a better understanding of their networks can help them share and collect reliable information in a more effective and efficient manner. While for microblogging services, a better understanding of their customers can help them provide better services and gain more profit. User related studies include identifying special users and user community detection.

A significant amount of work has been devoted to identifying special users [2, 20, 73, 127]. Aral and Walker identified influential users in Facebook and found that younger users were more susceptible to influence than older users, men were more influential than women, women influenced men more than they influenced other women, and married individuals were the least susceptible to influence in the decision to adopt the product offered [2]. Java et al. studied the topological and geographical properties of the Twitter network. Their findings verified the homophily phenomenon – that users with similar intentions connected with each other [56]. Kwak et al. conducted a similar study on the entire Twittersphere and observed some notable properties of Twitter,

such as a non-power-law follower distribution, a short effective diameter, and low reciprocity, which deviated from known characteristics of human social networks[73].

Besides studies on mining influential users, a large body of work has focused on user classification and community detection [79, 103, 117, 23]. Community detection is used to discover groups in a network where individuals' group memberships are not explicitly given [119], which can benefit many social media mining tasks such as social targeting and personalization [119]. Mislove et al. found that users were often friends with those who shared their attributes, and communities formed in the network around users who shared certain attributes [94].

Relation-Related tasks focus on mining relations among users and aim to reveal a finegrained and comprehensive view of social relations [119]. Key problems include link prediction and tie strength prediction and so on. Since they are the main tasks in this thesis, we will introduce related work in the later section.

A plethora of techniques have been developed for various content mining tasks such as user behavior analysis and topics discovering [55, 118, 58]. Benevenuto et. al. [11] examined how frequently people connected to OSN sites and for how long. They developed a analysis strategy to characterize user activity in OSNs. Maia et. al. proposed a methodology for characterizing and identifying user behaviors in online social networks [87]. There are also studies comparing different user behaviors on different microblogging services, like Twitter and Weibo [35]. Sakaki et al. [107] proposed to utilize the real-time nature of Twitter to detect a target event; while Mathioudakis and Koudas [90] presented a system, TwitterMonitor, to detect emerging topics in Twitter content.

However, to the best of our knowledge, the problem of triadic closure prediction has not been systematically studied.

2.1.2 Triadic Closure Study

Triad/Triangle is one of the fundamental subgraphs to study network structures in order to trace some hot social science issues like political movements, professional organizations and religious denominations and wide applications in social network analysis, anomaly detection and web mining. Various literatures have shown researcher's great interests in Triads.

The study of “triad” was pioneered by Simmel in 1908 [115]. He suggested that dyad, a group of two people, was the simplest group while triad, a group of three people, was quite different from dyad. In a dyad, if one person withdraws, the group can no longer exist. For example, divorce, the brokage of a marriage, which means the end of the relationship between a couple. In a triad, however, the dynamics is quite different. If one person withdraws, the group still lives on. The relationship also differs from dyad, which not only contains the direct relationship from "dyad", but also has an indirected relationship from common friends. Socialogical studies mainly focus on a single triad, like a family – mother, father and a child. Various social phenomena come out from triads, such as two-against-one phenomenon, which suggests that in a closed triad, two of its three members have the tendency to unite against the other one [18].

Since then, sociologists have worked out several profound theories on triadic relationships in social networks. Heider [50] developed the theory of social balance, in which the balance state was reached when there were three positive relationships, or two negatives with one positive, in a social triad. Essentially, the balance theory explains the real-world social phenomenon, that is “A friend of my friend is my friend” and “The enemy of my enemy is my friend.” Krackhardt and Handcock [70] applied this theory to explain the evolution of triangle closures. Davis et al. [22] took the theory of social status in directed networks. Status theory posits that by reversing the direction and flipping the sign to positive of each negative edge in a triad, a social triad complies with status theory if the resulting triad is acyclic. Recently, Leskovec et al. [78] suggested an alternative theory of status that provided a different organizing principle for signed networks. The social theories developed on triad structure successfully characterize the nature of social behaviors among three people in social networks. However, previous studies are limited by explaining the phenomena of triadic relationships in a static way or relying on the sign of relationships. In this thesis, to the best of our knowledge, we are the first to propose to examine the dynamics status of triadic relationships among three people over time without the sign of relationships.

In computer science, Backstrom et al. pointed out that the density of triangles, was a good indicator to show community growth. The community with higher ratio of closed to open triads, it was unlikely to grow [4]. Welser et al. suggested that the number of triangles of a user could be examined to identify a social role of the user in the network [126]. Becchetti et al. also discovered the different patterns that spam and non-spam hosts behaved in the web, and the triangle counting was a good feature to assess the user-provided

content [9]. Eckmann and Moses found that triangles tended to aggregate to generate groups with various kinds of size, thus to uncover hidden thematic layers [30].

Triad census, an enumeration of all triads, was one of the well known methods that utilizes such subgraphs [124, 34]. In graph theory, one useful measure – clustering coefficient [125] is calculated based on triad census, which measures the degree to which nodes tend to cluster together. High clustering coefficients imply a high proportion of triads (triangles) in the network. It has been pointed out that there is a close relationship between a high density of triads and the existence of community structure, especially in social networks, where the density of triads is remarkably high [97, 96, 120, 33]. The popularity of triads counting have also triggered the prosperous triangle counting algorithms [7, 57, 8, 9, 13, 60, 64, 101, 84].

In bioinformatics, motif, the recurrent subgraphs are always used to investigate large networks at the smallest scale [88, 121, 31]. Unlike subgraphs, motifs are aimed to characterize the network by the difference between the network structure and a random network with the same size and degree distribution, which means that recurrent patterns occur much more frequently than in randomized networks. Milo et al. [93, 92] defined 13 types of three-node connected subgraphs as "motifs" and emphasized their importance in uncovering the basic building blocks of most networks.

Moreover, much work has demonstrated that triadic closure can be identified as one of the fundamental dynamical principles in network formation and evolution [75, 81, 67]. 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, can be well used to model network evolution. Klimek et al.[67] and Li et al. [81] both declared that triadic closure could be identified as one of the fundamental dynamic principles in social multiplex network formation/evolution. [26, 27, 75, 137, 108, 42, 131] also provided some triadic-closure-based network generation models. Wu et al. [129] proposed a triadic closure based model to study the evolution of scientific citation networks. The network generated from triadic closure exhibited scaling laws for several structural characteristics [67], emergence of community structure, together with fat-tailed distributions of node degree and high clustering coefficients [12]. In addition, triadic closure can benefit many applications in social networks,

such as characterizing tie strength [116], influence diffusion [134], and spam detection [8].

Our study distinguishes from the previous ones in that we focus on the dynamics of triadic relationships over time after triadic closure, and the transitions from open triads to closed ones.

However, the triadic closure process itself is less well studied [68, 138, 54, 53]. Romero and Kleinberg [105] studied the problem of triadic closure and developed a methodology based on preferential attachment for studying the directed triadic closure process in directed networks. Zignani et al.[138] studied the triadic closure problem on undirected networks like Facebook and Renren. Lou et al. [86] investigated how a reciprocal link was developed from a parasocial relationship, and how the relationships further developed into triadic closure, in a Twitter dataset. Fang and Tang [32] recovered the formation process of a closed social triad in social networks. Doroud et al. examined the evolution of the triad to verify network properties in an efficient and inexpensive manner [28]. As far as we know, those studies only focus on certain triads, and none of these works systematically studied triadic closure formation and prediction in real large-scale directed networks.

2.1.3 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 of users are, the more likely they are to be linked. Various similarity measures of users are considered, such as preferential attachment [95], and the Katz measure [65]. Lichtenwalter et al. presented a flow-based method for link prediction [83]. A survey of unsupervised link prediction research can be found in [82, 37].

There are also a number of works that employ supervised approaches to predict links in social networks, such as [83, 5, 77]. Backstrom et. al. proposed a supervised random walk algorithm to estimate the strength of social links [5]. Leskovec et. al. employed a logistic regression model to predict positive and negative links in online social networks [77].

However, unlike link prediction studies, we focus mainly on triadic closure, which means we target at 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 integrate social theories into the semi-supervised learning model.

These problems are not well addressed in the literature. By measuring triad's transition during network evolution, Juszczyszyn et al. defined the triad transition matrix with probabilities of transitions between triads and showed how it can help to discover and quantify the dynamic patterns of network evolution and furthermore to predict link [61]. Golder and Yardi leveraged two structural characteristics – transitivity and mutuality to predict tie formation [41]. However, these works only used limited information from triangles to predict new established link, without considering other attributes of nodes.

2.1.4 Tie Strength Prediction

The low cost of link formation in social networks like Twitter and Weibo can lead to various relationship strengths (e.g., acquaintances, friends and mixed) [130]. As networks evolve, the strength of social ties is not constant over time. Some ties may become "strong ties" at first and then weaken over time, while other social ties begin as "weak ties" and become stronger. Users with stronger strength are likely to share greater similarity than those with weak strength; therefore with a better understanding of tie strength of can help social networks sites better serve their customers, where the problem of tie strength prediction arise, and the dynamics of social relationships and communities have attracted increasing attention [69, 4, 6, 63, 62].

Many researchers have adopted tie strength as an analytic framework for studying individuals and organizations [43, 45, 111] and paid a lot of attention to measuring the tie strength of social relations. Using survey data on friendship ties, Marsden et al. constructed and validated measures of tie strength [89]. Krackhardt validated that a "Simmelian tie" can strengthen the relationships between the individuals in social triads or groups [69]. Gilbert and Karahalios proposed a predictive model to map social media data to tie strength and distinguished them into strong and weak ties [40]. Jones et al. used online interaction data (specifically, Facebook interactions) to successfully identify real-world strong ties [59]. Xiang et al. developed an unsupervised model to estimate relationship strength from interactions [130].

On the other hand, less efforts are devoted to tie strength dynamics. Saramäki et al. found that the distribution of people that distributed their social in-

vestment over different social ties among their ego networks tended to persist over time [109]. Patil et al. presented a model to predict whether a group will remain stable or shrink over time [102]. Burke and Kraut investigated the factors that associated with tie strength dynamics. They found that tie strength increased with both one-on-one communication, such as posts, comments, and messages, and through reading friends' broadcasted content, such as status updates and photos [14]. However, most of them focus on understanding the dynamics status of social ties and communities, or measuring the tie strength in social networks, the structural factors associated with tie strength dynamics are not well addressed. Our work is the first to investigate the dynamics status of triadic relationships from a microscopic view in social networks.

2.2 Social Theories Revisited

2.2.1 Social Balance

The social balance theory was developed by Heider in the 1940s [49], and subsequently cast in graph-theoretic language by Cartwright and Harary [19].

Helder's social theory points us:

- Friend of my friend is my friend;
- Enemy of my friend is my enemy;
- Friend of my enemy is my enemy;
- Enemy of my enemy is my friend.

Social balance theory is applied in signed triad (with positive or negative links). Figure 2.1 shows such an example of social balance theory. If we look at any two people in the group in isolation, the edge between them can be labeled + or -; that is, they are either friends or enemies. Balanced triads with three positive edges exemplify the principle that "the friend of my friend is my friend," whereas those with one positive and two negative edges capture the notions that "the friend of my enemy is my enemy," "the enemy of my friend is my enemy," and "the enemy of my enemy is my friend."

Based on this reasoning, the triads with one or three + as balanced, since they are free of these sources of instability, and the triads with zero or two + as unbalanced. It was thought that unbalanced triads are sources of stress or psychological dissonance, people strive to minimize them in their personal relationships, and hence they will be less abundant in real social settings than balanced triads [29, 66].

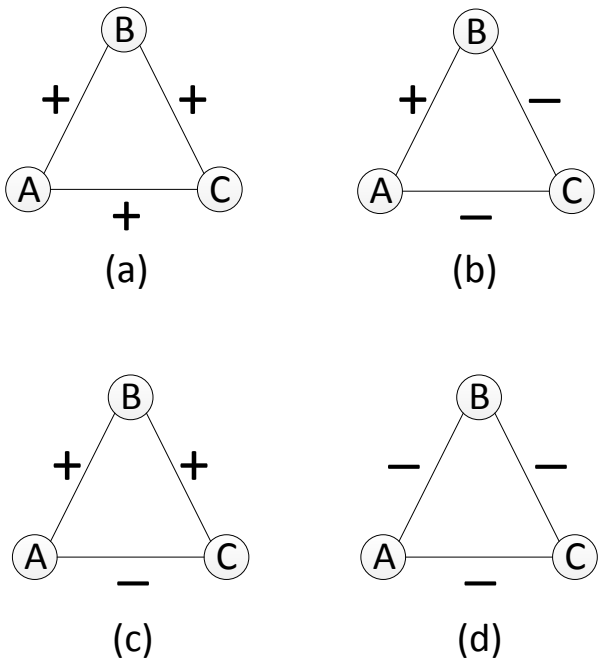


Fig. 2.1 An illustration of balance theory. (a) and (b) are balanced, while (c) and (d) are unbalanced.

2.2.2 Structural Holes

The theory of structural holes was originally developed by Ronald Stuart Burt [15], which was defined in the following way: a structural hole is a relationship of nonredundancy between two contacts. An illustration of structural holes is shown in Figure 2.2. In the figure, there is no relationship between user A and user C, which has no redundancy, so we can say that there is a structural hole between user A and user C, and user B is called a broker or structural hole spanner.

The theory suggests that individuals would hold positional advantage / disadvantages from filling the "holes" between people or groups that are otherwise disconnected [15, 85, 104]. Ron Burt showed in his studies that businessmen who maintained many structural holes had a significantly higher rate of success in a competitive marketplace [15–17]. Ahuja found that increasing structural holes had a negative effect on innovation from a longitudinal study of firms in the international chemicals industry [1].

Structural hole spanners play a key role in the information diffusion [133, 85]. Lou and Tang revealed that 25% of information diffusion was controlled

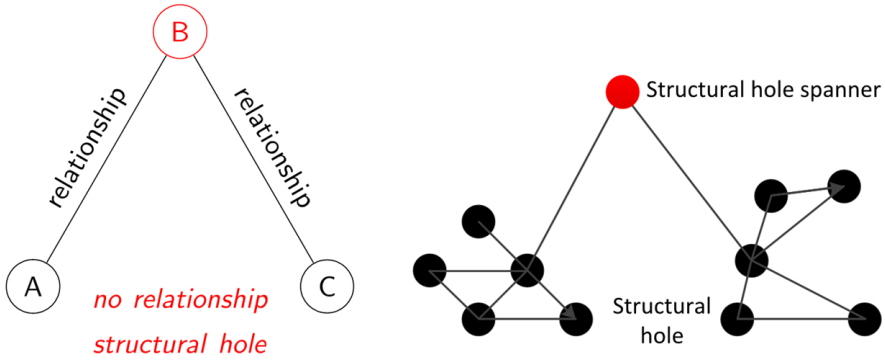


Fig. 2.2 An illustration of structural holes.

by 1% of users serving as structural hole spanners, who were bridges between otherwise disconnected communities in a network [85].

2.2.3 Strong Ties and Weak Ties

In social networks, strong ties always refer to close friends and family while weak ties refer to acquaintances and co-workers. Creating a weak tie is the first and the easiest step in any relationship. With interactions weak ties may somehow develop into strong ties.

Strong ties are very important in severe changes and uncertainty, as they constitute a base of trust that can reduce resistance and provide comfort in the face of uncertainty [71]. However, weak ties are also important resources in occupational mobility [43] and information diffusion [44], which was stated in "The strength of weak ties theory" developed by famous sociologist Mark Granovetter [43]. People can benefit more from weak ties than strong ties when they are looking for jobs because their strong ties know all the stuff they know, so there is no new information advantage [43]. The same principle holds for weak ties in information flows. More novel information flows to individuals through weak ties rather than strong ties. Because our close friends tend to move in the same circles that we do, the information they receive overlaps considerably with what we already know. Acquaintances, by contrast, know people that we do not, and thus receive more novel information [44]. Indeed, users of enterprise social networks are particularly motivated to cultivate a network of weak ties and to seek out new people [24].

As networks evolve, the strength of social ties is not constant over time. Some ties may become "strong ties" at first and then weaken over time, while other social ties begin as "weak ties" and become stronger later on.

The idea behind the strength of weak ties theory is somehow close to structural holes theory. According to weak ties theory, the stronger the tie between two people is, the more likely their contacts will overlap so that they will have common ties with the same third parties. This implies that bridging ties are a potential source of novel ideas. Therefore, Granovetter argues that strong ties are unlikely to transfer any novel information [43], which is close to structural hole theory, where structural holes is used for the separation between non-redundant contacts. Granovetter claims that whether a contact would serve as a bridge depends on a tie's strength. While Burt considers the opposite direction of causality, where a user serves as a bridge, there is no redundancy information.

Chapter 3

Triadic Closure Formation

In this chapter, we will focus on triadic closure formation on a real online social network and aim to figure out the underlying factors that trigger triadic closure [54, 53].

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 Romero and Kleinberg [105], Lou et al. [86]). 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. A screen capture of Weibo profile is shown in Figure 3.1.

Finally, the resultant dynamic networks span a period from September 29th, 2012 to October 29th, 2012. Table 4.1 gives statistics of the dataset. In addition, considered that our dataset is a sample of the whole network, we validate the crawled dataset to address sampling issues in Chapter 6.1.4.

We construct a network based on the following relationships, which is different from a co-author network or friendship network. The former is a

Fig. 3.1 A screen capture of Weibo.

Table 3.1 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

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 [73].

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 as a timestamp (denoted as $T_1, T_2, ..., T_7$). We followed the work in Lou et al. [86], where they used four days as a timestamp period to study triadic closure patterns in Twitter. In

addition, we also investigated other timestamps in Chapter 6.1.4 to see the effects of timestamps.

The number of newly formed links per timestamp period is shown in Figure 3.2(a), and the number of newly formed open triads per timestamp period is shown in Figure 3.2(b). In Figure 3.2(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. Figure 3.3 and Figure 3.4 shows the number of open and closed triads distribution in each timestamp. We can see that open triad 3 has the largest percentage - 94.9%.

In order to obtain fair and balanced observations among the limited samples, we only consider the triadic closures generated in 8 days after the open triad formed. Here we choose 8 days as a time windows, this is because: as shown in Figure 3.2(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 3.2(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 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

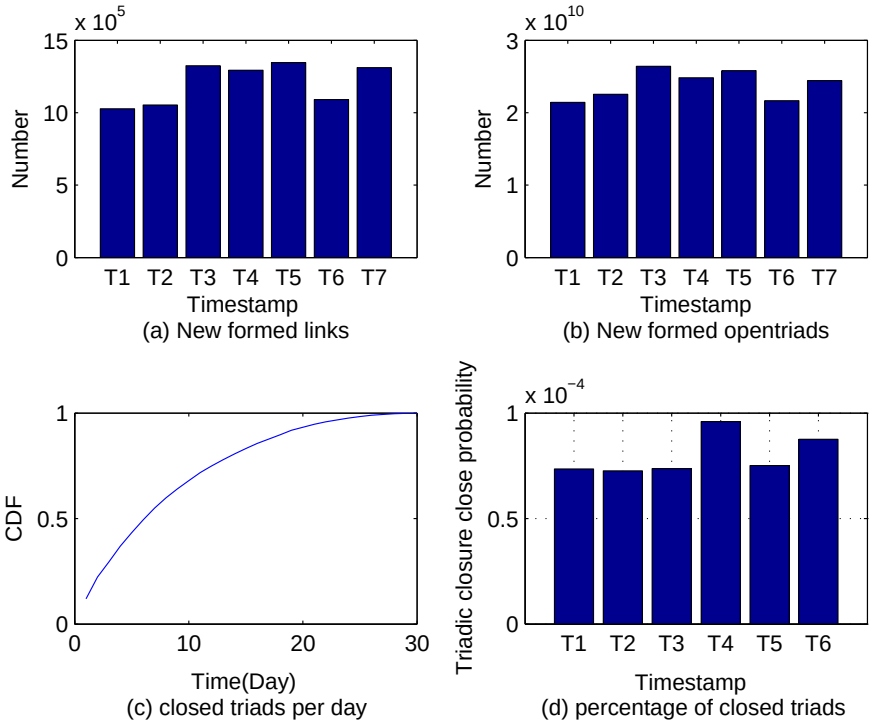


Fig. 3.2 Overall observation. (a) Y-axis: the number of new formed links in different timestamp periods. (b) Y-axis: the number of new formed open triads in different timestamp periods. (c) Y-axis: Cumulative distribution function of new formed triadic closures per day. (d) Y-axis: probability that open triads form triadic closures.

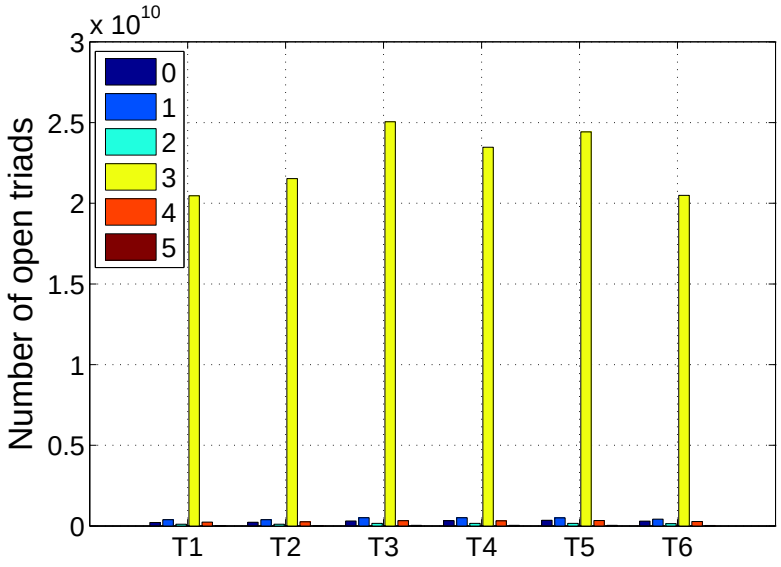


Fig. 3.3 Open triad distribution.

