7-2018

Exploring offline friendships on the social information network: Network characteristics, information diffusion, and tie strength

Felicia NATALI
Singapore Management University, felician.2013@phdis.smu.edu.sg

Exploring Offline Friendships on the Social Information Network: Network Characteristics, Information Diffusion, and Tie Strength

Felicia Natali

Singapore Management University
2018

Exploring Offline Friendships on the Social Information Network: Network Characteristics, Information Diffusion, and Tie Strength

by

Felicia Natali

Submitted to School of Information Systems in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Information Systems

Dissertation Committee:

Feida ZHU (Supervisor / Chair)
Associate Professor of Information Systems
Singapore Management University

Ee-peng LIM (Co-supervisor)
Professor of Information Systems
Singapore Management University

Robert J. KAUFFMAN
Professor of Information Systems
Singapore Management University

Kathleen M. CARLEY
Institute for Software Research
Carnegie Mellon University

Singapore Management University
2018

Exploring Offline Friendships on the Social Information Network: Network Characteristics, Information Diffusion, and Tie Strength

Felicia Natali

Abstract

The rapid increase in online social networking services over the last decade has presented an unprecedented opportunity to observe users' behaviour both on a societal and individual level. The insight gained from analysing such data can help foster a deeper understanding of social media users and the flow of information, while also offering valuable business applications. User relationships are among the most studied aspects of online behaviour. These relationships are not homogeneous. Past research has shown that people use social networks to both socialize and source information. Hence, different types of links – used to socialize, gain information, or both – are formed among users. While much research has focused on how users are connected online in general, it is crucial to explore how users interact with those present in offline social networks on the online social networks. Questions such as, "What would speed up the diffusion of a piece of information?" can be better answered from an integrated offline-online perspective. My thesis explores the behaviour of offline friends on the social information network in three main areas. I especially focus on social information networks, and use Twitter as a case study. In the first study, I explore and compare network characteristics on Twitter among offline friends and online friends. In the second study, I explore information diffusion in the same setting. In the last study, I investigate whether we can use the measurement of tie strength among friends on Twitter as a substitute for, or a complement to the measurement of tie strength among friends in the offline world.

Table of Contents

List of Figures

List of Tables

Acknowledgments

My deep gratitude goes to Dr. Feida Zhu, my advisor to whom I update my progress regularly. Thank you for your encouragement, patience and guidance. I would like to thank Dr. Ee-peng Lim, my co-supervisor who has given some of my works great pointers that have made these works more well-rounded and technically informed. I would also like to express a deep gratitude to Dr. Robert J. Kauffman, who has given me a detailed counsel that has increased the clarity and readability of this thesis significantly. Thank you for all your guidance and advice. To Dr. Kathleen M. Carley, whom I work with during my days at CMU, I would like to express a big thank you for all your encouragement and guidance. You and the CMU team have been a great inspiration. I am also grateful to Dr. Mark Kamlet who supervises us in CMU. Thank you for your supervision and knowledge sharing.

In addition, a deep appreciation for my siblings, all friends, and acquaintances who have made my PhD journey less lonely and isolated, and who have shared their precious knowledge with me.

Dedicated to my kind parents, who have given me the greatest love and support that anyone could ever give: *Yida Huang*, and *Libing Goh*. Thank you for listening patiently to my frustrations, and supporting me through many ups and downs.

Chapter 1

Introduction

The advent of the online social network has offered an unprecedented opportunity for researchers to study the dynamics of social interactions. Many of these studies explore the relationships between users of online social networks – a critical area of focus. At first, most of these studies assumed that online relationships were homogenous; slowly, however, more studies looked into the different relationship types present online. These studies sought to categorize users into communities on the online social network. However, the offline-online perspective does not determine the boundaries of these communities.

My study explores communities on the online social network from the offline-online perspective. It is motivated by the fact that offline relationships may offer new insights into the study of online relationships because a user's relationships on- and offline may overlap. To conceive how offline relationships may offer new insights into the study of online relationships, we need to first understand the differences between both online and offline social networks.

Online social network facilitate friendships with less physical and social boundaries. According to the National Geographic Encyclopedia[1], a physical boundary is a naturally occurring barrier between two areas. Unimpeded by means of transportation, people have more freedom to befriend others across regions, as long as

[1] https://www.nationalgeographic.org/encyclopedia/boundary/

the other parties accept their friend request. Therefore, physical boundaries are less consequential for online social networks.

Similarly, social boundaries are also less of an issue for online social networks. Social boundary is a structured system of relationships in which individuals are bound one to another by complex and ramifying ties (Cohen 1969). Therefore, social boundaries create rules or limits to identify reasonable, safe and permissible ways for different people to behave when they are involved. In the offline world, people relate and interact to one another based on social boundaries. However, this is not the case with online social networks. On online social networks, social contexts collapse (Vitak 2012; Vitak et al. 2012). Everyone is a friend by default, and they can see all your posts. One categorizes their friends manually provided that the online social network is equipped with the function to do so. For example, on Facebook, one can sort friends into different groups. Due to this particularity, information diffusion is more widespread online.