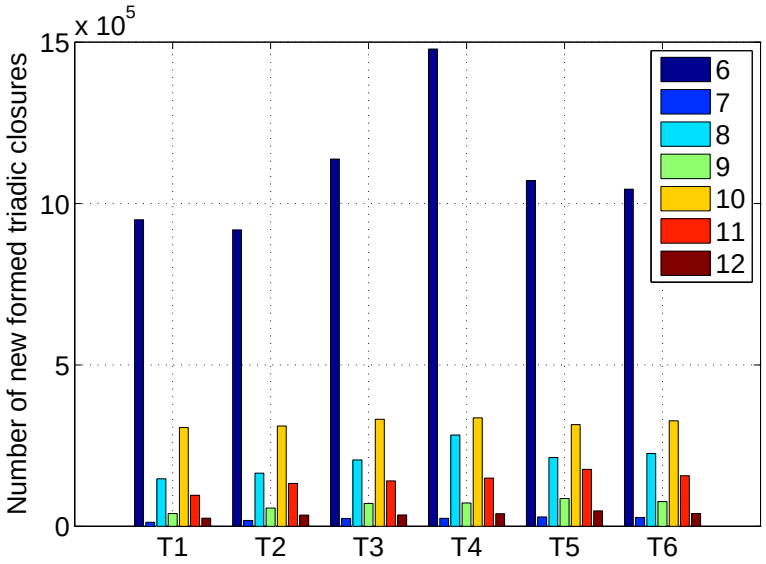


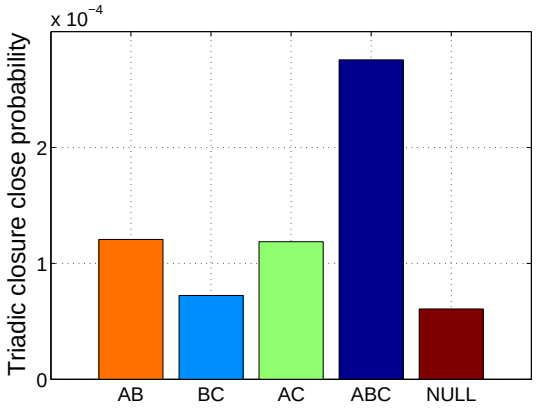
Fig. 3.4 Closed triad distribution.

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.

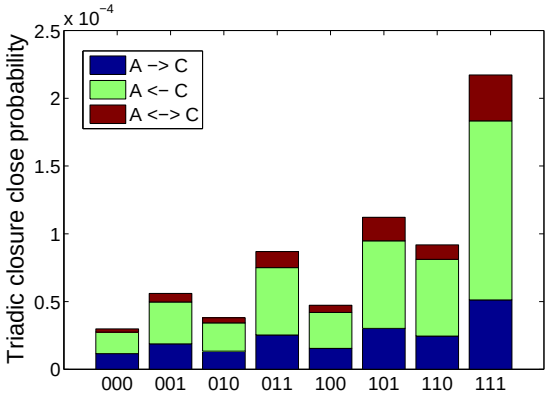
3.2.1 User Demographics

Location

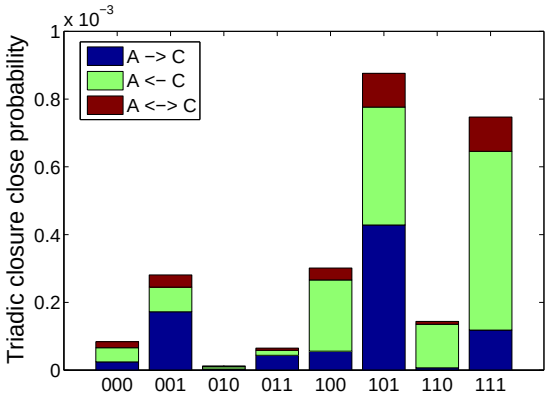
From user profiles, we can obtain location information (province and city that the user comes from). We test whether a user's location will influence the closure of a triad. We can see from Figure 3.5(a), if three users all come from the same province, the probability that the open triads will be closed 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.



(a) Location correlation



(b) Gender correlation



(c) Verified status correlation

Fig. 3.5 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.

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 3.5(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 3.5(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.

The different colors in the bar represent various occasions that who initiates a following action. For example, blue area means the first person will connect the third person. We can see from Figure 3.5(b), the third person is always more willing to close the triad.

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 3.5(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.

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 [93] as shown in Figure 1.2 – 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 3.6(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 3.6(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 3.6(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., $P_{5\to 11} > P_{5\to 12}$.

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 [100] to estimate the users' popularity in the network, based on which the top-1%-ranked users[1] 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 3.7(a)

[1]We follow the work [128] which has shown that less than 1% of Twitter users produce 50% of its content, and [86], which also uses the top-1%-ranked users to study triadic closure in Twitter.

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

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 3.7(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 each other to get more resource for themselves [106, 110, 98], 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, which is a measure of the individual's tendency to associate. 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 ones. 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 3.7(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 3.7(c), if we replace one user of "0" with a gregarious user ("1"), the probability that the triad becomes closed will double or triple.

Especially, in order to check whether gregariousness is correlated with activity, we conduct a random test. We generate a random version of users that allocate the same number of ties with gregarious users and find that at one timestamp the probability that three gregarious users close is 5.66% while the probability that random users close is 0.08%. We also test other cases and the results are shown in the Figure 3.8, which shows gregarious users are more likely to close.

Transitivity

Transitivity [124, 78] 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[2] is completed. The dataset was previously used for studying social influence in the diffusion process [135]. With this retweeting data, we study how triadic closure formation has been influenced by the retweeting behaviors.

[2]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.

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, 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 3.9 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.

3.2.4 Summary

The distribution of our observations is shown in Figure 3.10. We summarize our observations as below:

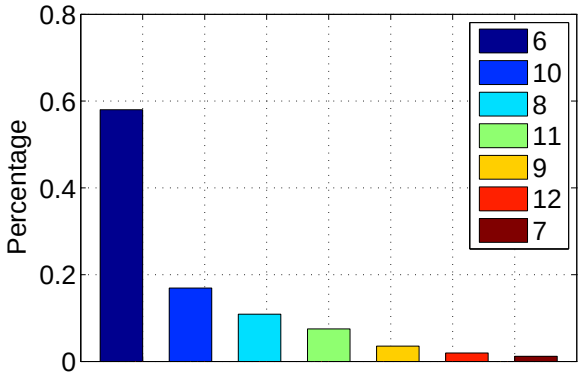
- Male users trigger triadic closure formation. The probability that three male users form a closed triad is $6\times$ as high as that of three female users.
- Gregarious users help form closed triads. The probability that three gregarious users form a closed triad is $39\times$ as high as that of three 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.

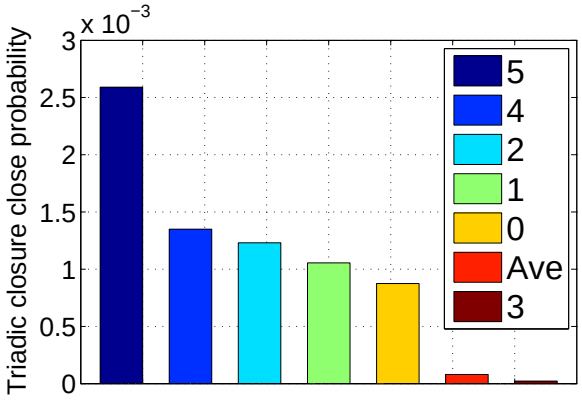
- Structural hole spanner is eager to close an open triad for more social resources ($> 10\times$ higher than that of three ordinary users). On the other hand, they are also reluctant to have two disconnected friends to be linked together.

- 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.

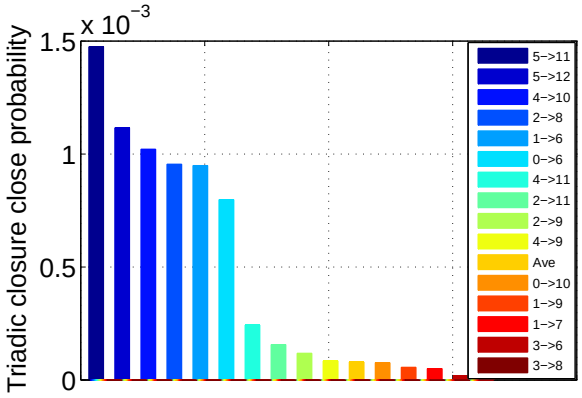
- In general, the closing action is often done by the third user (Figure 3.5(b), Figure 3.7(c)); since the third user is the last “active" user, he or she is more willing than the other users to connect the link. However, if the user has some social position, like “celebrity" or “resource holder," then ordinary users are more likely to connect with them (Figures 3.5(c), 3.7(a) and 3.7(b)) and close the triad.



(a) Distribution of close triads

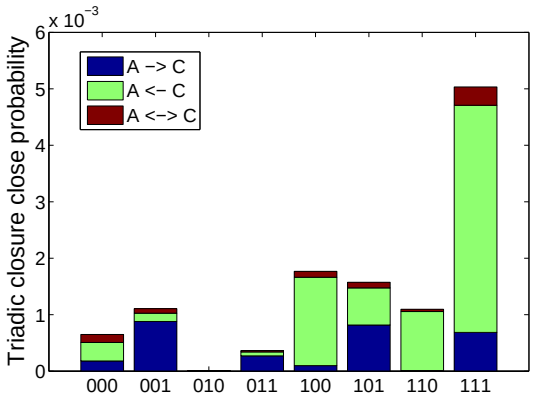


(b) Open triads that form triadic closure

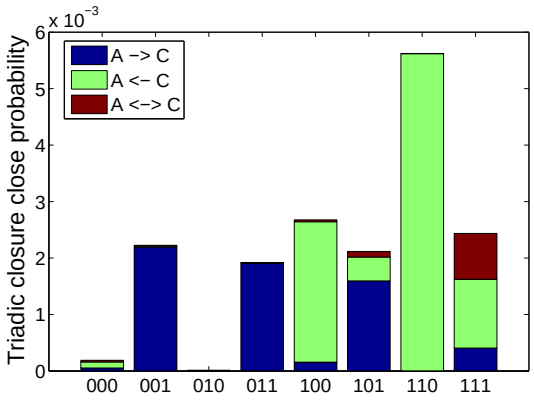


(c) Triads evolution

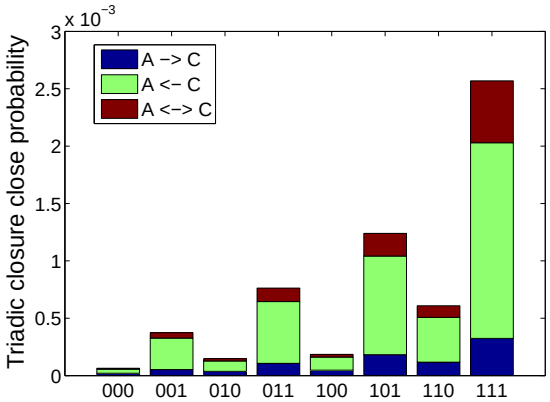
Fig. 3.6 Network Characteristics. (a) Y-axis: Percentage of newly formed closed triads. (b) Y-axis: probability that each open triad becomes closed. The number by the color bars means the index of open triads. (c) Y-axis: probability for each type of open triad (i.e., triad 0) to change from into each 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

Fig. 3.7 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.

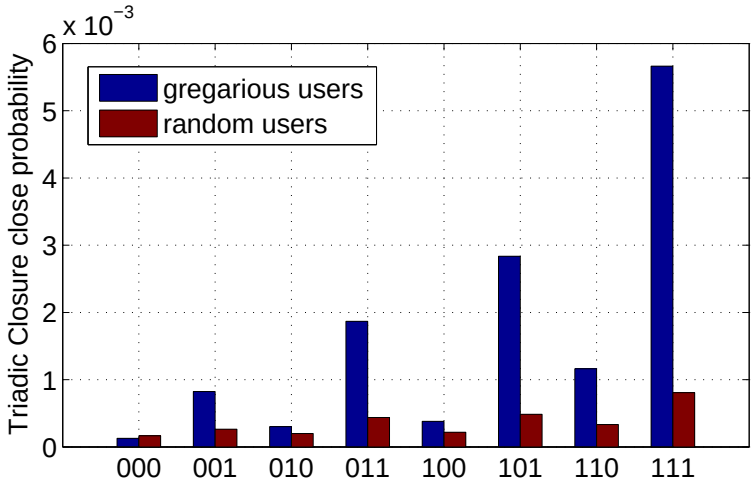


Fig. 3.8 Random test for gregarious users. Blue bars represent the case with gregarious users while red bars represent the case with random users. X-axis: represents the gregariousness of the triad. 0 indicates an ordinary user and 1 is used for a gregarious user.

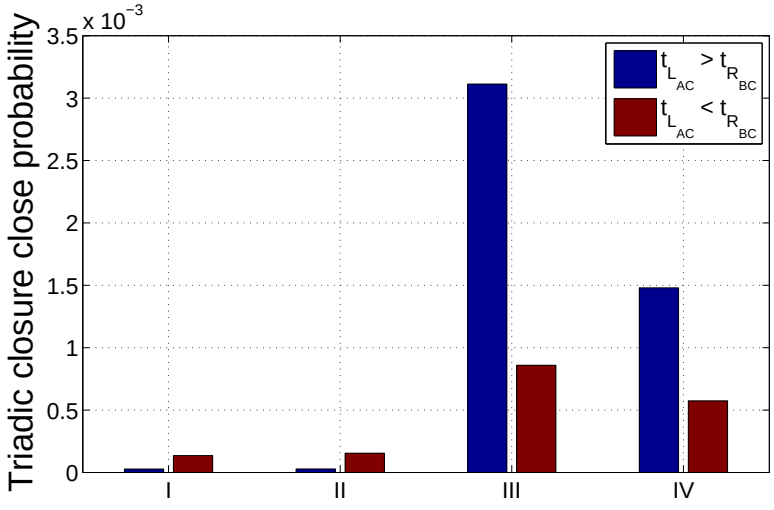


Fig. 3.9 Open triads that form triadic closures with social interaction information in different cases. X-axis: Cases. Y-axis: probability that open triads form triadic closures. $t_{L_{AC}}$ means the time that link AC is established, and $t_{R_{BC}}$ means the time that a retweet happens between user B and C.

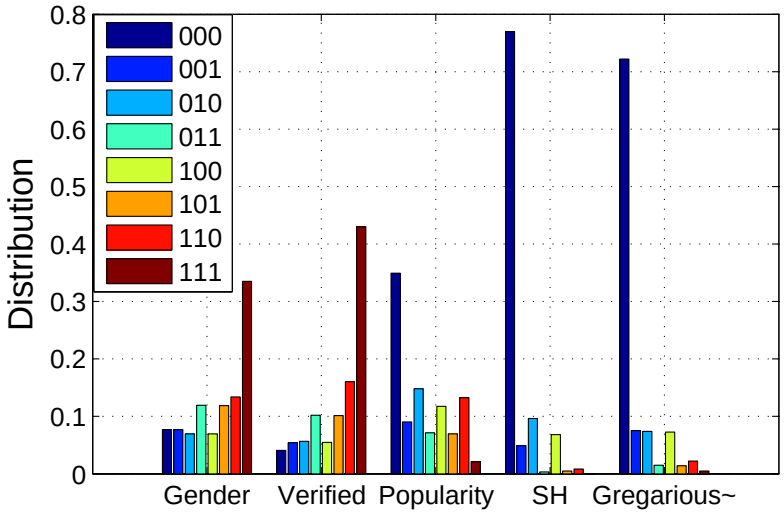


Fig. 3.10 Distribution of our observations. X-axis: Cases. Y-axis: probability for each type of closed triads. Legends: represents the status of the triad. 1 represents typical user(Male, verified user, popular user, structure hole spanner, and gregarious user) and 0 represents ordinary user.

Chapter 4

Tie Strength Dynamics

In this chapter, we first discern the degree to which triadic relationships in social networks are weakened. We then move on to examine how different factors influence the dynamics status of social triadic relationships in different networks. Specifically, given a closed triad $\mathscr{T}_{ABC}$ that becomes closed by the formation of e_{AC}, we investigate the following observations and factors:

- *Tie Dynamics in Triads:* What is the interaction dynamics of ties in social triads?
- *Social Tie:* How do the strength and reciprocity of the new tie e_{AC} and existing ties e_{AB} and e_{BC} affect the dynamics status of IF_e?
- *User Demographics:* How do users' demographic profiles – gender and social status – influence triadic tie strength dynamics?
- *Temporal Effects:* How long does the formation of link e_{AC} influence triadic tie strength dynamics?

4.1 Data Collection

We use two types of networks – social media and mobile networks. The social media network comes from Weibo, which is the dataset used in Chapter 3. The Weibo dataset we use contains more than 1 million users and more than 308 million following relationships (links) [54]. These users generated more than 12 million retweeting records from September 28^{th}, 2012 to October 29^{th}, 2012.

For each user, we have the demographic information – gender and verified status – in their online profiles.

Table 4.1 Data statistics for triadic tie dynamics.

Item	Weibo	Mobile
#Users	1,776,950	194,526
#Links	308,489,739	206,934
#Interactions	15,827,764	3,908,132
#Closed Triads	954,440	2,259,480
#Open Triads	241,364,986	35,314,058

The mobile network dataset is extracted from a subset of a collection of millions of mobile call detail records from an anonymous country. In this data, each user is anonymized by the data provider. We construct a sub-network by viewing each user as a node, and connecting a link between two users if they have at least one call from the observation window between August 1^{st}, 2008 and September 30^{th}, 2008. This resultant mobile network (Mobile) contains 194,526 users and 206,934 links. Table 4.1 details the statistics of the two networks.

4.2 Observation

Our problem is to observe tie strength dynamics in triads with directed links; there are in total 27 different types of directed triads. Here we consider the representative directed triangles with limiting e_{AB} and e_{BC} as reciprocal links, and leave the remaining cases for future work. The reason comes from the fact that reciprocal relationships are considered friendships or social relationships in social media [73, 86] or mobile communications [99, 27].

4.2.1 Tie Dynamics in Triads

We examine the ratios of strengthened and weakened ties in social networks and the degree to which a network as a whole presents weakened. We compare the results with interaction dynamics in both open triads and a network as a whole.

Suppose we have a set of dynamic networks $G = \{G^t = (V^t, E^t), t \in \{1, \cdots, t, \cdots, T\}\}$ with T observed days in real datasets, where $T = 90$ in the Weibo network and $T = 60$ in the Mobile network. We enumerate all combinations $< t, \Delta t >$ of t and Δt with $t \geq \Delta t$ and $t + \Delta t \leq T$. Given $< t, \Delta t >$, we study the closed triads that become closed at t and report the average

percentage of strengthened and weakened ties conditioned on Δt. Figure 4.1 plots the interaction dynamics of social triads in Weibo and Mobile.

In Figure 4.1, we can clearly see that in around 80% of closed triads the ties maintain a weakened state, triggered by the formation of the third link. While the number of the two kinds of strengthened ties is small, the increasingly strengthened ties (red lines) are consistently more numerous than the non-changing strengthened ones (green lines). Specifically, we observe that the interaction dynamics of link e_{AB} in Figure 4.1 (b,e) is a little weaker than that of e_{BC} in Figure 4.1 (c,f). Overall, the triadic relationships reveal that tie strength tends to become weaken in both the social media and mobile social networks.

We now compare the observations with the interaction dynamics in open triads to examine whether the dynamics status of a closed triad $\mathscr{T}_{ABC}$ arises from the formation of link e_{AC}. Similar to the experiments in Figure 4.1, we study the interaction dynamics of open triads for each $< t, \Delta t >$. Figure 4.2 plots the interaction dynamics of open triads in the Weibo and Mobile.

We can see the differences of corresponding interaction dynamics between closed triads (Figure 4.1) and open triads (Figure 4.2). In Weibo, the probabilities that links are weakened in open triads vs. closed triads are 60% vs. 80% in Figures 4.1 (a) and 4.2 (a), 15% vs. 70% in Figures 4.1 (b) and 4.2 (b), 50% vs. 80% in Figures 4.1 (c) and 4.2 (c). We can also see that in Mobile, the weakened probabilities in Figures 4.1 (d, e, f) are higher than those in Figure 4.2 (d, e, f). Conversely, the remaining strengthened cases in open triads are more than those in closed triads. In this sense, we conclude that the formation of the third link e_{AC} activates weakened ties in triads.

We finally examine the external environment of triadic tie strength dynamics in social networks. Essentially we plot the interaction frequencies over all links in Weibo and Mobile as a whole in Figure 4.3. In Weibo, we observe an upward trend in the average interactions of each link per day. In Mobile, we see that the line maintains a relatively strengthened trend. Therefore, we postulate that the network as a whole presents positive effects regarding triadic tie strength dynamics, which makes the fact that e_{AC}'s formation triggers the weakened ties in a triad $\mathscr{T}_{ABC}$ more pronouncedly.

4.2.2 Social Ties

We study how link e_{AC}'s attributes – such as tie strength and reciprocity – influence the dynamics status of ties in a closed triad $\mathscr{T}_{ABC}$. Although e_{AC}

belongs to the triad $\mathscr{T}_{ABC}$, we refer to it as an external tie because our goal is to study e_{AC}'s influence on e_{AB} and e_{BC} with its establishment.

Tie Strength of e_{AC}.

Tie strength represents the extent of closeness of social relationships. We measure the strength of social ties by the number of interactions between two users in Weibo (#retweets and #comments) and Mobile (#phone-calls) [39, 99, 27]. Such a definition suggests a way of answering the following question: How does the strength of the newly formed link e_{AC} affect the dynamics status of IF_e in $\mathscr{T}_{ABC}$? Figure 4.4 plots the influence of link e_{AC} with different tie strengths (indicated by different colors) in Weibo and Mobile. First, we observe that both networks present similar patterns of triadic tie strength dynamics. Second, surprisingly, we find that as e_{AC}'s tie strength increases, the likelihood that tie strengthens in closed triads $\mathscr{T}_{ABC}$ increases (*blue* to *red* to *green* lines). In other words, frequent interactions of the newly formed link between *A* and *C* promote the stronger ties in the closed triad $\mathscr{T}_{ABC}$.

Reciprocity of e_{AC}.

A reciprocal (two-way) relationship, usually developed from a parasocial (one-way) relationship, represents a stronger or trustful relationship between users in social media [73, 86] and mobile communications [99, 27]. We examine the extent to which the formation of link e_{AC} as a parasocial (one-way) and a reciprocal (two-way) relationship can affect the dynamics status of the closed triad $\mathscr{T}_{ABC}$. Figure 4.5 reports the results of triadic tie strength dynamics conditioned on the reciprocity of e_{AC}. From this figure, we see that while there are only slight differences between parasocial and reciprocal relationships, the parasocial e_{AC} consistently shows more positive effects on activating weakened ties in $\mathscr{T}_{ABC}$.

Tie Strength of e_{AB} and e_{BC}.

We investigate how the tie strength of e_{AB} and e_{BC} before the formation of link e_{AC} influence the tie dynamics in $\mathscr{T}_{ABC}$ after the establishment of e_{AC}. Relative to the external tie (e_{AC}), we refer to e_{AB} and e_{BC} as internal ties in $\mathscr{T}_{ABC}$. Our intuition is that with strong internal social ties, three users tend to maintain strengthened triadic relationships. Figure 4.6 details the results. Generally, we can see that in both Weibo and Mobile, internal tie strength and triadic

tie strength dynamics have a negative correlation; that is, ties in a social triad have a high probability to transit to a weakened state if it has strong internal ties before its closeness. The observation that is against our intuition indicates that three people with strong connections in open triadic relationships have the tendency to disperse their social focus and investment to the newly connected social tie that actually makes this open triad closed.

4.2.3 User Demographics

We investigate the interplay of triadic tie strength dynamics and user demographic profiles in social networks. Due to the unavailability of mobile users' demographic information, in this study we focus on Weibo. Specifically, we examine how users' gender and status correlate with the dynamics status of triadic relationships.

Gender.

Previous studies have revealed that females and males display different social behaviors and activities [27, 76]. Herein, we explore how people of different genders maintain their triadic social connections. Given a triad $\mathscr{T}_{ABC}$, we use a three-bit binary code *XXX* (*X=F* or *M* denotes a female or male user) to represent the gender information of its three users *A*, *B*, and *C*, respectively. We can enumerate eight different combinations of three users' gender. In Figure 4.7, we report the results of four special cases in our problem, i.e., *FFF, FMF, MFM,* and *MMM*. We can observe that different gender-based triads reveal different dynamics status. Generally, the decrease in tie strength among three males is more sharply than that among females. In opposite-gender triads, two females and one male (*FMF* in which the male serves as the bridge user *B*) have the tendency to maintain relatively strengthened relationships, compared with triads with two males and one female (*MFM*). Overall, we conclude that triadic relationships with more females (*FFF* and *FMF*) tend to be stronger compared with the relationships with more males (*MMM* and *MFM*).

Status.

We now look at the effects of users' social status [25] on triadic tie strength dynamics. Weibo.com provides a service for "celebrities" [1] to verify their real-world status, such as CEO, sports star, professor, and so on. We also use a

[1] We have also used degree and pagerank to category "celebrities" and got similar results.

three-bit binary code *XXX* ($X = 1$ or 0 denotes a verified celebrity or not) to represent the status information of three users A, B, and C in a triad. Figure 4.8 shows the dynamics status of triadic relationships conditioned on three users' social status. First, we can see that tie strengths between celebrities are more likely to be weakened as the closure of a triad than those between ordinary people. Similar to the gender-based observations above, we examine the triadic tie strength dynamics conditioned on the status of the bridge user B (*101* vs. *010*). The results show that the triadic relationships maintained by one celebrity and two ordinary people (*010*) are stronger than those in the reverse case (*101*). Overall, we conclude that the triadic relationships among celebrities (*111* and *101*) tend to be weaker compared with the relationships among ordinary people (*000* and *010*).

4.2.4 Temporal Effects

We finally study how the length of the observation timeframe – Δt – influences the dynamics status of triadic relationships. We use the interaction frequency within the timeframe Δt to determine the triadic tie strength dynamics.

From Figures 4.1, 3, 5 – 4.8, they are all plotted conditioned on the length of Δt (x-axis). Generally, we can see that when examining triadic tie strength dynamics conditioned on the interaction dynamics of e_{AB+BC} – subfigures (a, d) in Figures 4.1, 3, 5 – 4.8, the dynamics status of triadic relationships remains relatively horizontal as the increase of Δt. Specifically, we notice that in Figures 4.1 – 4.8 the overall increasing trends of triadic tie strength dynamics conditioned on e_{AB+BC} – subfigures (a, d) – come from the balance between the decreasing trends of triadic tie strength dynamics conditioned on e_{AB} – subfigures (b, e) – and the increasing trends conditioned on e_{BC} – subfigures (c, f) – as the length of Δt increases, respectively.

4.2.5 Summary

We conduct a regression analysis for the effects of different factors – user demographics and social ties – on triadic tie strength dynamics in Table 4.2. In this analysis, the ordinary least square model is used to model the relationships between the dependent variable (triadic tie strength dynamics) and the independent variables (Columns 1). It can be seen that the reciprocity of link e_{AC} plays more important role in making ties a triad $\mathscr{T}_{ABC}$ stay weakened than its strength (2^{nd} line vs. 3^{rd} line). We also observe that in most cases, the

demographics (gender and status) of three users are highly correlated with the tie dynamics. The regression results are consistent with the observations above.

According to the correlation and regression analysis above, we provide the following intuitions related to triadic tie strength dynamics in social networks:

- The triadic relationships are strongly weakened in both online social media and mobile social networks.
- The stronger the third tie is, the less likely the first two ties are weakened; while the stronger the first two ties are, the more likely they are weakened.
- The decrease in tie strength among three males is more sharply than that among females.
- Tie strengths between celebrities are more likely to be weakened as the closure of a triad than those between ordinary people.

Table 4.2 Regression Analysis for triadic tie strength dynamics.

	Dynamics status on IF_{AB}	Dynamics status on IF_{BC}	Dynamics status on IF_{AB+BC}
Reciprocity of e_{AC}	0.796 (0.015)	$8.23e$-03^{**} (0.013)	0.321 (0.013)
IF_{AC}	$6.32e$-09^{***} (0.011)	$5.02e$-09^{***} (0.010)	$1.87e$-11^{***} (0.009)
IF_{AB+BC}	$2.04e$-04^{***} (0.004)	$4.83e$-05^{***} (0.003)	$1.67e$-07^{***} (0.003)
Gender of A	$5.49e$-03^{**} (0.019)	0.593 (0.017)	0.538 (0.016)
Gender of B	0.197 (0.020)	0.833 (0.018)	0.610 (0.017)
Gender of C	0.718 (0.019)	$2.55e$-03^{**} (0.017)	$3.14e$-03^{**} (0.016)
Status of A	0.048^{*} (0.019)	0.875 (0.017)	$0.087^{\#}$ (0.017)
Status of B	0.561 (0.020)	0.232 (0.018)	0.233 (0.017)
Status of C	0.695 (0.019)	$3.55e$-03^{**} (0.017)	0.010^{*} (0.016)
R^2	0.019	0.023	0.029

Note: Robust standard errors in parentheses.
R^2: the proportion of variance in the criterion that is explained by the estimated regression model. Two-sided p-value are reported and its significant level at: 0.1 (#), 0.05 (*), 0.01 (**), 0.001(***).

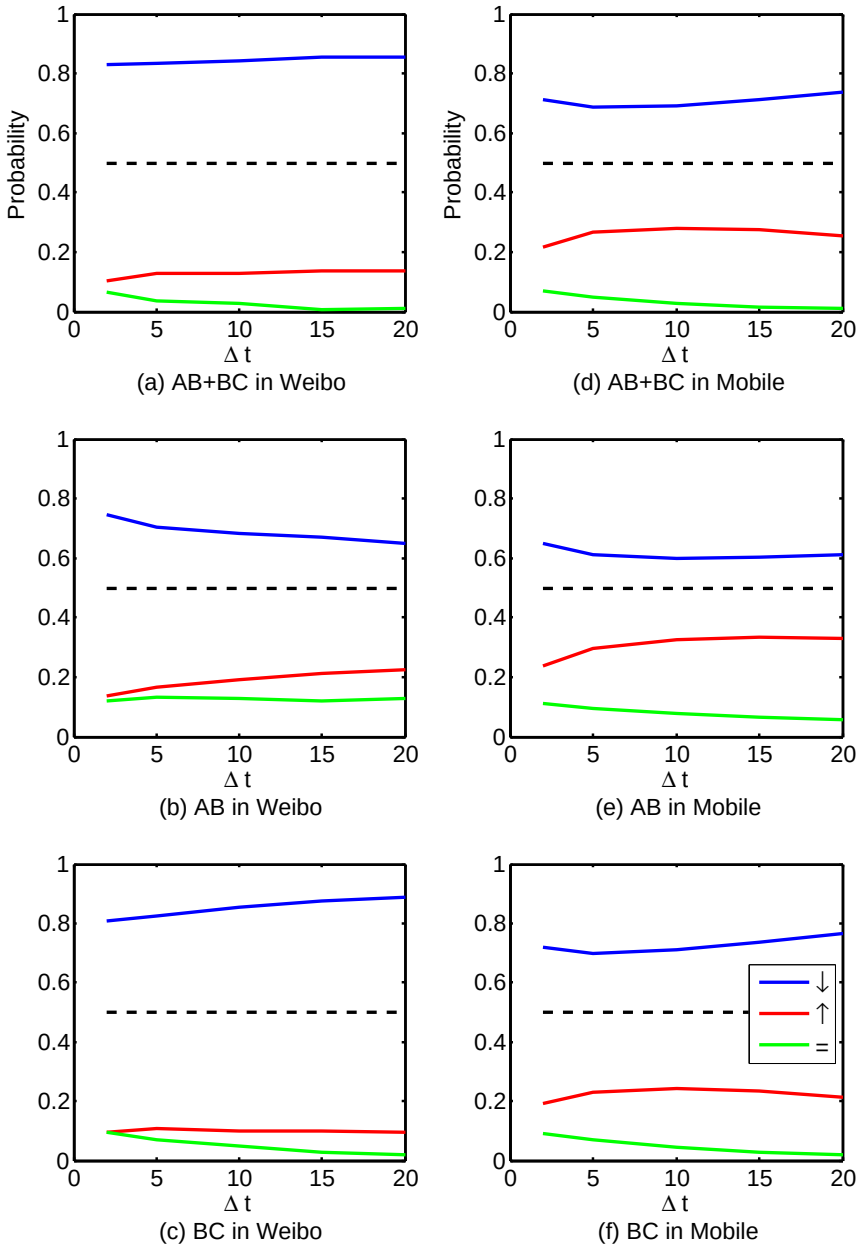


Fig. 4.1 Triadic tie strength dynamics of closed triads. x-axis: Δt; y-axis: Probability that a IF_e is strengthened (red and green lines) or weakened (blue line), conditioned on interaction dynamics of IF_{AB+BC} (a, d), IF_{AB} (b, e), and IF_{BC} (c, f).

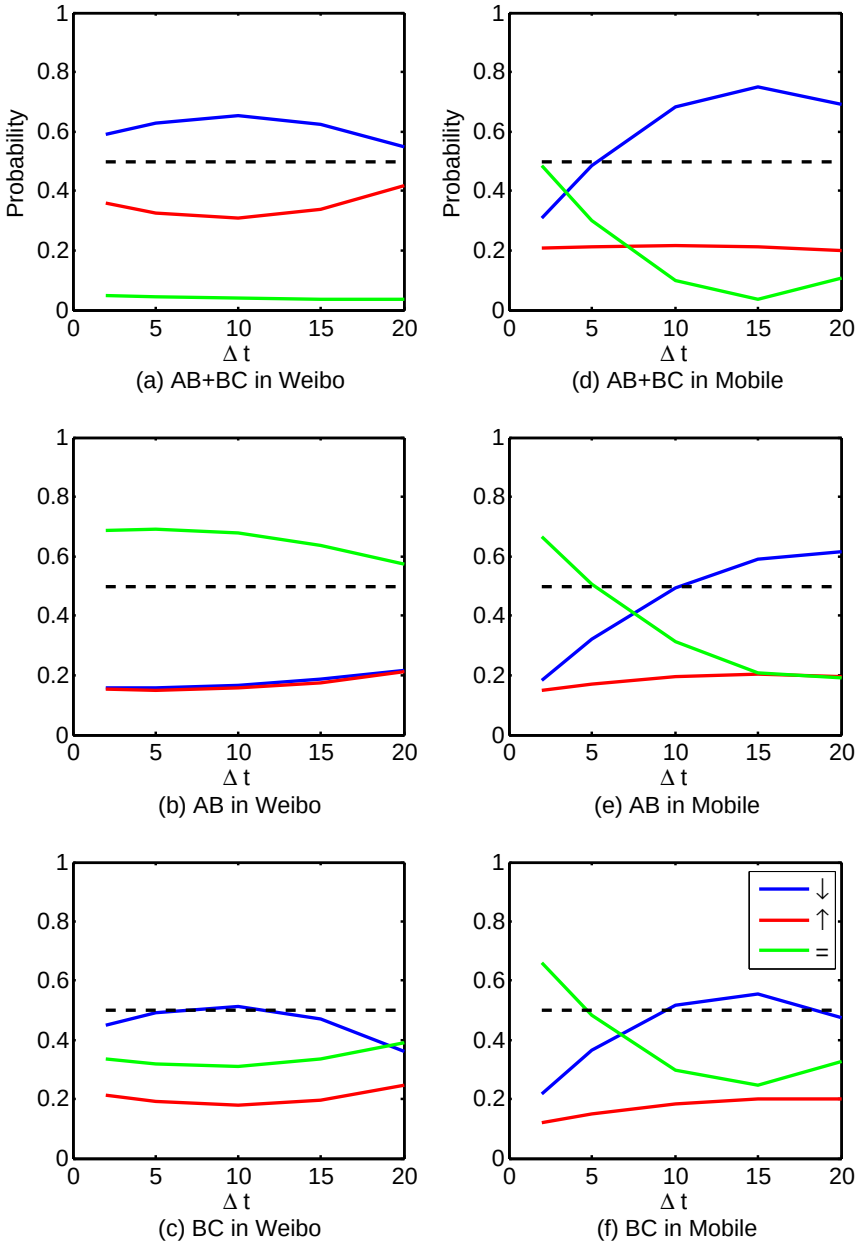


Fig. 4.2 Triadic tie strength dynamics of open triads. x-axis: Δt; y-axis: Probability that the interaction frequency of IF_e in an open triad is weakened (blue line), strengthened – increasing (red line), or strengthened – decreasing (green line), conditioned on the interaction dynamics of IF_{AB+BC} (a, d), IF_{AB} (b, e), and IF_{BC} (c, f).

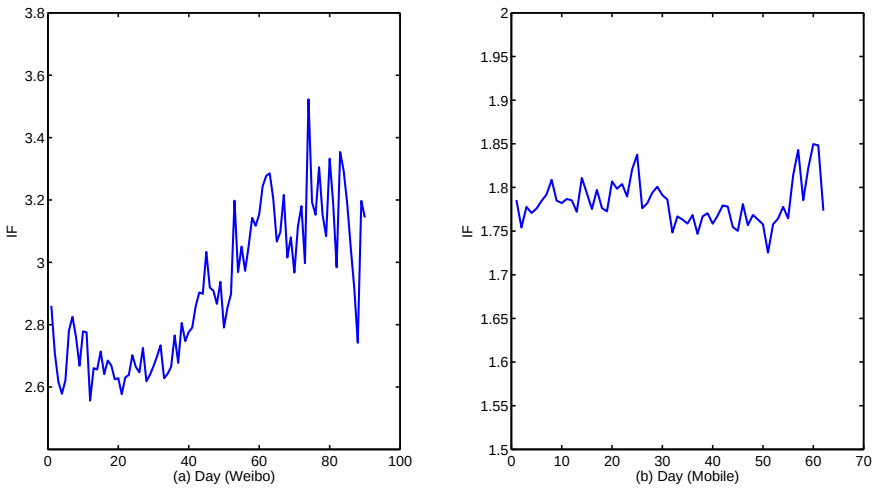


Fig. 4.3 Trends of the interaction frequency over time. *y*-axis: Interaction Frequency (#interactions per day). (a) Weibo network; (b) Mobile network.

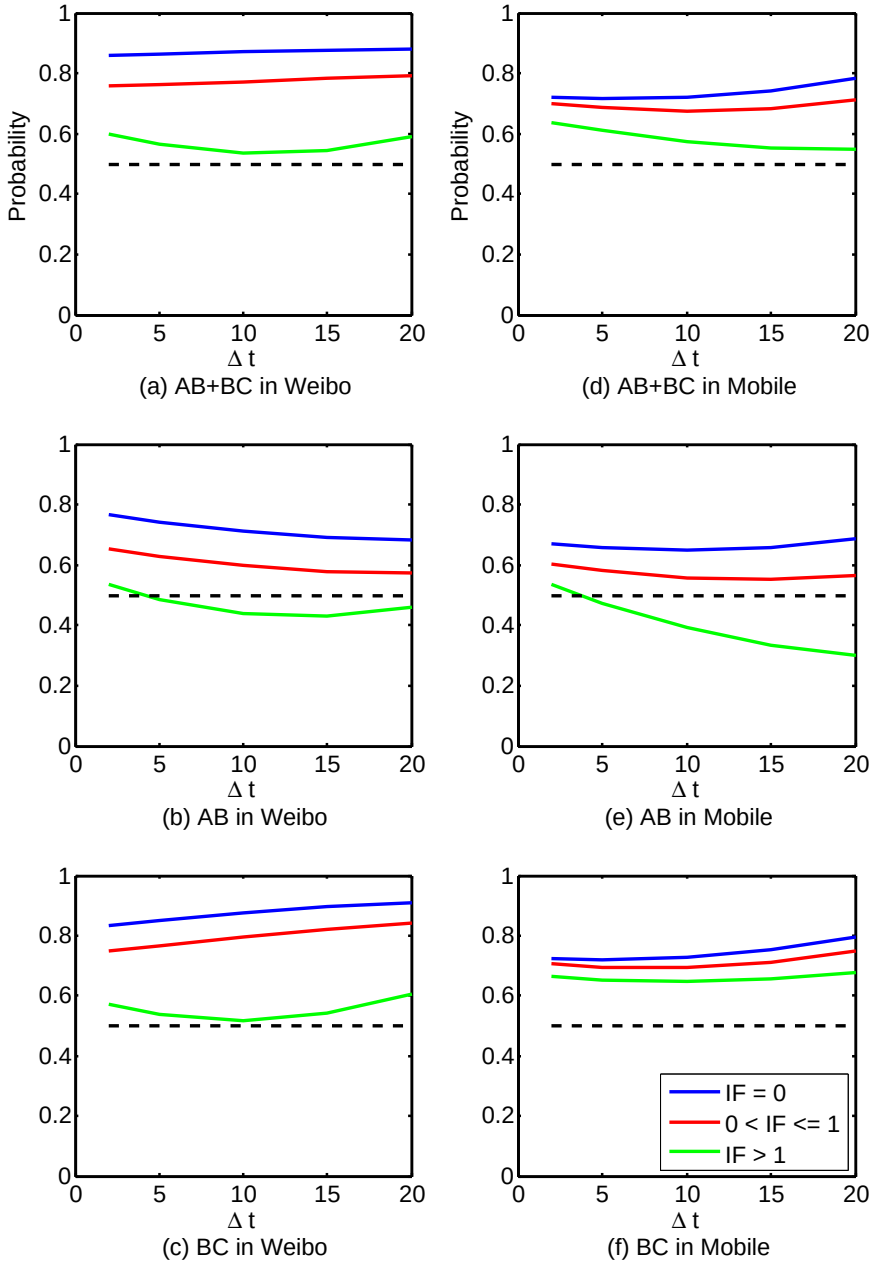


Fig. 4.4 External tie strength (e_{AC}). x-axis: Δt; y-axis: Probability that tie IF_e in $\mathscr{T}_{ABC}$ is weakened, conditioned on interaction dynamics of IF_{AB+BC} (a, d), IF_{AB} (b, e), and IF_{BC} (c, f). IF denotes the average number of interactions (#retweets and #comments in Weibo and #phone-calls in Mobile) per day.

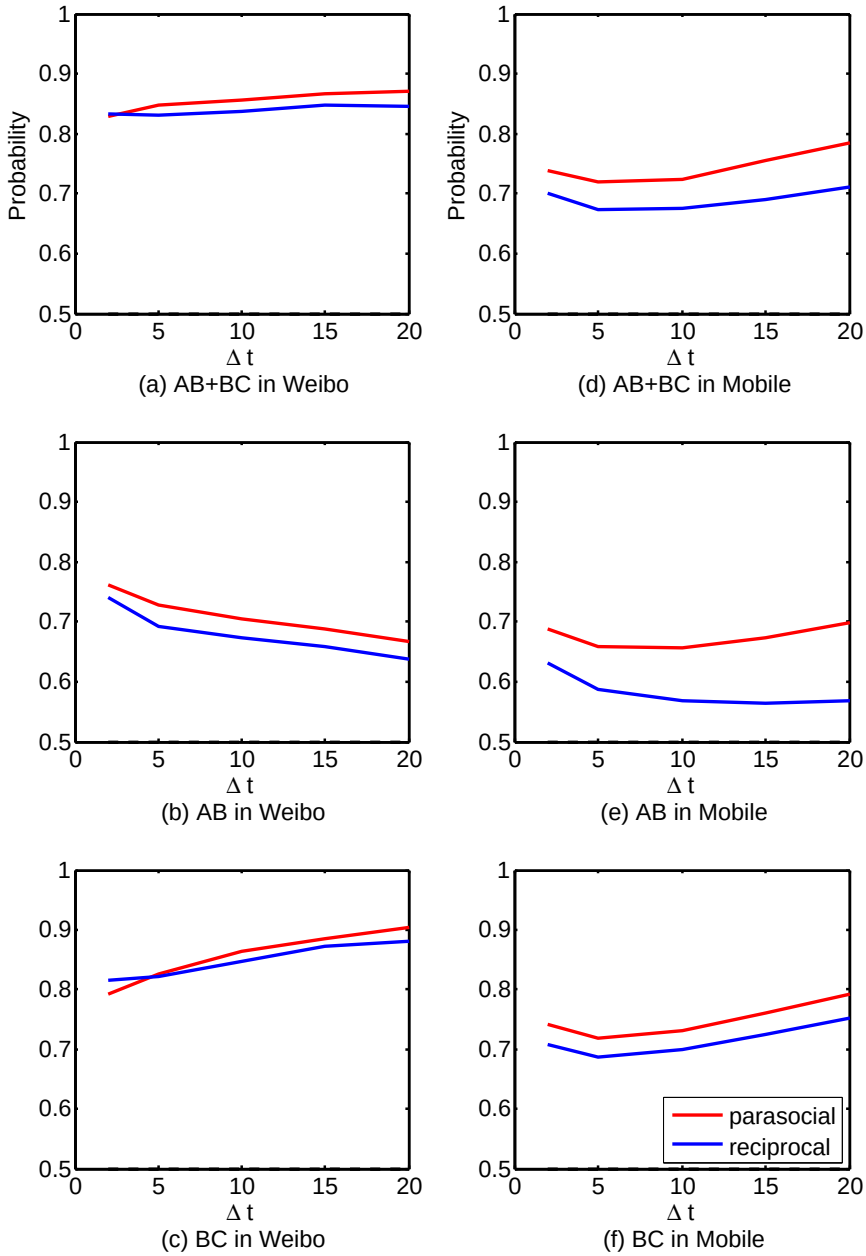


Fig. 4.5 Reciprocity of external tie e_{AC}. x-axis: Δt; y-axis: Probability that a tie IF_e in $\mathscr{T}_{ABC}$ is weakened, conditioned on interaction dynamics of IF_{AB+BC} (a, d), IF_{AB} (b, e), and IF_{BC} (c, f).