These differences are even more pronounced on a *social information network*, such as Twitter. Shi et al. (2014) defined Twitter as a *social broadcasting network*. A social broadcasting network is a social network in which the activities of socialization and information dissemination are so intermingled that the line between social networks and information networks becomes blurry. In this study we use the term social information network instead of social broadcasting network, as the word broadcast is typically reserved for media that can be viewed live, such as video or audio. Meanwhile, the word information includes all types of media, such as print media, audio, and video, as well as tweets.

Because of the differences between social information networks and offline social networks, we cannot expect the behaviour of offline friends on the social information network will precisely mimic their offline behaviour. Yet, at the same time, we also cannot expect that the actions of offline friends will be similar to those of friends who only know one another on social information networks such as Twitter, because offline friends benefit from face-to-face interactions in the real world.

Therefore, there are countless possible answers to the question, "What are the behaviors of offline friends on the social information network?", making investigation into the subject necessary.

In this thesis, we investigate the question in three areas, each of which is covered in one or two essays. Before opening the first essay, we provide literature studies on the existing research that has been done to bridge the gap between offline social networks and online social networks. In the first study (Essay 1A and Essay 1B), we compare the networking behavior of offline friends vs. online friends on Twitter. The second study (Essay 2A and Essay 2B) explores the application of offline friends on Twitter that may help the marketing and business world. Precisely, we study how offline friends vs. online friends retweet on Twitter. The last study (Essay 3) investigates whether we can use the measurement of tie strength online as a substitute for or complement to the measurement of tie strength offline. In other words, we aim to discover whether tie strength is explained differently in the offline world and on Twitter. The thesis ends with a conclusion summarizing our studies and providing further discussion on the subject.

By connecting a user's online social network and his or her offline social network, researchers will know whether they are able to rely on online social networks to substitute for offline network data, which is usually scarce and sourced from surveys that rely on imperfect recall. Moreover, more informed marketing efforts on Twitter could be made by knowing whom are likely offline friends on Twitter, and how they speed up information diffusion.

Chapter 2

Background and Literature Review

2.1 Definition of "Offline" and "Online" Friends in This Study

In this study, we investigate the behaviour of offline friends on social information networks. Throughout this study, the terms offline friends and online friends will be used frequently, and so we will define them below.

Offline friends are friends who build their friendships outside the Internet (Boase and Wellman 2006; Mesch and Talmud 2006). As we are dealing with friends on Twitter network in this study, what we define as *offline friends* are friends who are connected on a social information network, but who also know one another offline. Offline friends do not refer to the absence of an online linkage, but the existence of an offline relationship besides the presence of an online one.

Meanwhile, *online friends* are ties that are created and maintained through the Internet (Boase and Wellman 2006; Mesch and Talmud 2006). Therefore, offline friends and online friends may overlap. In this study, we define online friends differently from the previous studies to fit our research purpose. Here, online friends are connected on a social information network but are strangers offline. They are also not necessarily "friends" online, since these types of connections can include

news organizations or public figures like celebrities. So the term online friends does not imply the existence of friendship, but simply to the existence of an online linkage and the absence of any offline relationship.

2.2 Twitter as a Social Information Network

Previous studies have referred to Twitter as a social information network (Kwak et al. 2010; Shi et al. 2014).

The power of Twitter as a new information-sharing medium was widely acknowledged by the research world after the study conducted by Kwak et al. (2010). In the study, it was discovered that some characteristics of Twitter, namely effective diameter, reciprocity, and follower distribution, deviated from known characteristics of human social networks. Moreover, the power of Twitter as a medium of information diffusion was revealed through the discovery that any retweeted tweet would reach an average of 1,000 users no matter how many followers the user who sent the original tweet counted. This makes Twitter a viable social information network.

Social information network that has blurred the boundaries of social networks and information networks (Shi et al. 2014) provides a more open and malleable environment than the offline social world for people to communicate and share information. Such a different environment can lead to different behaviours among offline friends from what have been observed offline. On the other hand, additional mediums of communication outside the virtual world may generate different actions among offline friends compared to online friends.

These two potential differences in behaviour propel us to investigate the behaviours of offline friends on social information networks.

2.3 Previous Studies

Ours is not the first study that tries to bridge the gap between the offline and online worlds. Previous studies have attempted to do so in two ways: By comparing offline and online social networks, and exploring offline friendships on social networks online. Our line of work fits in with the latter.