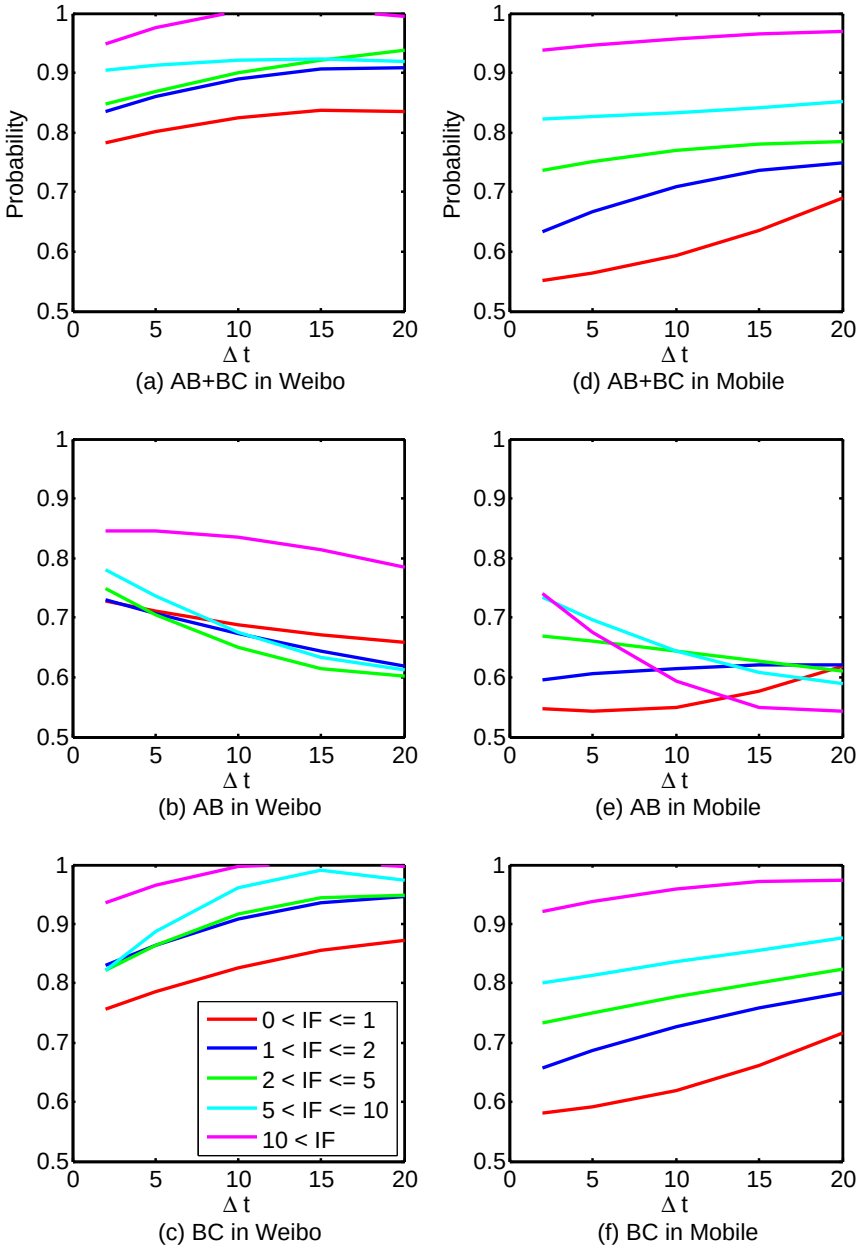


Fig. 4.6 Internal tie strength (e_{AB} and e_{BC}). x-axis: Δt; y-axis: Probability that a tie IF_e in $\mathscr{T}_{ABC}$ is weakened, conditioned on interaction dynamics of IF_{AB+BC} (a, d), IF_{AB} (b, e), and IF_{BC} (c, f). IF denotes the average number of interactions (#retweets and #comments in Weibo and #phone-calls in Mobile) per day.

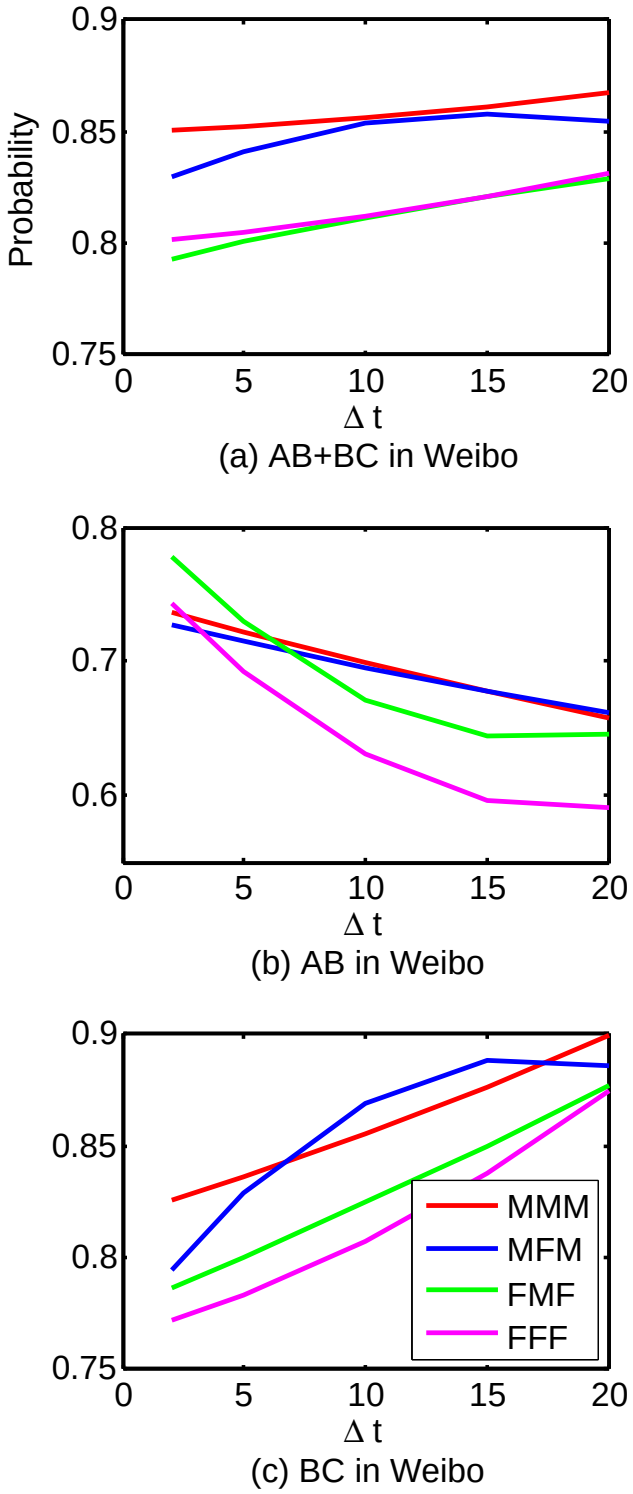


Fig. 4.7 User gender correlation in Weibo. F: female user; M: male user. x-axis: Δt;y-axis: Probability that a tie IF_e in $\mathscr{T}_{ABC}$ is weakened, conditioned on interaction dynamics of IF_{AB+BC} (a, d), IF_{AB} (b, e), and IF_{BC} (c, f).

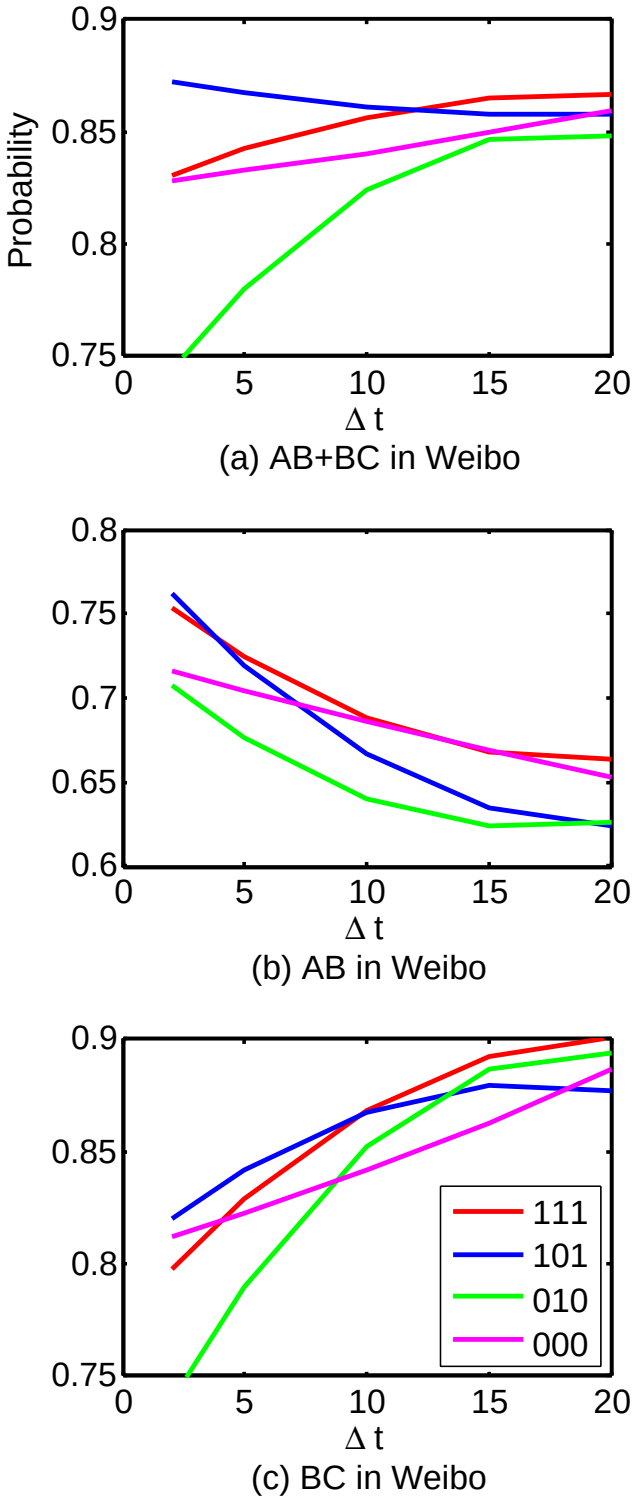


Fig. 4.8 User status correlation in Weibo. 1: verified celebrity; 0: ordinary user; x-axis: Δt;y-axis: Probability that a tie IF_e in $\mathscr{T}_{ABC}$ is weakened, conditioned on interaction dynamics of IF_{AB+BC} (a, d), IF_{AB} (b, e), and IF_{BC} (c, f).

Chapter 5

Prediction Model

In this chapter, we will first formulate our problems in a formal manner for triadic closure prediction and tie strength dynamics prediction problems. Then, we will propose kernel based factor models to solve these problems.

5.1 Problem Definition

In this part, we will formulate our problems in a formal manner.

5.1.1 Triadic Closure Prediction

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 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 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

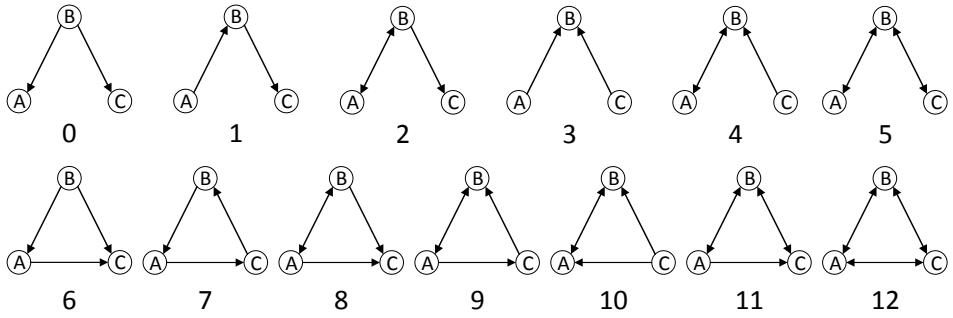


Fig. 5.1 Open Triads and Closed Triads. The number below is the index of each triad. Triad 0 – Triad 5 are open triads and Triad 6 – Triad 12 are closed triads. A, B and C represent users.

Table 5.1 How open triad forms triadic closure.

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

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 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 abovementioned 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.

We further consider the formation of implicit triads through social interactions alone. In other words, still considering retweeting as the interaction, if

retweeting happens between A and B, and between B and C, will retweeting happen between A and C? Before formally defining the problem, we introduce two new definitions of triads:

Definition 3 ***[Interaction Closed Triad]*** *For three users $\Delta = (A,B,C)$, if for any two users, there exists an interaction, then we say that Δ is an interaction closed triad.*

Definition 4 ***[Interaction Open Triad]*** *For three users $\Delta = (A,B,C)$, if an interaction happens between A and B and another interaction happens between B and C, but there is no interaction between A and C, we call the triad (A,B,C) an interaction open triad.*

Based on the definition of interaction open and closed triads, we can define the problem of Interaction Triadic Closure Prediction as below.

Problem 2 Interaction Triadic Closure Prediction *Given a network $G^t = (V,E)$ at time t and historical information regarding the formation of all interactions (e.g., all retweeting behaviors). For every candidate interaction 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 interaction open triad in G^t will become an interaction closed triad at a later time $t'(t' > t)$ – i.e.,*

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

where Y_I^t denotes the set of all values of the hidden variables at time t.

Notice that in Problem 2 we consider implicit triads formed by social interactions only; we do not consider relationships between individuals in the network. For example, at time t_1 user A retweets user B, and at time $t_2(t_2 > t_1)$ user C retweets user B. For Problem 2, we want to predict whether C will retweet A at time $t_3(t_3 > t_2)$, without considering whether there is a relationship between A and B, B and C, or A and C. While in Problem 2, our prediction is based on the relationship network of users.

Theoretically, Problem 1 and Problem 2 can be solved using the same technique. In the following sections, we will mainly concentrate on Problem 1.

5.1.2 Triadic Tie Strength Dynamics

Let $G^t = (V^t, E^t, I^t)$ denote a directed and weighted network at time t, where $V^t = \{v_i\}$ is the set of users, $E^t \subset V^t \times V^t$ is the set of links between users,

with each link denoted as $e_{ij} = (v_i, v_j) \in E^t$, and I^t represents the number of interactions of edges E^t during t. Then we can define a dynamic network $G = \{G^t = (V^t, E^t), t \in \{1, \cdots, t, \cdots, T\}\}$ over a timeframe T.

While extensive efforts have been devoted to explaining triadic relationships in social networks, such as social balance [50] and social status theories [22], little has been done to understand triadic interaction dynamics. In this work, we aim to model *the interaction dynamics of links e_{AB} and e_{BC} – when and after the third link e_{AC} is formed by which an open triad $\mathscr{O}_{ABC}$ becomes a closed triad $\mathscr{T}_{ABC}$.*

Suppose at timestamp t, the formation of edge e_{AC} turns an open triad $\mathscr{O}_{ABC}$ into a closed triad $\mathscr{T}_{ABC}$. Our goal is to trace the evolution of interactions in this triad when and after the transition happens. More specifically, we investigate the changes of the interaction frequencies of links e_{AB} and e_{BC} within a timeframe $[t - \Delta t, t + \Delta t]$, where Δt is an observation window. The interaction frequency IF_e is simply defined as the average interaction times, i.e., $IF_e^{[t-\Delta t, t+\Delta t]}/2\Delta t$.

Definition 5 ***Weakened or Strengthened Tie.*** *Given a closed triad $\mathscr{T}_{ABC}$ with the third link e_{AC} formed at timestamp t, and the interaction histories of e_{AB} and e_{AC}, in a future timestamp t' ($t' \geq t$), we call tie IF_e is weakened within the timeframe $[t' - \Delta t, t' + \Delta t]$ if the interaction frequency $IF_e^{[t', t'+\Delta t]}$ deceases significantly (with $p < 0.01$, t-test) compared with $IF_e^{[t'-\Delta t, t']}$; otherwise, IF_e is strengthened.*

In defining so, IF_e could be three different cases, including IF_{AB}, IF_{AC}, or IF_{AB+AC}. We formally define the problem of predicting triadic tie strength dynamics as follows.

Problem 3 Triadic Tie Strength Dynamics Prediction. *Given the set of closed triads $\mathscr{T}^t$ who become closed at timestamp t and a future timestamp t' ($t' \geq t$), the task is to learn a predictive function:*

$$f : (\{G, \mathscr{T}^t, \boldsymbol{X}\}) \rightarrow Y_{\mathscr{T}}^{t'}$$

,

where $Y_{\mathscr{T}}^{t'}$ denotes the dynamics status (strengthened vs. weakened) of IF_e within the timeframe $[t' - \Delta t, t' + \Delta t]$, with $y_i = 0$ indicating tie IF_e strengthens and $y_i = 1$ indicating that IF_e is weakened, and $\mathbf{X}$ is an attribute matrix associated with closed triads.

The problem is formalized in triads with directed links; there are in total 27 different types of directed triads. In this work we consider the representative directed triangles with limiting e_{AB} and e_{BC} as reciprocal links, and leave the remaining cases for future work. The reason comes from the fact that reciprocal relationships are considered friendships or social relationships in social media [73, 86] or mobile communications [99, 27].

5.2 Triadic Closure Prediction

Based on the observations in Chapter 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 [54, 53]. A similar model has been studied in [86] for reciprocal relationship prediction. We improve the model in [86] by using kernel-density estimate to better capture the similarity between open triads.

5.2.1 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 Chapter 3.2.1; 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 gener-

ative 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.

TriadFG-BF

Straightforwardly, we model these factor functions in a Markov random field, and by the Hammersley-Clifford theorem [47], 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 [74].

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

TriadFG-KF

Generally speaking, the binary feature function can discriminate closed triads and open triads. However, it cannot accurately capture correlation between

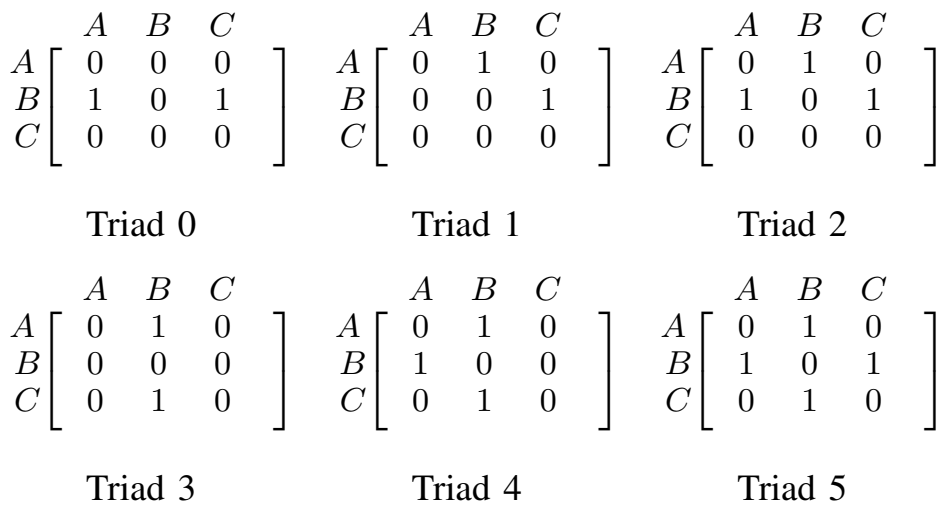


Fig. 5.2 Matrix representation of open triads.

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 [113, 132, 10]. In this paper, we use kernel-density estimate (KDE) [123] 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 5.2. Hence, we can define the similarity of triads using a Pearson's correlation coefficient as follows:

Definition 6 ***[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 [112]: non-negativity, identity of indiscernibles, symmetry, and triangle inequality, we define the triad similarity-based distance function as follows:

Definition 7 ***[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.0 The kernel-density estimates of structure information using the Gaussian kernel is shown in Figure 5.3 (green curve), the blue bars are the histogram of the distance to open triad 3.

Theoretically, we can use any smooth, strongly peaked function as a kernel, if the area under the curve of this kernel equals 1 – which is to make sure that the resulting KDE is normalized. In this work, we consider six commonly

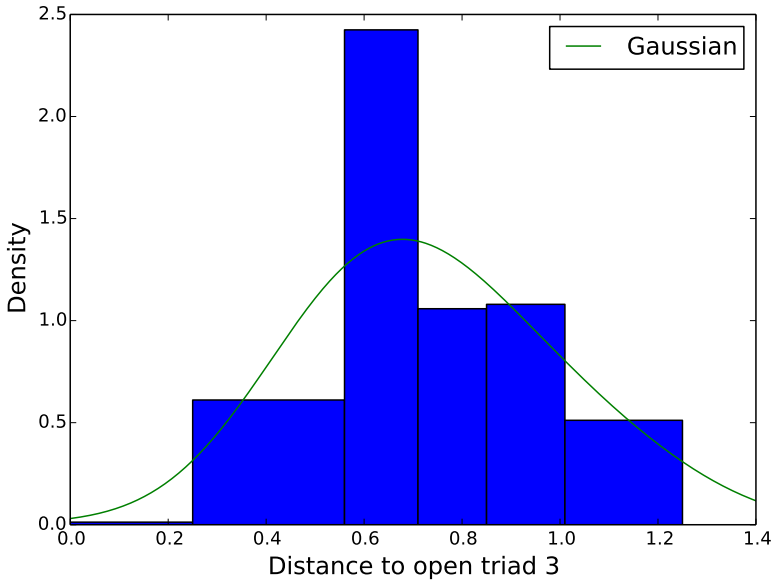


Fig. 5.3 Kernel density estimates within Gauss functions.

used kernel functions [123]. The curves of these functions are shown in Figure 5.4. The Tophat kernel, the Epanechnikov kernel, the linear kernel, and the cosine kernel are zero outside a finite range, whereas the Gaussian kernel and exponential kernel are nonzero everywhere, but negligibly small outside a limited domain.