2.3.1 Comparing the Offline Social Network and the Online Social Network

Most of the research that compares offline and the online social networks comes from the fields of psychology, sociology, and anthropology. The studies by Chan and Cheng (2004) and Antheunis et al. (2012) for example, compare the qualitative characteristics of friendships – closeness, frequency of interaction, interdependence, etc. – when answered through survey questions. Chan and Cheng (2004) compared offline and online friendships in terms of interdependence, breadth, depth, code change, understanding, commitment, and network convergence. They also analysed the differences between the two types of friendships over time. Meanwhile, Antheunis et al. (2012) compared the quality of online, offline, and mixed-mode friendships (i.e. friendships that originate online but extend to offline settings) among users of social networking sites. They also investigated the relative contribution of proximity, perceived similarity, and social attraction to the quality of each of these three types of friendships.

There are also studies that do not depend on conducting surveys, but instead crawl available network data found online. In these studies, researchers compared the structures of offline and online interaction networks. Dunbar et al. (2015) and Gonçalves et al. (2011) crawled and investigated Twitter and Facebook data, looking into whether these online networks also divided friends into layers of communications like the offline networks did.

2.3.2 Exploring Offline Friendships on the Online Social Network

The second type of work studies how offline friends behave on online social networks. Unlike the previous type of works, this type requires information on who are the offline friends on the online social network.

Some studies of this type investigated how the conduct of offline friends influenced how they acted online, particularly with regard to event-based social networks. Studies by Zuo et al. (2012) and Yin et al. (2014), for example, investigated how interactions in offline social networks correlated with those online. Other studies predicted specific offline relationships, such as those between significant others, family members, advisors and advisees, or managers and subordinates, based on online social networks. Examples of these include studies by Backstrom and Kleinberg (2014), Tang et al. (2011), and Xie et al. (2012) and through them, the effects of offline relationships on the online network structure were discovered. For example, Backstrom and Kleinberg (2014) found that mutual friends of romantic partners were not well-connected online. Meanwhile, Xie et al. (2012) discovered that random walk may explain the online network structure among offline friends.

Our study continues this line of work by,

1. Studying the behaviour of offline friends on a social information network, not an event-based social network.

2. Studying the behaviour of offline friends on the configuration of social information networks (besides those that have been explored previously).

3. Studying the behaviour of offline friends on information diffusion on a social information network.

4. Studying how tie strength is translated into offline behaviours and online behaviours.

In summary, this thesis is a comprehensive study on the behaviour of offline friends on a social information network, using Twitter as a case study.

Chapter 3

Essay 1A: Predicting Offline Friends on Twitter Using the Principles of Social Network Formation in the Offline World

3.1 Introduction

Network formation has been studied in both the offline social network and the online social network[1]. Before the emergence of the online social network, researchers investigated the offline social network. They discovered that the formation of the offline social network was characterized by a number of *dependencies*, also called *principles*. These principles were by no means arbitrarily generated but were empirically discovered or theoretically formulated in previous studies on social networks (Snijders 2011). When the online social network emerged, it was seen as a solution to the inconsistency and the high cost of procuring a large real life social networks data (Newman 2003). The principles of network formation that were previously

[1]This study is an extension of a published work *A Comparison of Fundamental Network Formation Principles between Offline and Online Friends on Twitter* (Natali and Zhu 2016). In this essay, *we* refers to the authors of the published study.

discovered in the offline social network are now studied in the online social network. Most of these studies reveal that the principles that apply to the offline social network – such as reciprocity, mutuality, preferential attachment, and homophily – also apply to the online social network (Golder and Yardi 2010; Kwak et al. 2010; Leskovec et al. 2008; Ellison et al. 2007). A provoking question then arises as to whether these similarities between the principles of offline and online network formation happen because "online social networks primarily support pre-existing social relations (Boyd and Ellison 2008)", particularly the existing offline contacts (Ellison et al. 2007).

To answer the question, we investigate how three fundamental principles of network formation proposed by Schaefer et al. (2010) apply among offline pre-existing social relations — referred to as *offline friends* — versus non pre-existing social relations — referred to as *online friends* — on Twitter. In this study, *offline friends* comprises of followers or followees on Twitter whom a user knows in the real world, whereas *online friends* comprises of followers and followees on Twitter whom a user does not know in the real world. As such, the set of offline friends and the set of online friends are mutually exclusive.

For your information, a user's followers are other users who follow the user on Twitter. Meanwhile, a user's followees are other users who are followed by the user on Twitter.

Since we only have the ground-truth data of a user's offline and online friends, we are making an assumption that all offline friends are pre-existing social relations, and all online friends are non pre-existing social relations. We believe this is a reasonable assumption to make because people maintain an online social network mainly to keep in touch with existing social relations that they have offline and meet new people online (Ellison et al. 2007).

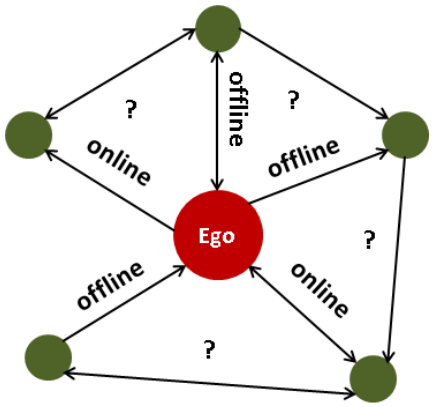


Figure 3.1: Ground truth ego networks.

3.2 Dataset

For our analysis, we use two datasets. The first dataset is the dataset by Xie et al. (2012). This dataset contains the data of 98 Twitter users and the list of his Twitter friends (followers or followees) whom he knows outside the Internet.

We also crawled the ego networks of these users in 2011. The illustration of the ground truth ego networks can be seen on Figure 3.1. From the illustration, the definition of an ego network can be understood clearly. An ego network includes a Twitter user – called *ego user* who is depicted by the red circle – and his followers and followees on Twitter. The edges among all of these users are crawled, producing a two-hop follow-networks that are bounded by the ego user and his followers/followees. In the ground truth data, we have the labels of who the offline friends among an ego user's followers or followees are. We procure these labels from the survey answers. However, there is a limitation to our ground truth data. The relationship types (offline or online) of the edges between the followers or followees of the ego users, are missing. These edges are marked by '?' in Figure 3.1. Our experiment and analysis will take into account this limitation.

Overall, the dataset has 20030 Twitter users (ego users and their followers/followees) and 23225 edges labeled as an offline or an online friend. We only use 49 ego networks (9380 users and 10153 labeled edges) for our observation. Based on our observation, we formulate rules to predict offline friendship and use the rest 49

ego networks for our prediction task.

The second dataset[2] is collected in 2015. This dataset contains the data of 41 Twitter users that include his ego network in 2015 and the list of his Twitter friends whom he knows in real life. The ego networks in this dataset consist of more private users than the ego networks in the previous dataset. Therefore, the ego networks crawled are not complete. We can only crawl the edges that come from or to public users. Overall, the dataset has 8696 Twitter users (ego users and their followers/followees) and 6170 edges labeled as offline or online friends. In this survey, we do not ask a user to label all their Twitter friends but only a sample of at most 100 of their friends. We do so because we use this dataset for another experiment in Chapter 7 and ask users more questions regarding their relationships with their sampled friends. To avoid low answers quality due to user fatigue, we sampled friends that a user needs to label. All the data is used as a test dataset except when using the machine learning algorithms.

3.3 Fundamental Principles of Network Formation among Offline versus Online Friends

Social networks are formed through multiple principles. Snijders (2011) listed some of the important ones in his work, they are: reciprocity, homophily, transitivity, degree differentials (popularity), and hierarchies. Schaefer et al. (2010) particularly picked up three principles — reciprocity, popularity, and triadic closure — to study the process of network formation among preschool children. They proposed that these principles were general because they had been proven to apply to the purest offline networks available, that is the children networks. Meanwhile, relationships formation in the offline networks of older people might be contaminated by pre-existing relationships and their cumulative socialization effect.

[2]The second dataset is added for the completion of this thesis.

Through longitudinal study using the SIENA modeling framework (Snijders 2001), they discovered that reciprocity, popularity and triadic closure shaped the formation of pre-school children's networks. As most children regularly interact with their peers for the first time in preschool, and they do not have prior social experience that might contaminate their motivation in creating social ties with their friends, the principles that govern their network formation are considered fundamental. Therefore, we choose these three principles to investigate in this study.

3.3.1 Reciprocity

Reciprocity means requiting a benefit received (Gouldner 1960). Since friends enjoy equality in right, privileges, and obligations (Laursen and Hartup 2002), reciprocity becomes the basis of friendship. On Twitter, reciprocity can happen when two users reply each other, mention each other, follow each other, etc. In this study, we focus on reciprocity that has a direct impact on a Twitter follow-network dependency, that is, reciprocity when two users follow each other. Although reciprocity is one of the basic principles of moral codes in a society which enables social stability (Gouldner 1960), it may not necessarily assume such a fundamental role when it comes to online friends in an online society. Therefore, in this study, we answer the following research question:

Research Question 1. *Does reciprocity as the basis of Twitter follow-network formation happen as often among online friends as among offline friends?*

Figure 3.2 shows the distribution of reciprocated links among offline and online friends. To answer the research question, we perform chi-square test of independence to check whether reciprocity depends on the type of friendship (offline or online). Our result shows that reciprocity depends on the type of friendship with odds ratio 11.02 (χ^2 = 2553.8, p-value < 0.001). Offline friends are 11 times more likely to reciprocate on Twitter.