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).

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:

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

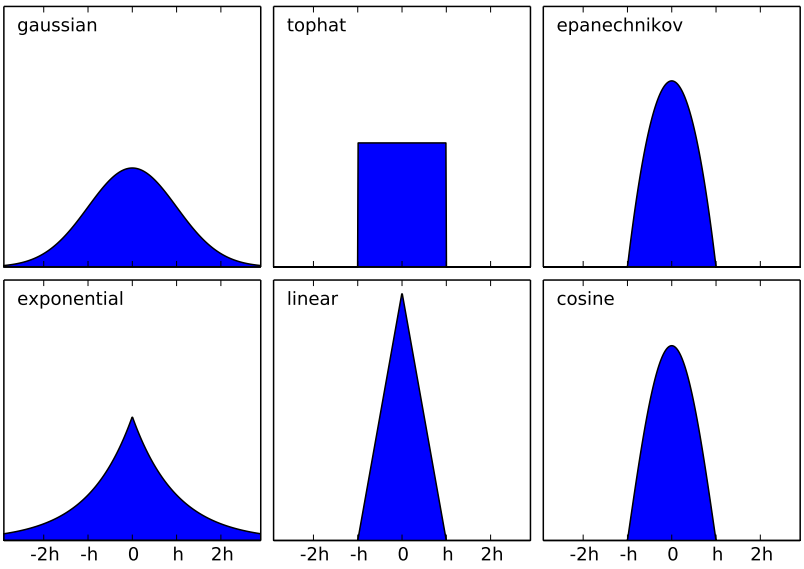


Fig. 5.4 Six kernel functions.

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 Chapter 5.2.2); 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 5.5. 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.,

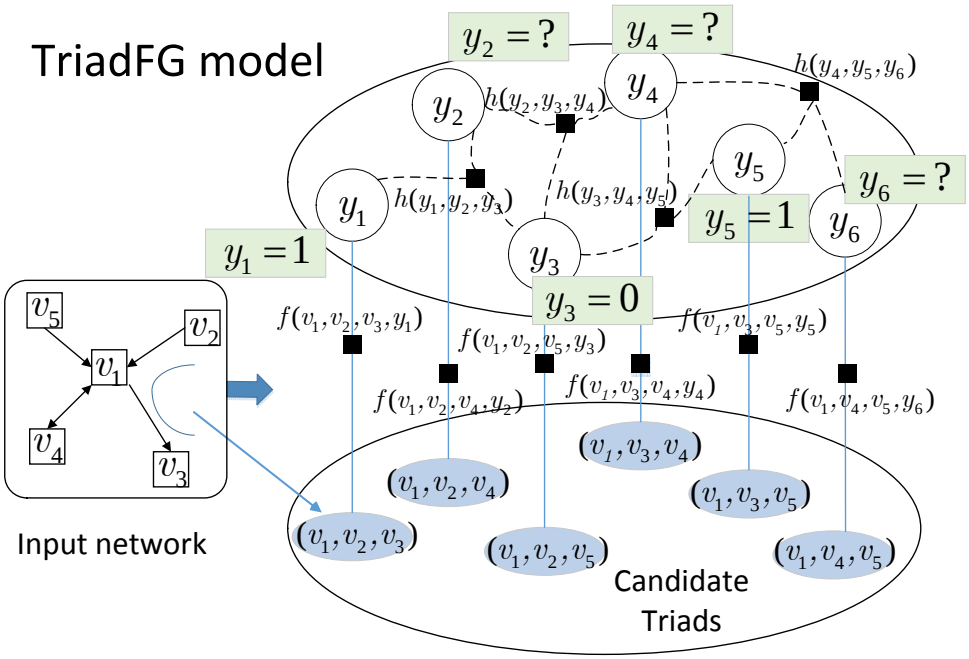


Fig. 5.5 Graphical representation of the TriadFG model. There are five users in the input network. Candidate open triads are illustrated as blue ellipses in the bottom right. White circles indicate hidden variables y_i. $f(v_1, v_2, v_3)$ represents the attribute factor function, and $h(.)$, the correlation function among triads.

$f(v_1, v_2, v_3)$. 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 5.5.

Comparison of Different Methods

Now we intuitively compare the three methods in this paper. In our model, we extract the feature functions from our observations. The main differences among these three approaches lie mainly in how we instantiate the structure feature and the probability of generating attributes $\mathbf{x}_i$, given the label y_i, say $P(\mathbf{x}_i|y_i)$. For TriadFG-BF, we choose binary function for each feature. For TriadFG-KF, we instantiate this probability $P(\mathbf{x}_i|y_i)$ within kernel-density estimation; but for TriadFG-EKF, we instantiate the feature function $f_j(x_{ij}, y_i)$ with kernel-density estimation. The kernel-density estimation for structure information is shown in Figure 5.6. From the figure, we can see that kernel-density estimate yields the green curve, so that for every open triad, we have

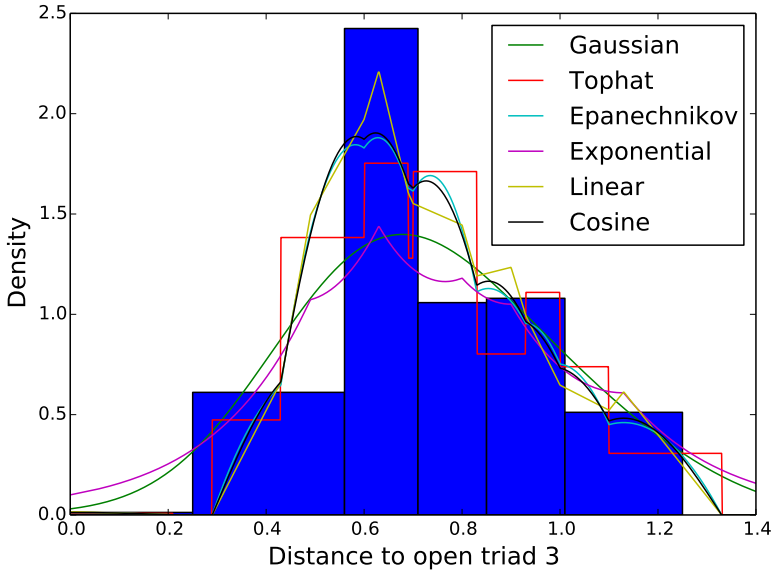


Fig. 5.6 Kernel density estimates within different kernel functions.

Table 5.2 Comparison of different methods.

Method	Feature function	Instantiate function
TriadFG-BF	$F1 = \begin{cases} 1, & \text{for open triad 5;} \\ 0, & \text{for others.} \end{cases}$	$g1 = \frac{1}{Z_1}\exp\{\alpha_j F1\}$
TriadFG-KF	$F2 = \sum_{i=1}^{N} \frac{1}{\beta} k(\frac{x-x_{ij}}{\beta})$	$g2 = \frac{1}{Z_1}\alpha_j F2$
TriadFG-EKF	$F3 = \sum_{i=1}^{N} \frac{1}{\beta} k(\frac{x-x_{ij}}{\beta})$	$g3 = \frac{1}{Z_1}\exp\{\alpha_j F3\}$

one estimation value, which gives more information for the estimation task, resulting in better prediction performance. We summarize the main differences in the Table 5.2.

5.2.2 Feature Definitions

Feature Definitions for Triadic Closure Prediction

We now depict how we define the factor functions in our models. According to the observations in previous sections, for problem 1 to predict triadic closure, we define 10 features of four categories: Network Structure(N), Demographics(D), Verified Status(V), and Social Information(S).

Network Structure. According to Figure 3.6(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 is not.

Feature Definitions for Triadic Closure Prediction with Interaction Information

We now introduce how we define the factor functions in our models for predicting triadic closure with interaction information. According to the observations in previous sections, for problem 2 to predict triadic closure with interaction information, we define 11 features of five categories: Network Structure(N), Demographics(D), Verified Status(V), Social Information(S), and Social Interaction(I). Except features in social interaction, features used for Problem 1 are identical to that for Problem 2. We simplify list all the features as follows.

Network Structure. 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 is not.

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.

5.2.3 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 [86].

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.

Algorithm 1: Learning algorithm for the TriFG model.

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

Initialize $\theta \leftarrow 0$;
repeat
Perform LBP to calculate marginal distribution of unknown variables $P(y_i|x_i,G)$;
Perform LBP to calculate the marginal distribution of triad c, i.e., $P(y_c|\mathbf{X}_c,G)$;
Calculate the gradient of μ_k according to Eq. 5.21 (for α_j with a similar formula):

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

Update parameter θ with the learning rate η:

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

until *Convergence*;

5.3 Tie Strength Dynamics Prediction

Our goal is to examine the extent to which the triadic tie dynamics status can be predicted in social networks. To do so, we propose a unified model to capture not only triads' attributes but also social and temporal correlations. In this section, the TRIST framework – a KDE-based Factor Graph (KFG) – is proposed for predicting triadic tie strength dynamics.

5.3.1 KDE-based Factor Graph (KFG)

Given a dynamic network $G = \{G^t = (V^t, E^t), t \in \{1, \cdots, t'\}\}$ and the attribute features $\mathbf{X}$ of candidate triads, we define an objective function by maximizing the conditional probability of triadic tie dynamics state Y, i.e., $P(Y|\mathbf{X},G)$. In a factor graph [72], the global probability can be factored as a product of local factor functions that capture both the attribute features $\mathbf{X}$ and structural and temporal correlations in the dynamic network G. Herein, we design three types of factor functions to model the observations in Chapter 4.2.

- Attribute factor $f(\mathbf{x_i^t}, y_i^t)$: The probability of a tie's dynamics state y_i^t at time t given the attribute vector $\mathbf{x_i^t}$ associated with each triad $\mathscr{T}_i$.

- Temporal factor $g(y_i^t, y_i^{t'})$, $t < t'$: The probability of a tie's dynamics state $y_i^{t'}$ at time t' given its state y_i^t at time t.
- Social factor $h(y_i^t, y_j^t)$: The probability of a tie's dynamics state y_i^t at time t given the dynamics status state y_j^t of a triad $\mathscr{T}_j$.

Thus, we can define the joint distribution over the triadic tie strength dynamics Y given G as

$$\begin{aligned} P(Y|\mathbf{X}, G) &\propto \prod P(\mathbf{X}|Y) \cdot P(Y|G) \\ &= \prod_{t=1}^{t'} \prod_{i,j}^{G} f(\mathbf{x_i^t}, y_i^t) g(y_i^t, y_i^{t'}) h(y_i^t, y_j^t). \end{aligned} \tag{5.13}$$

We illustrate the graphical representation of our proposed model in Figure 5.7. The bottom left part is the input network. From the input social network, we generate three closed candidate triads, including $\mathscr{T}_{v_1,v_2,v_3}$, $\mathscr{T}_{v_1,v_3,v_4}$ and $\mathscr{T}_{v_1,v_4,v_5}$. In the prediction model, these three candidate triads are modeled as yellow ellipses. The attribute features defined over the candidate triads at each timestamp are captured by the $f(\cdot)$ factor functions. The temporal correlations between triads at different timestamps are modeled by the $g(\cdot)$ factor functions. The social correlations between different triads are captured by the $h(\cdot)$ factor functions. Based on all the considerations, we construct the factor graph at the top level of Figure 5.7.

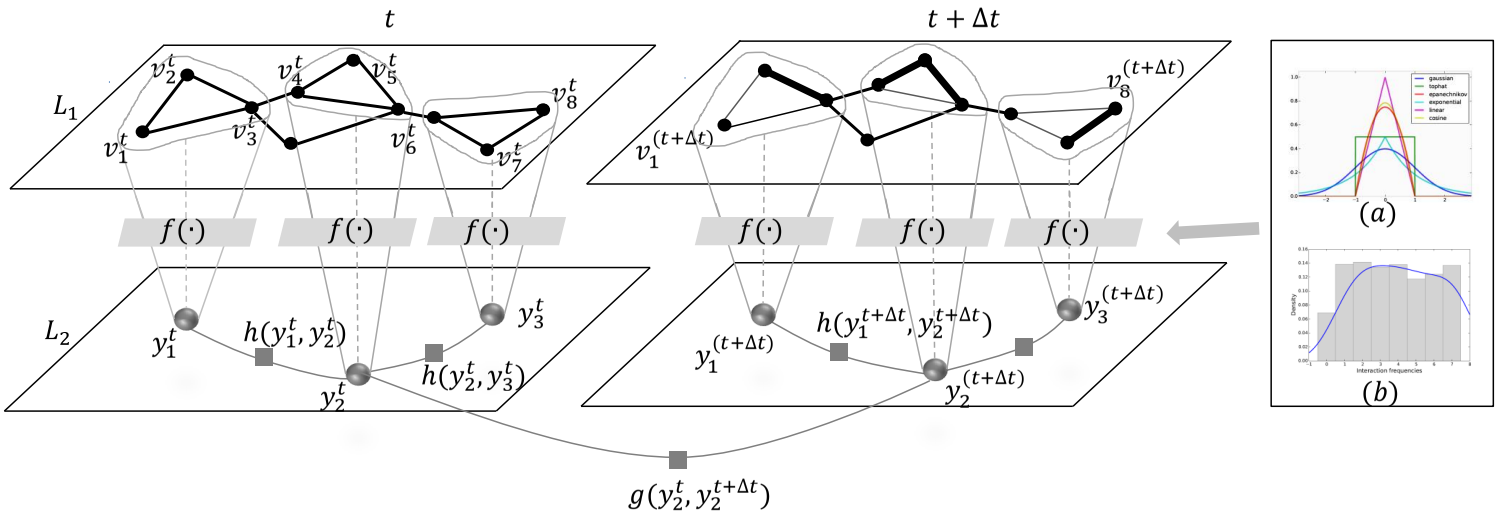


Fig. 5.7 Graphical representation of the TRIST model. Open triads become closed at time t. Three candidate triads in network layer L_1 are mapped into three hidden variable nodes $y_i^t, i = 1,2,3$ in factor graph layer L_2. We observe two timestamps t and $t+\Delta t$ ($\Delta t > 0$). The broader the line is, the stronger the tie is. $h(\cdot)$ represents the social correlation function between two triads, and $g(\cdot)$ represents the temporal correlation function between the two statuses of one triad at two timestamps. In order to get attribute factor function, we applied kernel density estimation $f(\cdot)$. An example of the histogram and kernel density estimation of the tie strength of e_{AB} and e_{BC} using a Gaussian kernel is shown in (b). y-axis: density; x-axis: interaction frequencies of ties e_{AB} and e_{BC}. In (a), we present different kernel functions.

KDE-based attribute factor.

Straightforwardly, we initialize the defined factors in a Markov random field based on the Hammersley-Clifford theorem [47]. In particular, we initialize the attribute factor as

$$f(\mathbf{x_i^t}, y_i^t) = \frac{1}{Z_1} \exp\{\sum_{k=1}^{K} \alpha_k \Phi(x_{ik}^t, y_i^t)\}, \tag{5.14}$$

where α_k is the weight of the k^{th} attribute, K is the number of features, and $\Phi(\cdot)$ is the attribute feature function.

There are various ways to instantiate the attribute feature function $\Phi(\cdot)$, i.e., binary function [54, 86]. However, a binary feature function cannot accurately capture correlations and similarities among features and would lose important information. In order to better capture these similarities, we propose using kernel density estimation to instantiate attribute factor functions.

Kernel density estimation (KDE) [123], without any priori information on the probability distribution of the dataset, using a non-parametric way to estimate random variables' density functions, is an attractive technique to obtain estimations [36, 136, 80]. To obtain a kernel density estimation, we first place a kernel – a smooth, strongly peaked function – at the position of each data point, and then add up the contributions from all the estimations to obtain a smooth curve. For instance, Figure 5.7 shows an example of the histogram (gray part) and kernel density estimation with a Gaussian kernel (blue curve) for the interaction-strength feature. Specifically, for a Gaussian kernel, we have the following definition:

$$\Phi(x_{ik}^t, y_i^t) = \sum_{n=1}^{N_k} \frac{1}{\lambda} \kappa(\frac{x_{ik} - x_n}{\lambda}),$$

By using kernel density estimation, the attribute feature function can be initialized as

$$f(\mathbf{x_i^t}, y_i^t) = \frac{1}{Z_1} \exp\{\sum_{k=1}^{K} \sum_{n=1}^{N_k} \alpha_k \frac{1}{\lambda} \kappa(\frac{x_{ik} - x_n}{\lambda})\}, \tag{5.15}$$

where $\kappa(\cdot)$ is the kernel function with a peak at x_n , λ is the kernel bandwidth, and N is the number of data points of each feature. We thus get the real values of the attribute features.

Table 5.3 Kernel functions.

Kernel	Function
Gaussian kernel	$\kappa(x) = 1/\sqrt{(2\pi)}\exp(-1/2\,x^2)$
Tophat kernel	$\kappa(x) = 1/2$ if $\lvert x\rvert \leq 1$
Epanechnikov kernel	$\kappa(x) = 3/4(1-x^2)$ if $\lvert x\rvert \leq 1$
Exponential kernel	$\kappa(x) = \exp(-x)$
Linear kernel	$\kappa(x) = 1-x$ if $\lvert x\rvert \leq 1$
Cosine kernel	$\kappa(x) = \pi/4\cos(\pi/2\,x)$ if $\lvert x\rvert \leq 1$

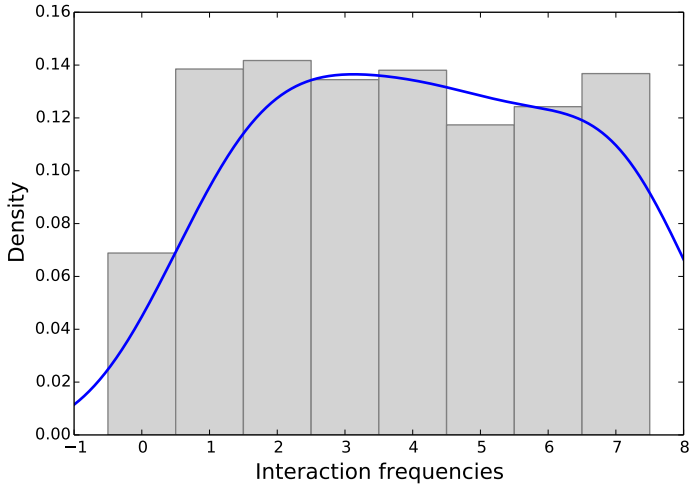


Fig. 5.8 Kernel density for attribute factors.

Theoretically, we can use any smooth, strongly peaked function as a kernel, if the area under the curve of this kernel equals 1 – which is to make sure that the resulting KDE is normalized. In this work, we consider six commonly used kernel functions [123] as shown in Table 5.3. The curves of these functions are shown in Figure 5.4. The Tophat kernel, the Epanechnikov kernel, the linear kernel, and the cosine kernel are zero outside a finite range, whereas the Gaussian kernel and exponential kernel are nonzero everywhere, but negligibly small outside a limited domain. By default, we use a Gaussian kernel in our TRIST framework and discuss the effects of different kernel choices in experiments.

It is worth mentioning that a KDE-based attribute factor can be used to study graph kernels [122] as well. Intuitively, graph kernels can be considered as functions measuring the similarity of pairs of graphs, making the whole family of kernel methods applicable to graphs [114, 122, 3]. In our work, we aim to measure similarity between different candidate closed triads, which also provides a way to measure graph similarity.

Temporal factor.

For the temporal factor, we model the interrelations between different weakened states of one triad at different timestamps. Specifically, we define it as:

$$g(y_i^t, y_i^{t'}) = \frac{1}{Z_2} e^{\sigma_1 (t'-t)} \exp\{\beta_i \Psi(y_i^{t'}, y_i^t)\}, \tag{5.16}$$

where $e^{\sigma_1 (t'-t)}$ is a triad-independent time-increase factor, σ_1 is a pre-defined parameter, $\Psi(\cdot)$ is a temporal feature function defined as $\Psi(\cdot) = (y_i^{t'} - y_i^t)^2$ and β is the weight of $\Psi(\cdot)$. In the model, we only consider the dependency of triadic tie strength dynamics between two subsequent timestamps as the same assumption in a hidden Markov model [38], Markov Random field [47] and Kalman Filters [48]. The triad-independent time-increase factor $e^{\sigma_1 (t'-t)}$ is defined as an exponential function, so that its parameter σ_1 can be simply absorbed by combining it with β_i. Thus, the above temporal factor function can be rewritten as

$$g(y_i^t, y_i^{t'}) = \frac{1}{Z_2} \exp\{\beta_i (t'-t)(y_i^{t'} - y_i^t)^2\}. \tag{5.17}$$

Social factor.

Intuitively, a triad's dynamics status state may be influenced by other triads; e.g., its neighborhood. For example, closed triad $\mathscr{T}_{ABC}$ and closed triad $\mathscr{T}_{ABD}$ share a common link e_{AB}, then the dynamics status state of $\mathscr{T}_{ABC}$ is highly correlated with that of $\mathscr{T}_{ABD}$. In addition, individuals have a tendency to associate and bond with similar others according to homophily theory [91]. Individuals in homophilic relationships share common characteristics, which makes similar triads have the similar dynamics status patterns [91].

Similarly, in order to capture the correlations between two triads, we define the social factor function as:

$$h(y_i^t, y_j^t) = \frac{1}{Z_3} \exp\{\gamma_{ij} (y_i^t - y_j^t)^2\}, \tag{5.18}$$

where γ_{ij} is the weight of the feature function, representing the influence degree of y_i on y_j.

Finally, by integrating Eq. (5.15), (5.17), and (5.18) into Eq. (5.13), we can obtain joint probability as follows:

$$P(Y|\mathbf{X},G) \propto \frac{1}{Z}\exp\{\sum_{t=1}^{t'}\sum_{i}^{G}\sum_{k=1}^{K}\alpha_k\Phi_k(x_{ik}^t,y_i^t) + \sum_{t=1}^{t'}\sum_{i}^{G}\beta_i(t'-t)(y_i^{t'}-y_i^t)^2 + \sum_{t=1}^{t'}\sum_{i,j}^{G}\gamma_{ij}(y_i^t-y_j^t)^2\}, \tag{5.19}$$

where $Z = Z_1Z_2Z_3$ is a normalization factor to guarantee that the result is a valid probability.