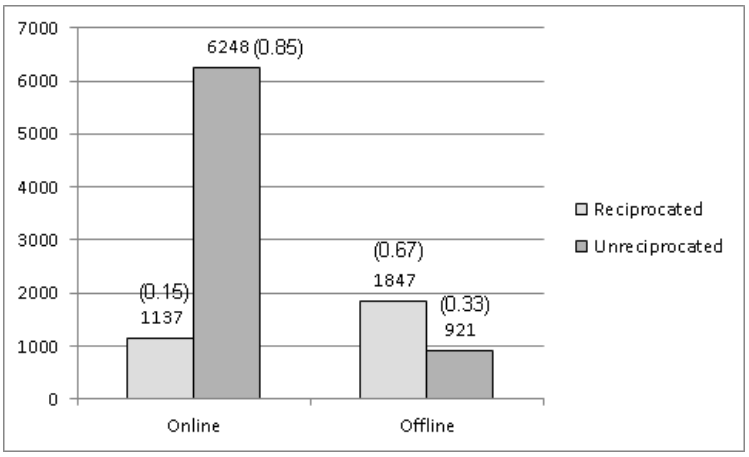


Figure 3.2: Reciprocated links among offline and online friends.

Based on this observation, we create our first conjecture to predict offline friendship. We divide our first conjecture into two, a conjecture with a fixed output and a conjecture with a probability output. A conjecture with a fixed output states that, given two online friends, A and B, on Twitter:

Conjecture 1. IF A and B reciprocate on Twitter THEN A and B are offline friends.

Meanwhile, we take the proportion of reciprocated and unreciprocated ties between offline friends as the output for the conjecture with a probability output. If friends reciprocate, then $\frac{1847}{1847+1137} = 62\%$ of the time they are offline friends. If friends do not reciprocate, then $\frac{1137}{1847+1173} = 38\%$ of the time they are offline friends.

3.3.2 Popularity

Popularity means the state of having many connections. An individual's popularity increases as the idealized qualities imposed by society increase, e.g. wealth, beauty, and social skill (Adler et al. 1992). These idealized qualities increase one's attractiveness and invite connections. As popularity allows a person to access more resources (Coie and Dodge 1983), popularity also entails higher popularity. The theoretical account of this phenomenon was elaborated by Price (1976). This phenomenon is called *the-rich-get-richer phenomenon*, or *preferential attachment*

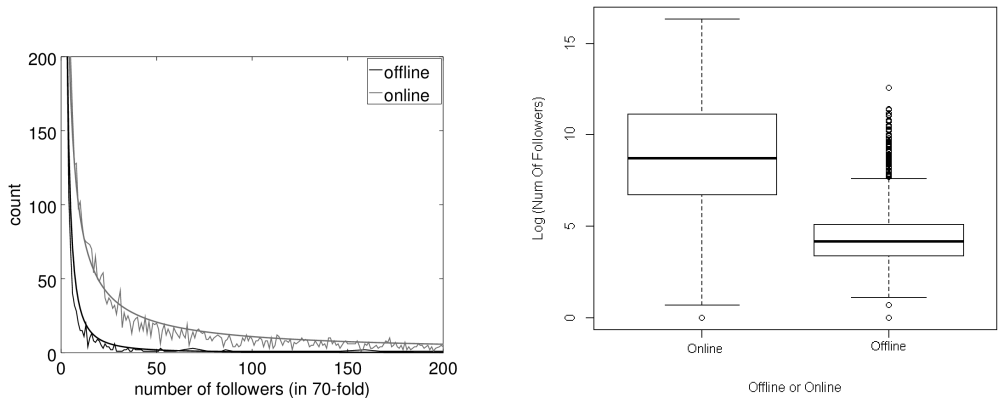
(Barabási and Albert 1999). Therefore, popularity in itself is also an idealized quality that increases one's attractiveness. On Twitter, the number of followers is the simplest measure of popularity.

Although preferential attachment has been shown to exist in both the online social network (Leskovec et al. 2008) and the offline social network (Price 1976), we wonder whether the rate at which popularity increases a user's attractiveness among online friends differs from the rate at which it does among offline friends. In this study, we answer the following research question:

Research Question 2. *On Twitter, does preferential attachment happen among online friends at the same rate as it does among offline friends?*

We plot the distributions of the number of followers of offline friends and online friends. Although in general they follow the power law, there is too much fluctuation in the distributions, thus making it impossible to find the parameters that fit a power law curve closely. Therefore, we try several folds of number of followers and discover that the distributions of the number of followers (in 70-fold) of both offline friends and online friends fit the power law closely ($N = cx^{-\alpha}$ where N is the frequency of users with a specific number of followers, and x is the number of followers in 70-fold), but at different parameters c and α (c is 1482.16 and α is 1.70 among offline friends, c is 769.13 and α is 0.92 among online friends. See Figure 3.3(a)). The power law distributions show that preferential attachment exists (Price, 1976), and it happens at a faster attachment rate among offline friends judging by the larger α.

A stranger (an online friend) has a thicker tail, meaning he has a greater tendency to have a higher number of followers. The next question is, whether there is a number of followers at which a user is likely to be an online friend to anyone. According to previous studies, there may be. Kwak et al. (2010) discovered that homophily was not observed between a user who had more than 1000 followers and his reciprocal friends. Moreover, another study showed that 71% of top link



(a) Power law distribution of followers.

(b) Boxplot of followers.

Figure 3.3: The distribution of the followers of offline and online friends.

farmers (users who try to acquire large numbers of follower links to amass influence) on Twitter had more than 1000 followers (Ghosh et al. 2012). Link farmers usually reciprocate even those whom they do not know to amass social capital and promote their Twitter content. As a result, many of the users in their network are strangers. Our boxplot in Figure 3.3(b) also shows that a user who has more than 1000 followers (log 1000 = 6.9) is at around the 87th percentile of all offline friends. Meanwhile, such a user is only at around the 25th percentile of all online friends. Thus, we formulate our second conjecture to predict offline friendship. We divide our second conjecture into two, a conjecture with a fixed output and a conjecture with a probability output. The conjecture with a fixed output states that, given two online friends A and B on Twitter:

Conjecture 2. IF B has more than 1000 followers THEN A and B are not offline friends.

Meanwhile, the conjecture with a probability output inserts the number of followers into the power law distribution on Figure 3.3(a). The power law distribution produces the number of offline friends and online friends that can be expected from a user with such number of followers. We divide these number by the total number of offline and online friends in the training dataset respectively to get the probabilities. If the probability of an offline friend having such number of followers is larger than that of an online friend, we assume that the friend is an offline friend.

3.3.3 Triadic Closure

Triadic closure happens between offline friends because of the increased propinquity and the psychological need for balance between two individuals who share mutual friends (Schaefer et al. 2010). If we assume that a triadic closure in real life translates into a triadic closure online, it is likely that triadic closure happens between offline friends on Twitter. On the other hand, as the pressure towards closure may not be as strong among online friends due to the lack of propinquity, we ask the following research question:

Research Question 3. *Are triadic closures on Twitter as likely to happen among online friends as they are among offline friends?*

We answer the research question by the following logit function:

$$\ln\left(\frac{p(\text{closure}=1)}{1-p(\text{closure}=1)}\right) = \beta_0 + \beta_1 I_1 + \beta_2 I_2 \tag{3.1}$$

I_1 is 1 if there is 1 offline friendship between any two users in a triad, I_2 is 1 if there are 2 offline friendships between any two users in a triad, and I_1 and I_2 are 0 if there is no offline friendship in a triad. The closure can be either an offline or an online link.

Section 3.2 has informed us on the limitation of the dataset, that is we only have the labels between an ego user and his followers/followees. Consequently, the maximum number of labels indicating the relationship type between two users in a triad (offline/online) is only two. Therefore, in the equation we only have I_1 and I_2. The triads that we consider as an input to the Equation 3.1 must have at least two sets of connected pairs of nodes. The type of the connection in the triads, that is which node follows which, does not matter. You can view the images of all the triads we consider in the Equation 3.1 in Appendix A.1.1.

The result shows that when offline friendship does not exist, a triadic closure is unlikely to happen (β_0 -3.36, p-value < 0.0001). β_0 -3.36 indicates that the

likelihood of closure is by default $\frac{1}{1+e^{-(-3.36)}} = 0.03$. When an offline friendship exists, the probability of a triadic closure increases (β_1 = 0.60, p-value < 0.0001). The probability becomes $\frac{1}{1+e^{-(-3.36+0.60)}} = 0.06$. When two offline friendships exist, the probability increases further (β_2 = 1.41, p-value < 0.0001). Specifically, the probability is now $\frac{1}{1+e^{-(-3.36+0.60+1.41)}} = 0.21$. From the result, we expect that when three offline friendships exist in a triad, an online triadic closure is even more likely to happen even though the ground-truth data that we have does not allow us to validate our expectation. In summary, when offline friendships exist in a triad, a triadic closure online is more likely to happen.

From this observation, we formulate the third conjecture to predict offline friendship. We divide our third conjecture into two, a conjecture with fixed output and a conjecture with probability output. The conjecture with fixed output states that given A-B-C, an online closed triad on Twitter,

Rule 3.3.1. *IF A and B are offline friends AND B and C are offline friends, THEN A and C are offline friends.*

Meanwhile, to derive the conjecture with a probability output, we perform another logistic regression (Equation 3.2) that calculates the likelihood of an edge being offline given the ratio of the number of closed triads ($|C(e_{ij})|$) and open triads ($|P(e_{ij})|$) in which the edge is involved.

$$\ln\left(\frac{p(e_{ij} = \text{offline})}{1 - p(e_{ij} = \text{offline})}\right) = \beta_0 + \beta_1 \frac{|C(e_{ij})|}{|P(e_{ij})|} \tag{3.2}$$

Running Equation 3.2 gives us $\beta_0 = -1.27$ (p-value ≤ 0.001) and $\beta_1 = 4.37$ (p-value ≤ 0.001). Conjecture with a probability output inserts the ratio of the number of closed triads and open triad in which an edge is involved into the equation that we have derived to get the probability of the edge being an offline friendship.