5.3.2 Feature Definitions

We now describe how we define the factor functions in our model. According to the analysis in the previous section, we define factor functions of three categories: an attribute factor, a temporal factor and a social factor.

Attribute factor. Attribute factors include social tie and user demographics. For social tie, we define a feature for e_{AC}'s reciprocity indicating whether e_{AC} is reciprocal or not, a feature for the tie strength of e_{AC} denoting the interaction frequency between users A and C, and also a feature for tie strengths of e_{AB} and e_{BC}. The tie strength features are modeled via kernel density estimation. For Weibo dataset, we also define six features based on user demographics – gender and status of three users – with their kernel density estimates respectively (Cf. Eq. (5.15)).

Temporal factor. Temporal factors are used to model the interrelations between the states of one triad at different timestamps (Cf. Eq. (5.17)).

Social factor. We define one correlation function for social factors to see whether any two triads have the same triadic tie strength dynamics patterns (Cf. Eq. (5.18)).

5.3.3 Learning and Prediction

Model Learning.

Given the joint probability in Eq. (5.19), we have the following log-likelihood objective function

$$\mathcal{O}(\theta) = \log P_\theta(Y|\mathbf{X},G). \tag{5.20}$$

The task of model learning is to estimate a parameter configuration $\theta = (\{\alpha_k\}, \{\beta_i\}, \{\gamma_{ij}\})$ that maximizes the log-likelihood objective function – i.e.,

$$\theta = \arg\max \mathcal{O}(\theta).$$

To solve the maximization problem, we employ a gradient descent method. The idea is that each parameter θ is assigned an initial value, and then the gradient of each parameter with regard to the objective function is calculated. For example, the gradient of the parameter α_k with regard to Eq. (5.20) is written as:

$$\frac{\mathcal{O}(\theta)}{\alpha_k} = \mathbb{E}[\Phi_k(x_{ik}^t, y_i^t)] - \mathbb{E}_{P_{\alpha_k}(Y|\mathbf{X})}[\Phi_k(x_{ik}^t, y_i^t)], \tag{5.21}$$

where $\mathbb{E}[\Phi_j(x_{ik}^t, y_i^t)]$ is the expectation of factor function $\Phi_k(x_{ik}^t, y_i^t)$ given the data distribution; and $\mathbb{E}_{P_{\alpha_k}(Y|\mathbf{X})}[\Phi_k(x_{ik}^t, y_i^t)]$ is the expectation of factor function $\Phi_k(x_{ik}^t, y_i^t)$ under the distribution $P_{\alpha_k}(Y|\mathbf{X})$ estimated by the model. Before calculating variables' marginal distribution, we need first get factor functions for different factors; i.e., we run KDE to get estimates for attribute factors, and get temporal factors from two subsequent timestamps in the whole period we observed.

Similarly, the gradients of parameters β_i and γ_{ij} can be derived. Finally each parameter is updated with a learning rate η:

$$\theta_{m+1} = \theta_m + \eta \cdot \frac{\mathcal{O}(\theta)}{\theta}, \tag{5.22}$$

where m is the iteration time. The learning algorithm is summarized in Algorithm 2.

Prediction.

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., $Y^\star = \arg\max \mathcal{O}(Y|\mathbf{X}, G, \theta)$. Specifically, we use the learned model to calculate the marginal distribution of each candidate triad with unknown variable and finally assign each candidate triad with a label of the maximal probability.

Algorithm 2: Learning algorithm for the TRIST model.

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

Initialize $\theta \leftarrow 0$;
repeat
  repeat
    Get temporal factors between timestamp t and its last timestamp $t-1$;
    Update t;
  until $t \geqslant T$;
  Run KDE to get estimates of attribute factors:

$$f(\mathbf{x_i^t}, y_i^t) = \frac{1}{Z_1} \exp\{\sum_{k=1}^{K} \sum_{n=1}^{N_k} \alpha_k \frac{1}{\lambda} \kappa(\frac{x_{ik} - x_n}{\lambda})\}$$

  Perform LBP to calculate marginal distribution of unknown variables $P(Y, Y^L|G)$;
  Perform LBP to calculate the marginal distribution of known variables $P(Y|G)$;
  Calculate the gradient of α_j according to Eq. (5.21) (for β_i and γ_{ij} with a similar formula); Update parameter θ with the learning rate η with Eq. (5.22);
until *Convergence*;

Chapter 6

Experiments and Discussions

6.1 Triadic Closure Prediction

6.1.1 Experiment Setup

We use the dataset described in Chapter 3.1 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

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. § 5.2.1).

TriadFG-KF. Represents the proposed TriadFG model with kernel feature functions (Cf. § 5.2.1).

TriadFG-EKF. Represents the proposed TriadFG model with exponential kernel functions (Cf. § 5.2.1).

Table 6.1 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**

For SVM and Logistic, we use Weka[46]. 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.

Evalution Measures

We evaluate the performance of different approaches in terms of accuracy, precision, recall, and F1-Measure.

6.1.2 Triadic Closure Prediction

Prediction Performance

We now list the performance results for different methods in Table 6.1. 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.

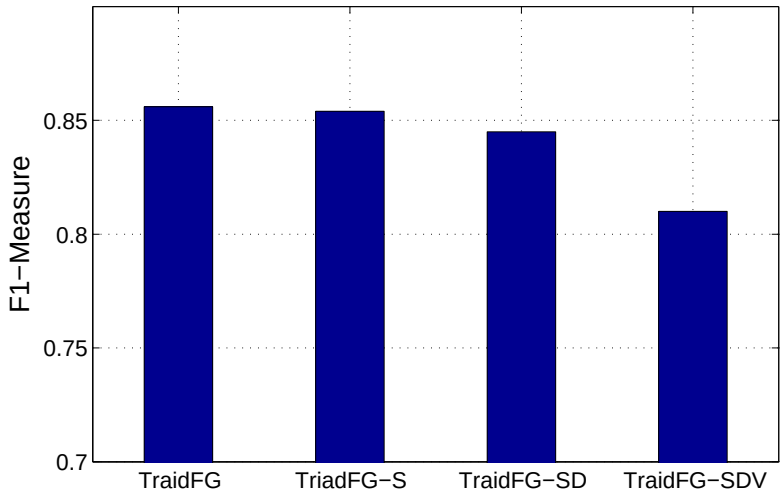


Fig. 6.1 Factor contribution analysis. TriadFG-S denotes ignoring social information when we use TriadFG model, TriadFG-SD denotes ignoring social information and demographics while TriadFG-SDV denotes further ignoring verified status information.

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. 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 6.1, 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 6.2. 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

Table 6.2 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

Table 6.3 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**

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.

6.1.3 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 6.3. 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 interaction 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). We first rank the individual factors by removing 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. 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 6.2, 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.

6.1.4 Discussion

Convergence Property

We now conduct an experiment to see the effect of the number of the loopy belief propagation iterations, shown in Figure 6.3 (for TriadFG-KF and TriadFG-BF; TriadFG-EKF has similar properties). We can see that on average, the performance of the algorithm becomes stable after about 120 iterations, which suggests that the learning algorithm converges well.

Computation Time.

We then conduct an experiment to see the computation cost in terms of time, shown in Table 6.4. We can see that Logistic Regression runs the fastest. For TriadFG, it converges the slowest. This is because factor graph inference is relatively complicated, so it takes more time to converge. However, its F1-Measure is significantly better than others (shown in Table 6.1).

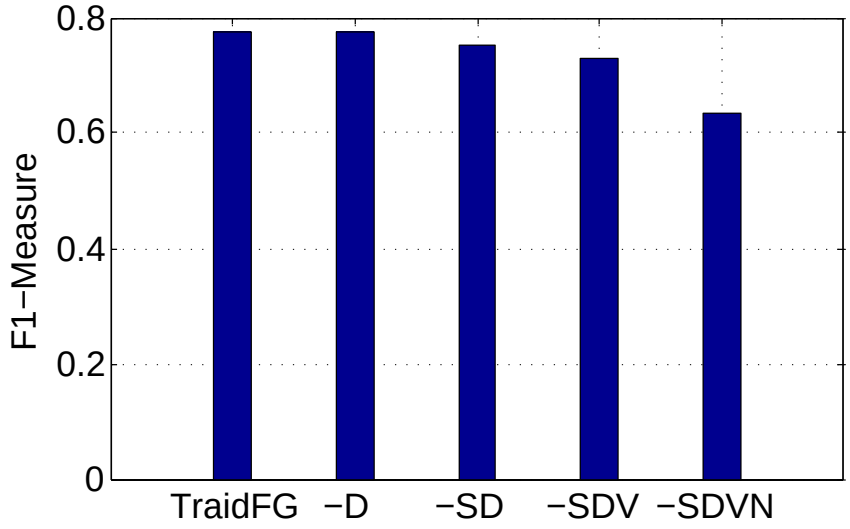


Fig. 6.2 Factor contribution analysis. TriadFG-D denotes ignoring Demographics when we use the TriadFG model; TriadFG-SD denotes ignoring Social Information and Demographics; while TriadFG-SDV denotes also ignoring Verified Status information; and TriadFG-SDVN denotes further ignoring Network sSructure information.

Table 6.4 Computation Time.

Algorithm	Logistic	SVM	TriadFG		
			-BF	-KF	-EKF
Time	0.385	13.1075	12.302	10.356	11.044

Effects of Different Kernel Functions

Now we will see whether kernel functions will play some role in triadic closure. Specifically, we compare six different kernels: Gaussian, tophat, epanechnikov, exponential, linear, and cosine.The kernel density estimate within different kernel functions is shown in Figure 5.6. The prediction performance within different kernel functions is shown in Table 6.5. We can see from the table that all the kernel functions performs almost the same, but the Gaussian kernel function performs slightly better than the others in terms of F1-measure.

Effects of Training Sets Ratios

Now we will see whether the ratio of training sets plays some role in triadic closure. Specifically, we compare five different training sets: 66%, 50%, 40%, 30%, and 20%. The prediction performance within different training sets is shown in Table 6.6. We can see from the table that TriadFG-KF is very

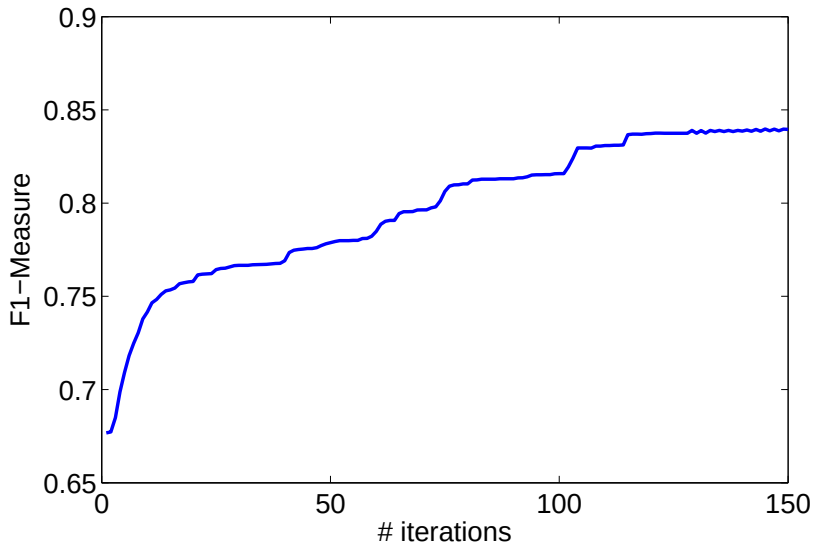


Fig. 6.3 Convergence analysis of the learning algorithm.

sensitive to the ratio of training sets, while other algorithms are insensitive to the ratio of training sets.

Prediction on Specific Users

Now we will see prediction performance with different types of users. Specifically, we select two types of users: verified users and popular users. The prediction performance is shown in Table 6.7. We can see from the table that for specific type of users, especially for popular users, the prediction performance is much better than that for random ones. This may be partly because for these specific users, we get enough features for training, which is quite different from the random case in which some feature functions may be sparse.

Effects of Dataset Size

In order to verify whether dataset size would influence our observations, we sample several subsets. In the subset, we random select 5 users, and then their followees and followees' followees, from our original dataset. We repeat this process 3 times and get 3 sample subsets.

We repeat our observations as in Chapter 3.2, and find most observations are consistent, as the observations of the subset are similar to those of the full dataset. For example, for location distribution, the distribution of three samples is shown in Figure 6.4(a). When compared with Figure 3.5(a), we can see that the three observations have almost the same pattern. We also check the other

Table 6.5 Triadic closure prediction performance within different kernel functions.

Method	Kernel	Accuracy	Precision	Recall	F1-score
TriadFG -KF	Gaussian	0.8426	0.8102	0.8613	**0.8482**
	Tophat	0.8370	0.8308	0.8548	0.8420
	Epanechnikov	0.8418	0.8343	0.8594	0.8463
	Exponential	0.8430	0.8386	0.8551	0.8464
	Linear	0.8370	0.8311	0.8537	0.8418
	Cosine	0.8384	0.8307	0.8594	0.8446
TriadFG -EKF	Gaussian	0.8444	0.8360	0.9084	**0.8564**
	Tophat	0.8307	0.7894	0.9108	0.8457
	Epanechnikov	0.8385	0.7994	0.9106	0.8513
	Exponential	0.8450	0.8103	0.9063	0.8556
	Linear	0.8342	0.7961	0.9070	0.8478
	Cosine	0.8348	0.7975	0.9078	0.8490

Table 6.6 Triadic closure prediction performance within different training sets.

Method	66%	50%	40%	30%	20%
Logistic	0.7316	0.733	0.734	0.738	0.708
SVM	0.7344	0.736	0.736	0.740	0.740
TriadFG-BF	0.7890	0.7819	0.7799	0.7778	0.7831
TriadFG-KF	0.8482	0.6014	0.2568	0.0466	0.0067
TriadFG-EKF	0.8564	0.8501	0.8482	0.8505	0.8515

observations and find nearly the same characteristics. Due to space limitations, we omitted the detailed statistics here.

Effects of Crawling Bias

It has been empirically observed that incomplete BFS is biased toward high-degree nodes, which may affect the measurements. In this section, we will check whether this crawling method will affect our observations. We select weights of random samples from the crawled users, where selection probability is inversely correlated to their in-degree.

We select three random samples, and repeat our observations as in Chapter 3.2. We find most observations are consistent, as the observations of the subset are similar to those of our crawled dataset. For example, for location distribution, the distribution of three samples is shown in Figure 6.4(b). Comparing them with Figure 3.5(a), we can see that the three observations have almost the same pattern.

Table 6.7 Triadic closure prediction performance on specific users.

User Type	Methods	Accuracy	Precision	Recall	F1-score
Verified Users	SVM	0.854	0.861	0.854	0.852
	Logistic	0.855	0.861	0.855	0.853
	TriadFG-BF	0.8355	0.788	0.9416	0.858
	TriadFG-KF	0.7655	0.7021	0.9655	0.813
	TriadFG-EKF	0.8815	0.8556	0.933	0.8926
Popular Users	SVM	0.993	0.993	0.993	0.993
	Logistic	0.993	0.993	0.993	0.993
	TriadFG-BF	0.9979	0.9958	1.0000	0.9979
	TriadFG-KF	0.9967	0.9935	1.0000	0.9967
	TriadFG-EKF	0.9981	0.9962	1.0000	0.9981

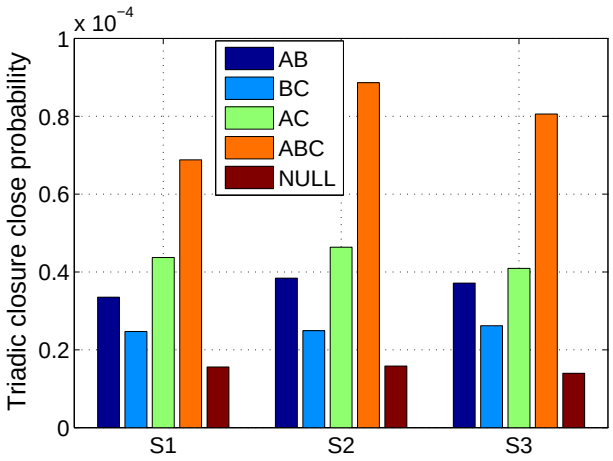
Table 6.8 Triadic closure prediction performance with one-day-timestamp

Algorithm	Accuracy	Precision	Recall	F1-score
Logistic	0.730	0.761	0.730	0.722
SVM	0.732	**0.770**	0.732	0.723
TriadFG-BF	0.6932	0.6275	0.9537	0.7570
TriadFG-KF	0.5781	0.5430	**0.9959**	0.7028
TriadFG-EKF	**0.7630**	0.7029	0.9127	**0.7942**

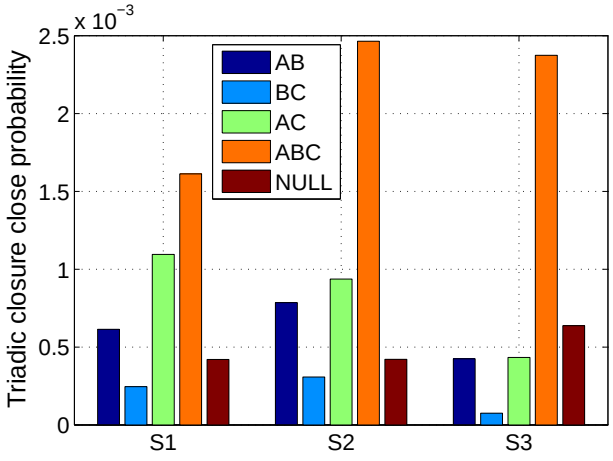
Although distributions are slightly different for some correlations, they do not affect our prediction performance, since our TriadFG-model leverages the feature functions that qualify the similarities between features to do prediction. In this sense, we can draw obvious features from our observations, regardless of observations of crawled samples or bias-corrected sub-samples, since there are clear differences between them.

Effects of Timestamp

In our sampled data, the network is dynamic. On average, there are around 6 billion new open triads every day. If we consider each new tie as a new event, the computation cost will increase beyond our computation capability, so we use a time window – every four days as a timestamp period. In order to see the effects of timestamp periods, here we choose one day as a timestamp period to predict triadic closure. The performance is shown in Table 6.8, which shows that the prediction performance is much worse than when we use four days as a timestamp period. One possible explanation could be that when the time window is smaller, the correlation factors of each triad are less independent, thus making the prediction performance worse.



(a) dataset size correlation



(b) crawl bias correlation

Fig. 6.4 Location correlation for samples. X-axis: Samples. Y-axis: probability that triadic closure occurs. Expressions attached to each bar indicate whether certain users are from the same city – e.g., *AB* represents only *A*, *B* are in the same city. *NULL* means users in a triad all come from different cities.

Qualitative Case Study

Now we present a case study to demonstrate the effectiveness of the proposed model. Figure 6.5 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 timestamp. In Figure 6.5(b), green dash lines indicate the

links are predicted by SVM however will not connect in the next timestamp. The red dash lines indicate the link will form in the next timestamp, but not predicted by SVM. In Figure 6.5(c), the green dash lines indicate the links are predicted by our approach however will not connect in the next timestamp.

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.

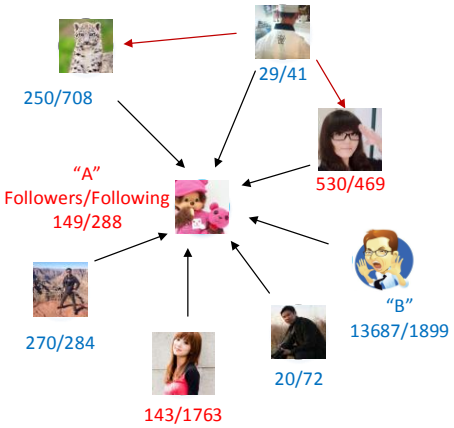
Comparison with Twitter Observations

We compare the results with a similar study about popularity within triads on Twitter [52] and find:

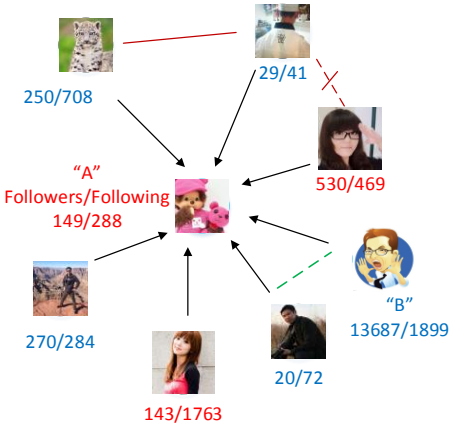
- Both results demonstrate the phenomenon of "the rich get richer" – i.e., $P(1XX) > P(0XX)$, which validates the mechanism of preferential attachment in both networks (Twitter and Weibo).

- 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, in China, Weibo is a combination of Twitter and Facebook, and integrates the features of both, which better helps users interact with each other, and ordinary users have more chances to connect with others.

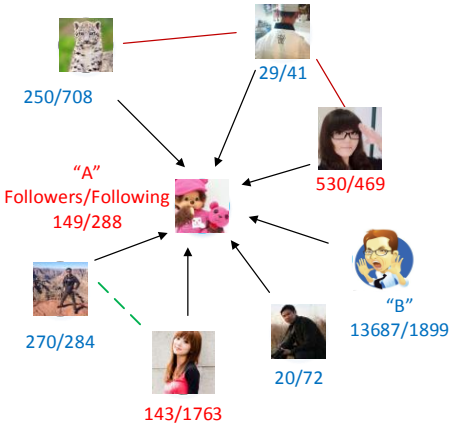
- The probability $P(111)$ for popular users in Weibo is much higher than that in Twitter. In Twitter, $P(111)$ is twice as high as $P(000)$; while in Weibo, $P(111)$ is eight times as high, which implies that popular users in China have more closeness connections.



(a) Ground Truth



(b) SVM



(c) Our approach (TriadFG)

Fig. 6.5 Case study for triadic closure prediction.

6.2 Triadic Tie Strength Prediction

6.2.1 Experiment Setup

Settings

We use the dataset described in Chapter 4.1 in our experiments. The task is to predict whether ties in the candidate triads that become closed at t will become weakened within a timeframe Δt. We set up two prediction cases (Weibo1 and Weibo2) using the Weibo dataset and one case (Mobile) using the Mobile dataset. In Weibo1, we use a network over the first 10 days of the window for experiments ($t = 5$ and $\Delta t = 5$) and predict the dynamics status in day t' ($t' = 10$, $\Delta t = 5$) of newly closed triads (formed on the 5^{th} day). In Weibo2, we incrementally observe the network and use the first 20 days for prediction experiments ($t = 10$ and $\Delta t = 10$), predicting the dynamics status on day t' ($t' = 20$ and $\Delta t = 10$) of closed triads (formed on the 10^{th} day). Using the Mobile dataset, we construct an experimental case Mobile by using the same setting as that of the Weibo1 case. Similar to the observations in Chapter 4.2, for each prediction case, we use three different measures to determine the dynamics status of social triads, including the interaction frequencies of both links IF_{AB+BC} together, and also these two links separately IF_{AB} and IF_{BC}.

Evaluation

We use 50% of the observed closed triads as a training set and the remaining triads as a test set. We repeat the prediction experiments ten times, and report the average performance in terms of Accuracy, Precision, Recall, and F1-score.

Comparisons

We compare the proposed TRIST model with four classical classification baselines, including logistic regression, support vector machine (SVM), decision tree (C4.5), and Naïve Bayes. The TRIST model is 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. As to the baseline methods, we employ the open-source software Weka with default parameters.

SVM. It uses all the same features defined in the attribute factors of our TRIST model for each triad to train a corresponding classification model, and then use the classification model to predict triadic tie strength dynamics in the test data.

Logistic Regression. It uses all the same features defined in the attribute factors of our TRIST model for each triad to train a corresponding classification model, and then use the classification model to predict triadic tie strength dynamics in the test data.

Decision Tree. It uses all the same features defined in the attribute factors of our TRIST model for each triad to train a corresponding classification model, and then use the classification model to predict triadic tie strength dynamics in the test data.

Naïve Bayes. It uses all the same features defined in the attribute factors of our TRIST model for each triad to train a corresponding classification model, and then use the classification model to predict triadic tie strength dynamics in the test data.

TRIST. It represents the proposed model that trains a factor graph model with all three types of factors – attribute factors, social factors, and temporal factors.

TRIST-ST. It uses all the same features defined in the attribute factors of our TRIST model for each triad. The difference lies in that TRIST-ST leverages kernel-based attribute factors to capture the feature correlations.

TRIST-A, TRIST-T, and **TRIST-S.** They are three reduced versions of our TRIST model that ignore the corresponding factors.

6.2.2 Prediction Results

Performance

Table 6.9, Table 6.10 and Table 6.11 report the prediction performance of triadic tie strength dynamics as measured by interaction frequencies IF_{AB+BC}, IF_{AB}, and IF_{BC} respectively. Generally, the results in both Weibo and Mobile show that our TRIST model clearly outperforms the baseline methods in each case.

In terms of F1-score, the proposed TRIST model achieves a $(25\%, 44\%)$ improvement compared with baselines when triadic tie strength dynamics is measured by IF_{AB+BC}, a $(16\%, 44\%)$ improvement when measured by IF_{AB}, and a $(26\%, 42\%)$ improvement when measured by IF_{BC}. Similarly, the three tables show that the proposed TRIST model also outperforms other methods significantly in terms of other evaluation metrics – Precision, Recall, and Accuracy.

On average, by using the same prediction settings, the results in the Weibo1 prediction case in Table 6.9 reveal about $(10\%, 15\%)$ greater predictability of

Table 6.9 Prediction performance for IF_{AB+BC}.

Data	Method	Accu.	Prec.	Rec.	F1
Weibo1	Logistic	0.538	0.541	0.538	0.538
	SVM	0.540	0.547	0.540	0.537
	Decision Tree	0.527	0.530	0.527	0.527
	NaiveBayes	.535	.538	.535	.535
	TRIST	**0.844**	**0.940**	**0.756**	**0.838**
	TRIST-ST	0.541	0.559	0.579	0.569
Weibo2	Logistic	0.535	0.534	0.535	0.534
	SVM	0.537	0.540	0.537	0.537
	Decision Tree	0.532	0.532	0.532	0.532
	NaiveBayes	0.545	0.544	0.545	0.544
	TRIST	**0.809**	**0.889**	**0.723**	**0.797**
	TRIST-ST	0.540	0.562	0.599	0.580
Mobile	Logistic	0.601	0.558	0.601	0.510
	SVM	0.605	0.366	0.605	0.456
	Decision Tree	0.605	0.366	0.605	0.456
	NaiveBayes	0.604	0.365	0.604	0.455
	TRIST	**0.750**	**0.738**	**0.911**	**0.815**
	TRIST-ST	0.604	0.614	0.937	0.742

Table 6.10 Prediction performance for IF_{AB}.

Data	Method	Accu.	Prec.	Rec.	F1
Weibo1	Logistic	0.557	0.595	0.557	0.403
	SVM	0.557	0.632	0.557	0.402
	Decision Tree	0.553	0.545	0.553	0.438
	NaiveBayes	0.556	0.563	0.556	0.405
	TRIST	**0.859**	**0.944**	**0.756**	**0.840**
	TRIST-ST	0.554	0.727	0.006	0.013
Weibo2	Logistic	0.575	0.634	0.575	0.424
	SVM	0.573	0.328	0.573	0.417
	Decision Tree	0.556	0.502	0.556	0.462
	NaiveBayes	0.570	0.520	0.570	0.432
	TRIST	**0.797**	**0.821**	**0.702**	**0.757**
	TRIST-ST	0.553	0.818	0.007	0.014
Mobile	Logistic	0.574	0.572	0.574	0.573
	SVM	0.568	0.560	0.568	0.478
	Decision Tree	0.570	0.563	0.570	0.479
	NaiveBayes	0.568	0.562	0.568	0.562
	TRIST	**0.705**	**0.750**	**0.714**	**0.732**
	TRIST-ST	0.573	0.618	0.630	0.626

Table 6.11 Prediction performance for IF_{BC}.

Data	Method	Accu.	Prec.	Rec.	F1
Weibo1	Logistic	0.591	0.560	0.591	0.516
	SVM	0.592	0.351	0.592	0.440
	Decision Tree	0.589	0.565	0.589	0.555
	NaiveBayes	0.566	0.551	0.566	0.554
	TRIST	**0.864**	**0.972**	**0.782**	**0.867**
	TRIST-ST	0.580	0.585	0.949	0.724
Weibo2	Logistic	0.583	0.540	0.583	0.531
	SVM	0.614	0.604	0.614	0.491
	Decision Tree	0.590	0.558	0.590	0.553
	NaiveBayes	0.574	0.566	0.574	0.569
	TRIST	**0.826**	**0.947**	**0.749**	**0.836**
	TRIST-ST	0.597	0.603	0.939	0.735
Mobile	Logistic	0.608	0.557	0.608	0.516
	SVM	0.615	0.378	0.615	0.468
	Decision Tree	0.615	0.378	0.615	0.468
	NaiveBayes	0.615	0.570	0.615	0.468
	TRIST	**0.764**	**0.756**	**0.907**	**0.824**
	TRIST-ST	0.610	0.620	0.937	0.746

triadic tie strength dynamics in terms of Accuracy than those of Mobile cases. Meanwhile, we also note that the prediction performance with a short Δt (5 days) in the Weibo1 case is better than that with a relatively long Δt (10 days) in the Weibo2 case, in terms of both Accuracy and F1-score.

The reasons that TRISToutperforms Logistic Regression, SVM, Decision Tree, and naïve Bayes come from 1) the modeling of structural correlations of triadic tie strength dynamics captured by temporal and social factors in our TRIST model, and 2) the use of kernel densities for modeling attribute features.

Factor Contribution Analysis

To predict the dynamics status of triadic relationships, we devise three kinds of factors in our proposed model TRIST. To explore the contributions of different factors to the prediction task, we remove each type of factor and keep the remaining two in a series of experiments. For this purpose, we have three versions of our model, i.e., TRIST-A (removing attribute factors), TRIST-T (removing temporal factors), and TRIST-S (removing social factors). The results of the three reduced methods and TRIST are shown in Figure 6.6. Clearly, we can see that the removal of each type of factor results in a clear

drop in the prediction performance. By removing social factors, the F1-score of TRIST-S model deceases slightly compared to the original model TRIST in all prediction cases. However, the drops of F1-score when removing attribute or temporal factors (TRIST-A or TRIST-T) are much more significant than TRIST-S's performance drops, which indicates the importance of attribute and temporal factors in predicting triadic tie strength dynamics. Specifically, TRIST-A achieves better prediction performance than TRIST-T, which means that temporal factors are more telling than attribute factors for predicting the dynamics status of social triads in TRIST model. These experimental results further demonstrate the effectiveness of our TRIST model by modeling the attribute, temporal, and social correlations examined in Chapter 4.2.

6.2.3 Discussion

Training/Test Ratio

We analyze the effects of different training samples on the prediction performance of triadic tie strength dynamics. Figure 6.7 shows the prediction performance with different ratios of training samples in Weibo1, Weibo2, and Mobile. First, we observe that as training ratios increase, the TRIST model gradually achieves better performance in terms of both F1-score and Accuracy and reaches a stable performance when using only 5% to 10% training samples. Second, we can see that the results in different prediction cases show similar trends to those obtained by increasing the number of training samples, in three sub-figures. The results indicate a positive effect of the size of training data on predicting triadic tie strength dynamics in social networks. At the same time, we also conclude that the predictability of triadic tie strength dynamics can be largely revealed by a small set of labeled triads.

Convergence

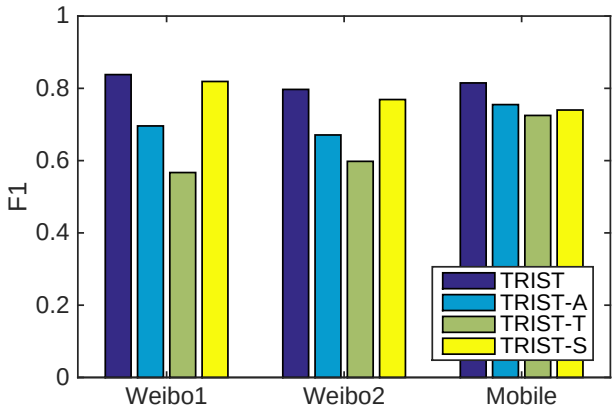
We conduct experiments to see the convergence of our TRIST model – the number of iterations of our learning algorithm. The convergence properties of TRIST with different kernel density functions by using the default training data (50%) are plotted in Figures 6.8 (a), (b), and (c). From Figures 6.8 (a) and (b) we observe similar convergence patterns, that is, the TRIST model gradually reaches convergence states within 50 to 60 iterations. From Figure 6.8 (c) we observe that the TRIST model immediately converges when the number of iterations approaches around 70. From the top three sub-figures

of Figure 6.8 we also notice that different kernels have limited effects on TRIST's convergence. By using different ratios of training data, we report the convergence of our proposed model with the default kernel function (Gaussian) in Figures 6.8 (d), (e), and (f). In the bottom three sub-figures of Figure 6.8, we again find that the model converges gradually in Weibo cases, and reaches convergence suddenly in Mobile case within 100 iterations. We also find that the TRIST model with 10% training data (red line) has the lowest convergence rate compared with the model with more training data. Overall, Figure 6.8 demonstrates that the proposed TRIST model can reach the convergence state quickly.

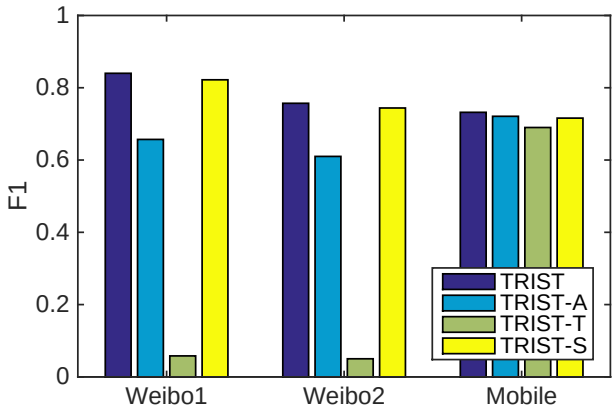
Effects of Kernels

We demonstrate the power of TRIST's kernel density by comparing the reduced version of our model TRIST-ST with baseline methods that use the same set of features. The difference lies in that in TRIST-ST, the kernel density is incorporated to smooth the discrete feature values. In Table 6.9 – 6.11, we can see that in most cases, TRIST-ST yields better prediction performance than the alternative methods. For example, by our methodology, the predictive power of TRIST-ST significantly outperforms the other four methods, with a $(5\%, 23\%)$ increase of F1-score in the Mobile case.

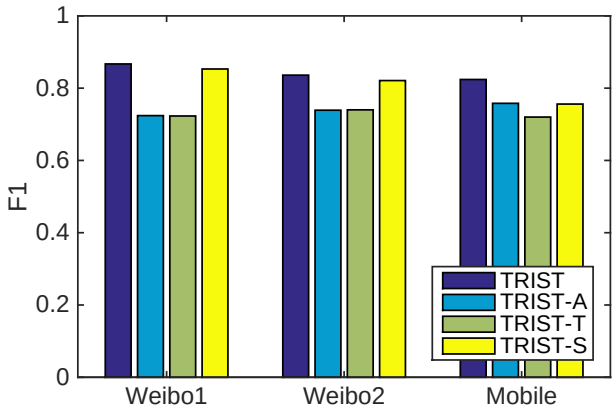
We also examine the effects of different kernel density functions on the prediction power of our method. Specifically, we use six different kernels: Gaussian, Tophat, Epanechnikov, exponential, linear, and cosine. The details of these six kernels are introduced in Chapter 5.3.1. The prediction results on triadic tie strength dynamics as measured by IF_{AB+BC} in Weibo1, Weibo2, and Mobile cases are shown in Figure 6.9. Clearly, we can see that the TRIST model with different kernel functions yields similar prediction performance. We conclude that our TRIST model is robust with respect to the choice of different kernel functions.



(a) IF_{AB+BC}



(b) IF_{AB}



(c) IF_{BC}

Fig. 6.6 Factor contributions for triadic tie strength dynamics Prediction. *x*-axis: different prediction cases; *y*-axis: Prediction performance in terms of F1-score. TRIST is the proposed model; TRIST-A is the reduced version of TRIST without modeling attributed factors; TRIST-T is the reduced version of TRIST without modeling temporal factors; TRIST-S is the reduced version of TRIST without modeling social factors.

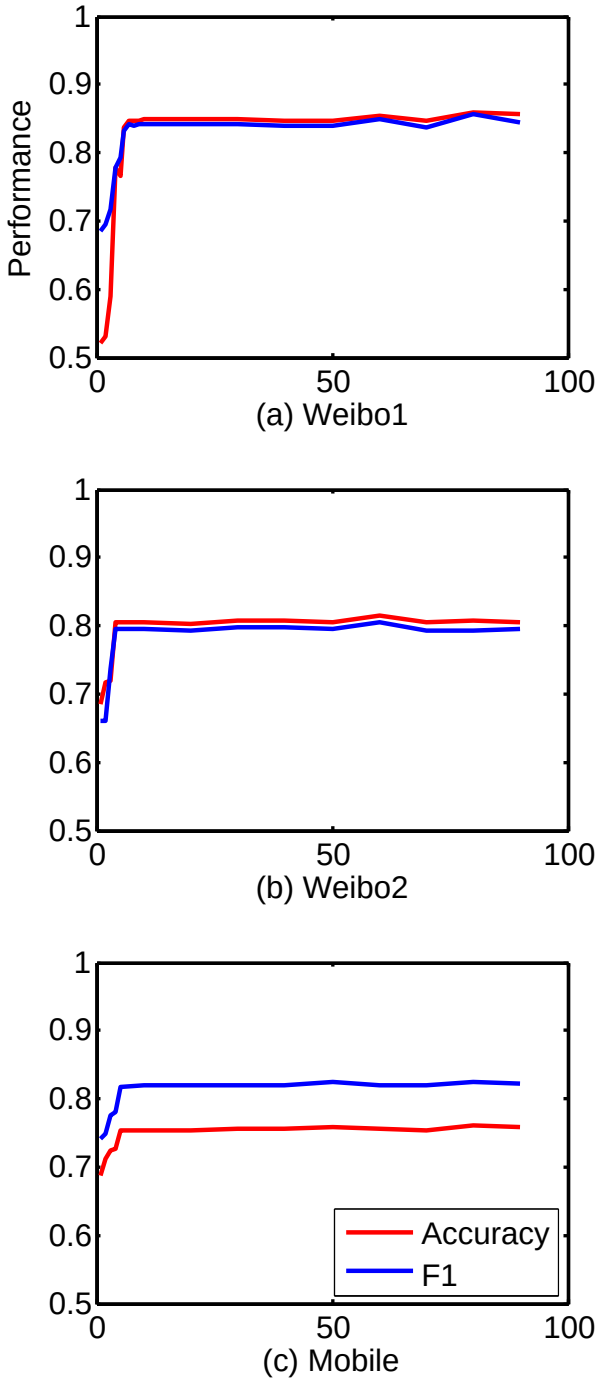


Fig. 6.7 Performance of triadic tie strength dynamics prediction with different percentages of training data. x-axis: different ratios of training set; y-axis: prediction performance in terms of both F1-score and Accuracy. The experiments are conducted for predicting triadic tie strength dynamics as measured by IF_{AB+BC} when using TRIST model in Weibo1 (a), Weibo2 (b), and Mobile (c) cases.

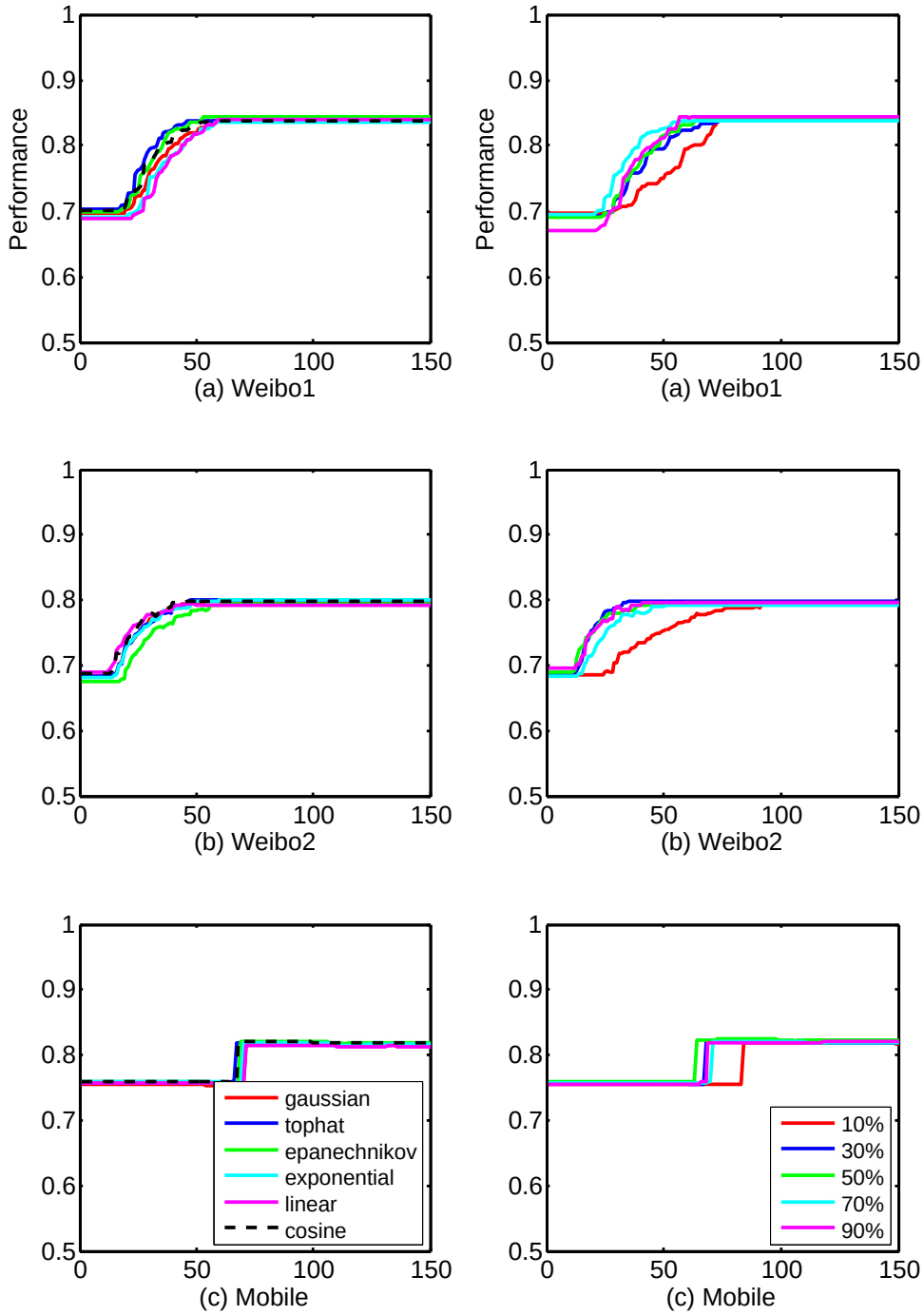


Fig. 6.8 Convergence of TRIST. x-axis: the iteration time m in Eq. 5.22 of our learning algorithm; y-axis: F1-score. (a,b,c). Convergence rates with different kernel functions in different prediction cases (50% training data); (d,e,f). Convergence rates with different training data in different prediction cases (Gaussian kernel).