3.4 Practical Application: Predicting Offline Friendship on a Twitter Network

A hands-on practical application from the above observation is the formulation of rules for offline friendship prediction on a Twitter network which we will investigate in this work. We will compare the accuracy of using Schaefer's principles and other algorithms. All of these algorithms are described below.

3.4.1 Schaefer's Principles

We apply the three principles proposed by Schaefer et al. namely reciprocity, popularity, and triadic closure to predict offline friends on a Twitter ego network. The three conjectures with fixed results are combined into Algorithm 1. The algorithm for the conjectures with probability output are similar except that conjecture one and two in Algorithm 1 do not give fixed results given the condition, but probabilities given in Section 3.3.1 and Section 3.3.2 respectively. We also insert these principles into the artificial neural network algorithm to see whether machine learning improves accuracy.

3.4.2 Individual Conjectures from Schaefer's Principles

We also apply each principle proposed by Schaefer et al. separately to ensure that the success of the three principles in predicting offline friends on Twitter is not caused by the domination of the success of any one principle. First, we only use reciprocity to predict offline friends. Next, we use the principle of popularity to predict offline friends. Lastly we consider the principle of triadic closure. We try both the principles with a fixed output and a probability output.

Algorithm 1: Offline friendship prediction

Data: a Twitter user, u_i

Result: u_i's offline friends, C_i

u_i has a set of friends on Twitter S_i where $S_i = \{f_1, f_2, f_3...\}$
Let C_i be the set of u_i's offline friends
for *each friend* $f_j \in S_i$ **do**
    Apply *Conjecture 1*: If u_i and f_j reciprocates on Twitter then $f_j \in C_i$
    for *each friend* $f_j \in C_i$ **do**
        Apply *Conjecture 2*: If f_j has a number of followers larger than 1000 then $f_j \notin C_i$
    end
end
Apply *Conjecture 3*: Offline friends of an offline friend are offline friends
$temp = \{u_i\}$
while $temp.size$ *!= 0* **do**
    for *each friend* $f_j \in C_i$ **do**
        Let S_j be the set of f_j's friends on Twitter where $S_j \subset S_i$
        Let C_j be the set of f_j's offline friends where $C_j \subset S_i$
        for *each friend* $f_g \in S_j$ **do**
            Apply *Conjecture 1*: If f_j and f_g reciprocates on Twitter then $f_g \in C_j$
            for *each friend* $f_g \in C_j$ **do**
                Apply *Conjecture 2*: If f_g has a number of followers larger than 1000 then $f_g \notin C_j$
            end
        end
        $temp = temp \cup C_j$
    end
    $temp = temp \setminus \{C_i, u_i\}$
    $C_i = C_i \cup temp$
end

3.4.3 Machine Learning Algorihtms

We use various popular machine learning algorithms to predict offline friends namely logistic regression, naive bayes, support vector machine, and artifical neural network. However, we do not use Schaefer's principles as the input features.

For the input to the algorithms, we extract other network and interactive features on Twitter as predictors. They are indegree centrality, outdegree centrality, closeness centrality, node betweenness centrality, edge betweenness centrality, number of tweets, number of followers, number of followees, number of mentions, number of replies, and LDA-topic similarity.

3.4.4 Xiewei's Algorithms

Xiewei's algorithm (Xie et al. 2012) creates a matrix of a user's ego network and assigns a probability of walk from a user to his Twitter followers that decreases polynomially as a user's number of followers increases. Therefore, a user who has 1000 followers has a lower probability of walk to anyone than a user who has 100 followers. When the probability of walk to a friend is higher than the probability of walk to another friend who has the median number of followers, the friend is regarded as an offline friend. The process is performed iteratively to include offline friends of offline friends as offline friends.

3.4.5 Results

The prediction results are shown in Table 3.1 and 3.2. Overall, Schaefer's principles perform well and beat the machine learning algorithms. Individually, each of these principles does not perform as well. Reciprocity is a simple principle that beats most of the machine learning algorithms. Popularity has the same F-score as the F-score achieved by applying the principle of reciprocity, however precision and recall achieved are less balanced.

Algorithm 1 is none other than combining the principle of reciprocity, popularity, and triadic closure in the following way: (reciprocity $\cap$ popularity) $\cup$ TC (≥ 2 reciprocal edges) in which the friends involved in the triad are not popular. Simply put, two Twitter users who follow one another, have a number of followers less than 1000, or are involved in a closed triad with at least two reciprocal edges in which