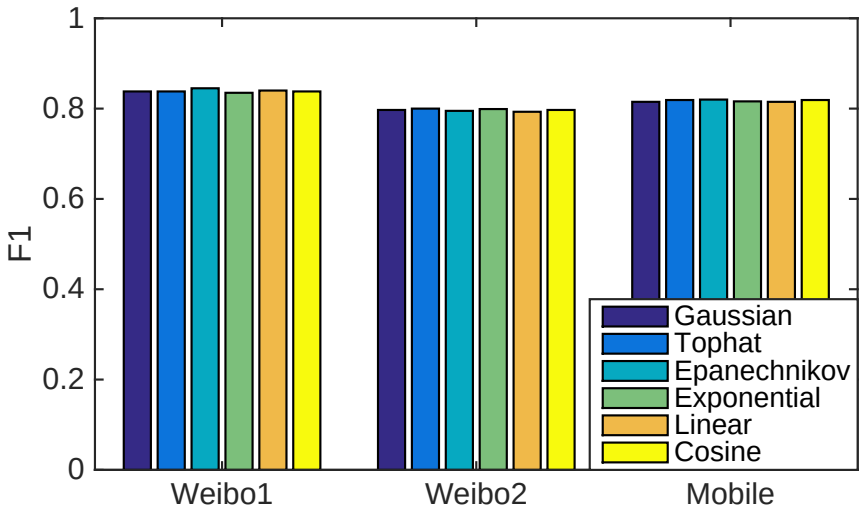


Fig. 6.9 Effects of different kernel functions in TRIST. X-axis: Different test cases. Y-axis: F1-score. The TRIST model with different kernel functions yields similar prediction performance.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this thesis, we study the network structure from a microscopic view, particularity, triadic closure and its influence in social networks. We aim to systematically understand the whole process during and after triadic closure. Starting from open triads, we trace triadic closure formation in dynamic social networks, and then after triadic closure, we investigate the triadic closure's influence, in other words, will triadic closure influence the tie strength of a triad over time.

Employing a large microblogging network (Weibo) as the source in our study, we first focus on open triads closing process. By investigating the impact of different factors from three aspects: user demographics, network characteristics, and social perspectives, we observe that 1) male, celebrity and gregarious users are more willing to close open triads; 2) structural hole spanners are eager to close open triads for more social resources, but they are also reluctant to have two disconnected friends linked together.

After triadic closure, we trace the interaction dynamics within a triad in two social networks – microblogging and mobile networking – we find that the formation of the third link in a triad will demote the interaction strength of the other two links. In around 80% of closed social triads, the strength of the first two ties become weakened in both online social media and mobile social networks. We also uncover an interesting phenomenon, that is, both males and celebrities tend to maintain more weakened triadic relationships than females and ordinary users, although the latter are more likely to form closed triads.

We formally define a prediction problem for triadic closure. We propose a probabilistic factor model – TriadFG for modeling and predicting whether

three persons in a social network will finally form a triad. In this model, we consider attribute correlations such as user demographics and social correlations such as popularity. With better instantiating attribute factors, we also extend TriadFG model with kernel density estimates. Furthermore, in order to predict tie strength dynamics, we formalize the triadic tie strength dynamics prediction problem, and present a TRIST model to solve it by incorporating user demographics and temporal together with structural correlations.

Extensive experimental results show that our proposed models outperform benchmark methods by up to 30% in terms of F1-score. In addition, we demonstrate that our methodology offers a greater-than-82% potential predictability for inferring the dynamics status of social triads in both networks. Also, with the kernel density estimate, our models outperform baselines by up to 30% in terms of F1-score.

7.2 Future Work

We share some of our thoughts on the future work.

Firstly, we can explore the underlying mechanism of the evolution of triadic relationships. In this thesis, we consider one special case that the open triad with two reciprocal relationships. However, the tie strength dynamics in other triads is not well studied. Unlike the triad with reciprocal relationships, the triads with mixed parasocial and reciprocal relationships are far more complicated due to the direction of links. In general, there are 27 types of triad, each of which has its own evolving patterns.

What's more, the interplay among ties over time can be better examined. Few work has been done to explore how one tie influences the others. Although we have discovered certain correlations between the established third tie and existing ties, we are unable to quantitatively measure how much this influence is. With the knowledge of underlying influence mechanises, we can better understand network evolution.

Secondly, we need to connect the microscopic triadic principles with network scaling phenomena at the macro level. We would attempt to explain macro characteristics on the basis of aggregate effects of the micro characteristics, interpret social phenomena from the behavior of rational actors at the micro level, and identify a set of important micro-macro linkages between local behavior in networks and global network properties.

Finally, it is also necessary to examine the dynamics status of social triads in other types of networks, such as collaboration networks, organization networks, location-based social networks, finance networks and so on. It is worthwhile to explore these networks with our triadic closure models to obtain more insights from them. For example, tracing interaction dynamics within bankers can prevent financial crisis, and leveraging the understanding of interaction patterns among employees will achieve better organization performance.

Bibliography

[1] Ahuja, G. (2000). Collaboration networks, structural holes, and innovation: A longitudinal study. *Administrative science quarterly*, 45(3):425–455.

[2] Aral, S. and Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341.

[3] Bach, F. R. (2008). Graph kernels between point clouds. In *Proceedings of the 25th international conference on Machine learning*, pages 25–32. ACM.

[4] Backstrom, L., Huttenlocher, D., Kleinberg, J., and Lan, X. (2006). Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54. ACM.

[5] Backstrom, L. and Leskovec, J. (2011). Supervised random walks: predicting and recommending links in social networks. In *WSDM'11*, pages 635–644.

[6] Bahrami, B., Olsen, K., Bang, D., Roepstorff, A., Rees, G., and Frith, C. (2012). Together, slowly but surely: the role of social interaction and feedback on the build-up of benefit in collective decision-making. *Journal of Experimental Psychology: Human Perception and Performance*, 38(1):3.

[7] Bar-Yossef, Z., Kumar, R., and Sivakumar, D. (2002). Reductions in streaming algorithms, with an application to counting triangles in graphs. In *Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 623–632. Society for Industrial and Applied Mathematics.

[8] Becchetti, L., Boldi, P., Castillo, C., and Gionis, A. (2008). Efficient semi-streaming algorithms for local triangle counting in massive graphs. In *KDD'08*, pages 16–24.

[9] Becchetti, L., Boldi, P., Castillo, C., and Gionis, A. (2010). Efficient algorithms for large-scale local triangle counting. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(3):13.

[10] Ben-Hur, A. and Noble, W. S. (2005). Kernel methods for predicting protein–protein interactions. *Bioinformatics*, 21(suppl 1):i38–i46.

[11] Benevenuto, F., Rodrigues, T., Cha, M., and Almeida, V. (2009). Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 49–62. ACM.

[12] Bianconi, G., Darst, R. K., Iacovacci, J., and Fortunato, S. (2014). Triadic closure as a basic generating mechanism of communities in complex networks. *Physical Review E*, 90(4):042806.

[13] Buriol, L. S., Frahling, G., Leonardi, S., Marchetti-Spaccamela, A., and Sohler, C. (2006). Counting triangles in data streams. In *Proceedings of the twenty-fifth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 253–262. ACM.

[14] Burke, M. and Kraut, R. E. (2014). Growing closer on facebook: changes in tie strength through social network site use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 4187–4196. ACM.

[15] Burt, R. S. (1993). The social structure of competition. *Explorations in economic sociology*, 65:103.

[16] Burt, R. S. (2004). Structural holes and good ideas1. *American journal of sociology*, 110(2):349–399.

[17] Burt, R. S. (2007). Secondhand brokerage: Evidence on the importance of local structure for managers, bankers, and analysts. *Academy of Management Journal*, 50(1):119–148.

[18] Caplow, T. (1968). Two against one: Coalitions in triads.

[19] Cartwright, D. and Harary, F. (1956). Structural balance: a generalization of heider's theory. *Psychological review*, 63(5):277.

[20] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P. K. (2010). Measuring user influence in twitter: The million follower fallacy. *ICWSM'10*, 10:10–17.

[21] Coleman, J. (1990). *Foundations of Social Theory*. Harvard.

[22] Davis, J. A. and Leinhardt, S. (1972). The structure of positive interpersonal relations in small groups. In Berger, J., editor, *Sociological Theories in Progress*, volume 2, pages 218–251. Houghton Mifflin.

[23] De Choudhury, M., Diakopoulos, N., and Naaman, M. (2012). Unfolding the event landscape on twitter: classification and exploration of user categories. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 241–244. ACM.

[24] DiMicco, J., Millen, D. R., Geyer, W., Dugan, C., Brownholtz, B., and Muller, M. (2008). Motivations for social networking at work. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 711–720. ACM.

[25] Dong, Y., Tang, J., Chawla, N. V., Lou, T., Yang, Y., and Wang, B. (2015). Inferring social status and rich club effects in enterprise communication networks. *PLoS ONE*, 10:e0119446.

[26] Dong, Y., Tang, J., Wu, S., Tian, J., Chawla, N. V., Rao, J., and Cao, H. (2012). Link prediction and recommendation across heterogeneous social networks. In *ICDM '12*, pages 181–190.

[27] Dong, Y., Yang, Y., Tang, J., Yang, Y., and Chawla, N. V. (2014). Inferring user demographics and social strategies in mobile social networks. In *KDD '14*, pages 15–24. ACM.

[28] Doroud, M., Bhattacharyya, P., Wu, S. F., and Felmlee, D. (2011). The evolution of ego-centric triads: A microscopic approach toward predicting macroscopic network properties. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 172–179. IEEE.

[29] Easley, D. and Kleinberg, J. (2010). Networks, crowds, and markets. *Cambridge Univ Press*, 6(1):6–1.

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

[31] Eom, Y.-H., Lee, S., and Jeong, H. (2006). Exploring local structural organization of metabolic networks using subgraph patterns. *Journal of Theoretical Biology*, 241(4):823–829.

[32] Fang, Z. and Tang, J. (2015). Uncovering the formation of triadic closure in social networks. In *IJCAI'15*, pages –.

[33] Foster, D. V., Foster, J. G., Grassberger, P., and Paczuski, M. (2011). Clustering drives assortativity and community structure in ensembles of networks. *Physical Review E*, 84(6):066117.

[34] Frank, O. (1979). Sampling and estimation in large social networks. *Social networks*, 1(1):91–101.

[35] Gao, Q., Abel, F., Houben, G.-J., and Yu, Y. (2012). A comparative study of users' microblogging behavior on sina weibo and twitter. In *User Modeling, Adaptation, and Personalization*, pages 88–101.

[36] Gerber, M. S. (2014). Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115–125.

[37] Getoor, L. and Diehl, C. P. (2005). Link mining: a survey. *ACM SIGKDD Explorations Newsletter*, 7(2):3–12.

[38] Ghahramani, Z. and Jordan, M. I. (1997). Factorial hidden markov models. *Machine learning*, 29(2-3):245–273.

[39] Gilbert, E. (2012). Predicting tie strength in a new medium. In *CSCW '12*, pages 1047–1056. ACM.

[40] Gilbert, E. and Karahalios, K. (2009). Predicting tie strength with social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 211–220. ACM.

[41] Golder, S. A. and Yardi, S. (2010). Structural predictors of tie formation in twitter: Transitivity and mutuality. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 88–95. IEEE.

[42] Gong, N. Z., Xu, W., Huang, L., Mittal, P., Stefanov, E., Sekar, V., and Song, D. (2012). Evolution of social-attribute networks: measurements, modeling, and implications using google+. In *IMC'12*, pages 131–144.

[43] Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological theory*, 1(1):201–233.

[44] Granovetter, M. (2005). The impact of social structure on economic outcomes. *The Journal of economic perspectives*, 19(1):33–50.

[45] Granovetter, M. S. (1995). *Getting a job: A study of contacts and careers*. University of Chicago Press.

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

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

[48] Haykin, S. S., Haykin, S. S., and Haykin, S. S. (2001). *Kalman filtering and neural networks*. Wiley Online Library.

[49] Heider, F. (1946). Attitudes and cognitive organization. *The Journal of psychology*, 21(1):107–112.

[50] Heider, F. (1958). *The psychology of interpersonal relations*. Wiley.

[51] Holland, P. W. and Leinhardt, S. (1971). Transitivity in structural models of small groups. *Comparative Group Studies*.

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

[53] Huang, H., Tang, J., Liu, L., Luo, J., and Fu, X. (2015). Triadic closure pattern analysis and prediction in social networks. *IEEE Transactions on Knowledge and Data Engineering*, 27(12):3374–3389.

[54] Huang, H., Tang, J., Wu, S., Liu, L., et al. (2014). Mining triadic closure patterns in social networks. In *WWW'14*, pages 499–504.

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

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

[57] Jha, M., Seshadhri, C., and Pinar, A. (2013). A space efficient streaming algorithm for triangle counting using the birthday paradox. In *KDD'13*, pages 589–597.

[58] Jin, L., Chen, Y., Wang, T., Hui, P., and Vasilakos, A. V. (2013). Understanding user behavior in online social networks: A survey. *IEEE Communications Magazine*, 51(9):144–150.

[59] Jones, J. J., Settle, J. E., Bond, R. M., Fariss, C. J., Marlow, C., and Fowler, J. H. (2013). Inferring tie strength from online directed behavior. *PloS one*, 8(1):e52168.

[60] Jowhari, H. and Ghodsi, M. (2005). New streaming algorithms for counting triangles in graphs. In *International Computing and Combinatorics Conference*, pages 710–716. Springer.

[61] Juszczyszyn, K., Musial, K., and Budka, M. (2011). Link prediction based on subgraph evolution in dynamic social networks. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 27–34. IEEE.

[62] Kahanda, I. and Neville, J. (2009). Using transactional information to predict link strength in online social networks. *ICWSM*, 9:74–81.

[63] Kairam, S. R., Wang, D. J., and Leskovec, J. (2012). The life and death of online groups: Predicting group growth and longevity. In *WSDM '12*, pages 673–682. ACM.

[64] Kane, D. M., Mehlhorn, K., Sauerwald, T., and Sun, H. (2012). Counting arbitrary subgraphs in data streams. In *International Colloquium on Automata, Languages, and Programming*, pages 598–609. Springer.

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

[66] Khanafiah, D. and Situngkir, H. (2004). Social balance theory: revisiting heider's balance theory for many agents.

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

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

[69] Krackhardt, D. (1999). The ties that torture: Simmelian tie analysis in organizations. *Research in the Sociology of Organizations*, 16(1):183–210.

[70] Krackhardt, D. and Handcock, M. S. (2007). Heider vs simmel: Emergent features in dynamic structures. In *Statistical Network Analysis: Models, Issues, and New Directions*, pages 14–27. Springer.

[71] Krackhardt, D., Nohria, N., and Eccles, B. (1992). The strength of strong ties.

[72] Kschischang, F. R., Frey, B. J., and Loeliger, H. A. (2001). Factor graphs and the sum-product algorithm. *IEEE TOIT*, 47:498–519.

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

[74] Lafferty, J., McCallum, A., and Pereira, F. C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML'01*, pages 282–289.

[75] Leskovec, J., Backstrom, L., Kumar, R., and Tomkins, A. (2008). Microscopic evolution of social networks. In *KDD'08*, pages 462–470.

[76] Leskovec, J. and Horvitz, E. (2008). Planetary-scale views on a large instant-messaging network. In *WWW '08*, pages 915–924. ACM.

[77] Leskovec, J., Huttenlocher, D., and Kleinberg, J. (2010a). Predicting positive and negative links in online social networks. In *WWW'10*, pages 641–650.

[78] Leskovec, J., Huttenlocher, D., and Kleinberg, J. (2010b). Signed networks in social media. In *CHI'10*, pages 1361–1370.

[79] Leskovec, J., Lang, K. J., Dasgupta, A., and Mahoney, M. W. (2009). Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123.

[80] Li, L., Goodchild, M. F., and Xu, B. (2013a). Spatial, temporal, and socioeconomic patterns in the use of twitter and flickr. *cartography and geographic information science*, 40(2):61–77.

[81] Li, M., Zou, H., Guan, S., Gong, X., Li, K., Di, Z., and Lai, C.-H. (2013b). A coevolving model based on preferential triadic closure for social media networks. *Scientific reports*, 3.

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

[83] Lichtenwalter, R. N., Lussier, J. T., and Chawla, N. V. (2010). New perspectives and methods in link prediction. In *KDD'10*, pages 243–252.

[84] Lim, Y. and Kang, U. (2015). Mascot: Memory-efficient and accurate sampling for counting local triangles in graph streams. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 685–694. ACM.

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

[86] Lou, T., Tang, J., Hopcroft, J., Fang, Z., and Ding, X. (2013). Learning to predict reciprocity and triadic closure in social networks. *TKDD*, 7(2):5.

[87] Maia, M., Almeida, J., and Almeida, V. (2008). Identifying user behavior in online social networks. In *Proceedings of the 1st workshop on Social network systems*, pages 1–6. ACM.

[88] Mangan, S., Zaslaver, A., and Alon, U. (2003). The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *Journal of molecular biology*, 334(2):197–204.

[89] Marsden, P. V. and Campbell, K. E. (1984). Measuring tie strength. *Social forces*, 63(2):482–501.

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

[91] McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444.

[92] Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., and Alon, U. (2004). Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542.

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

[94] Mislove, A., Viswanath, B., Gummadi, K. P., and Druschel, P. (2010). You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 251–260. ACM.

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

[96] Newman, M. E. (2003). Properties of highly clustered networks. *Physical Review E*, 68(2):026121.

[97] Newman, M. E. and Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E*, 68(3):036122.

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

[99] Onnela, J. P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., and Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *PNAS*.

[100] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. *Stanford InfoLab*.

[101] Pagh, R. and Tsourakakis, C. E. (2012). Colorful triangle counting and a mapreduce implementation. *Information Processing Letters*, 112(7):277–281.

[102] Patil, A., Liu, J., and Gao, J. (2013). Predicting group stability in online social networks. In *WWW'13*, pages 1021–1030.

[103] Pennacchiotti, M. and Popescu, A.-M. (2011). A machine learning approach to twitter user classification. *ICWSM*, 11(1):281–288.

[104] Podolny, J. M. and Baron, J. N. (1997). Resources and relationships: Social networks and mobility in the workplace. *American sociological review*, pages 673–693.

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

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

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

[108] Sala, A., Cao, L., Wilson, C., Zablit, R., Zheng, H., and Zhao, B. Y. (2010). Measurement-calibrated graph models for social network experiments. In *WWW'10*, pages 861–870.

[109] Saramäki, J., Leicht, E. A., López, E., Roberts, S. G. B., Reed-Tsochas, F., and Dunbar, R. I. M. (2014). Persistence of social signatures in human communication. *PNAS*, 111(3):942–947.

[110] Sasovova, Z., Mehra, A., Borgatti, S. P., and Schippers, M. C. (2010). Network churn: The effects of self-monitoring personality on brokerage dynamics. *Administrative Science Quarterly*, 55(4):639–670.

[111] Schaefer, C., Coyne, J. C., and Lazarus, R. S. (1981). The health-related functions of social support. *Journal of behavioral medicine*, 4(4):381–406.

[112] Schweizer, B. and Sklar, A. (2011). *Probabilistic metric spaces*. Courier Dover Publications.

[113] Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

[114] Shervashidze, N., Petri, T., Mehlhorn, K., Borgwardt, K. M., and Vishwanathan, S. (2009). Efficient graphlet kernels for large graph comparison. In *International conference on artificial intelligence and statistics*, pages 488–495.

[115] Simmel, G. (1950). *The sociology of georg simmel*, volume 92892. Simon and Schuster.

[116] Sintos, S. and Tsaparas, P. (2014). Using strong triadic closure to characterize ties in social networks. In *KDD'14*, pages 1466–1475.

[117] Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., and Li, P. (2011). User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1397–1405. ACM.

[118] Tan, C., Tang, J., Sun, J., Lin, Q., and Wang, F. (2010). Social action tracking via noise tolerant time-varying factor graphs. In *KDD'10*, pages 1049–1058.

[119] Tang, J., Chang, Y., and Liu, H. (2014). Mining social media with social theories: a survey. *ACM SIGKDD Explorations Newsletter*, 15(2):20–29.

[120] Toivonen, R., Onnela, J.-P., Saramäki, J., Hyvönen, J., and Kaski, K. (2006). A model for social networks. *Physica A: Statistical Mechanics and its Applications*, 371(2):851–860.

[121] Vazquez, A., Dobrin, R., Sergi, D., Eckmann, J.-P., Oltvai, Z., and Barabási, A.-L. (2004). The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proceedings of the National Academy of Sciences*, 101(52):17940–17945.

[122] Vishwanathan, S. V. N., Schraudolph, N. N., Kondor, R., and Borgwardt, K. M. (2010). Graph kernels. *The Journal of Machine Learning Research*, 11:1201–1242.

[123] Wasserman, L. (2004). *All of statistics: a concise course in statistical inference*. Springer.

[124] Wasserman, S. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge University Press.

[125] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of "small-world" networks. *nature*, 393(6684):440–442.

[126] Welser, H. T., Gleave, E., Fisher, D., and Smith, M. (2007). Visualizing the signatures of social roles in online discussion groups. *Journal of social structure*, 8(2):1–32.

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

[128] Wu, S., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Who says what to whom on twitter. In *WWW'11*, pages 705–714.

[129] Wu, Z.-X. and Holme, P. (2009). Modeling scientific-citation patterns and other triangle-rich acyclic networks. *Physical Review E*, 80(3):037101.

[130] Xiang, R., Neville, J., and Rogati, M. (2010). Modeling relationship strength in online social networks. In *WWW'10*, pages 981–990.

[131] Xiaolin Shi, Lada A. Adamic, M. J. S. (2007). Networks of strong ties. *Physica A: Statistical Mechanics and its Applications*, 378:33–47.

[132] Yang, M.-H. (2002). Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods. In *FG'02*, pages 0215–0215.

[133] Yang, Y., Tang, J., Leung, C. W.-k., Sun, Y., Chen, Q., Li, J., and Yang, Q. (2015). Rain: Social role-aware information diffusion. In *AAAI*, pages 367–373.

[134] Zhang, J., Fang, Z., Chen, W., and Tang, J. (2015). Diffusion of "following" links in microblogging networks. *TKDE*.

[135] Zhang, J., Liu, B., Tang, J., Chen, T., and Li, J. (2013). Social influence locality for modeling retweeting behaviors. In *IJCAI'13*, pages 2761–2767.

[136] Zhang, J.-D. and Chow, C.-Y. (2013). igslr: personalized geo-social location recommendation: a kernel density estimation approach. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 334–343. ACM.

[137] Zheleva, E., Sharara, H., and Getoor, L. (2009). Co-evolution of social and affiliation networks. In *KDD'09*, pages 1007–1016.

[138] Zignani, M., Gaito, S., Rossi, G. P., Zhao, X., Zheng, H., and Zhao, B. Y. (2014). Link and triadic closure delay: Temporal metrics for social network dynamics. In *ICWSM'14*, pages 564–573.

List of figures

List of tables