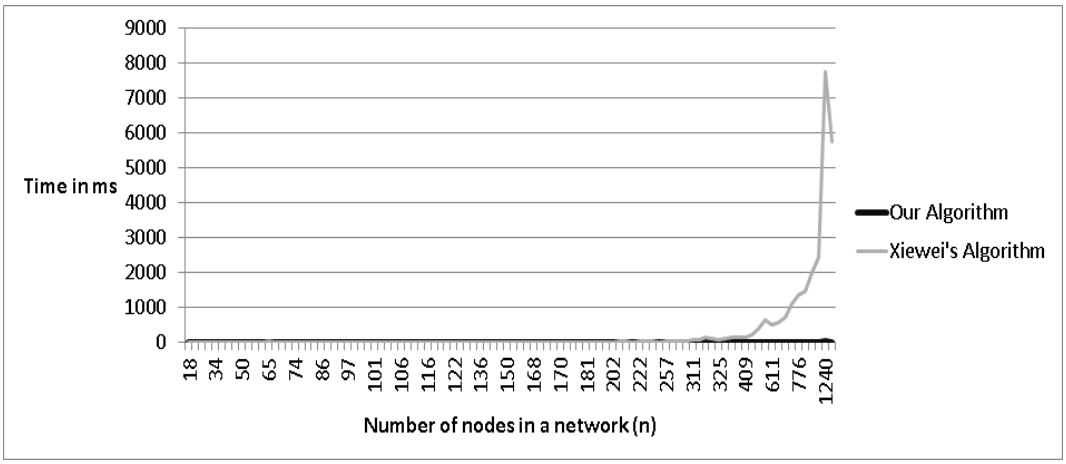


Figure 3.4: Milliseconds required to perform prediction

members are not popular, are offline friends.

Conjectures with a probability output reduce accuracy. For the new dataset, Schaefer's principles perform as well as Xiewei's random walk algorithm. The reason could be that missing edges in the new dataset affects the efficacy of Xiewei's random walk that depends very much on a global network structure, unlike Schaefer's principles that mostly depend on a local network structure. We have mentioned before in Section 3.2 that the new dataset has a lot more missing edges. Although the accuracy of Schaefer's principles loses to Xiewei's for the old dataset, Schaefer's principles reduce the time complexity from $O(n^2)$ to $O(n)$ (See Figure 3.4).

3.5 Conclusion

We have shown that some of the fundamental principles of social network formation, namely reciprocity, popularity, and triadic closure apply mainly to offline friends on Twitter. The results suggest that using an online social network as a substitute for a real life social network requires careful consideration as the dynamics that apply to the offline social network does not necessarily apply to the online friends. We also use the results of our observation to create an efficient algorithm for offline friendship prediction.

Although our algorithm loses in accuracy to the Xiewei's algorithm, it wins

Table 3.1: Prediction Results of Xiewei's Dataset

Algorithm		Precision	Recall	F-score
Schaefer's principles(F)		0.78	0.74	0.76
Schaefer's principles(P)		0.77	0.65	0.71
ANN with Schaefer's principles		0.75	0.73	0.74
Single conjecture	Reciprocity(F)	0.73	0.65	0.64
	Reciprocity(P)	0.47	0.41	0.44
	Popularity(F)	0.54	0.94	0.64
	Popularity(P)	0.46	0.78	0.58
	TC(F)	0.37	0.91	0.53
	TC(P)	0.37	0.29	0.33
M.L.	Logistic Regression	0.73	0.52	0.61
	Naive bayes	0.47	0.81	0.60
	SVM	0.78	0.36	0.50
	ANN	0.72	0.72	0.72
Xiewei's Random Walk Algorithm		0.77	0.88	0.82

(F) means conjecture with fixed output, (P) means conjecture with probability output

in other aspects. First, it is much more scalable. The time required to complete Xiewei's algorithm increases exponentially when the number of nodes in an ego network increases. Meanwhile, the time required to complete our algorithm remains stable. Second, our algorithm performs as well as Xiewei's algorithm in the presence of larger amounts missing data because it does not require as much edge data as Xiewei's algorithm does.

The limitation of this work mainly lies in the fact that the algorithm only focuses on Twitter. Future work can be directed to assess the applicability of the algorithm across various social networks in a larger dataset. Additionally, future work can analyse other principles of network formation beyond the three principles that we have analysed.

Table 3.2: Prediction Results of the New Dataset

Algorithm		Precision	Recall	F-score
Schaefer's principles(F)		0.87	0.94	0.90
Schaefer's principles(P)		0.88	0.87	0.88
ANN with Schaefer's principles		0.88	0.90	0.89
Single conjecture	Reciprocity(F)	0.92	0.76	0.83
	Reciprocity(P)	0.89	0.52	0.66
	Popularity(F)	0.74	0.99	0.85
	Popularity(P)	0.76	0.97	0.85
	TC(F)	0.78	0.93	0.85
	TC(P)	0.77	0.28	0.41
M.L.	Logistic Regression	0.86	0.95	0.90
	Naive bayes	0.92	0.74	0.82
	SVM	0.83	0.97	0.90
	ANN	0.88	0.93	0.91
Xiewei's Random Walk Algorithm		0.88	0.89	0.89

(F) means conjecture with fixed output, (P) means conjecture with probability output

Chapter 4

Essay 1B: Going Beyond Triads: Discovering Social Cliques on Twitter Follow-Networks

4.1 Introduction

These days, news spread across social media is prevalent (Gottfried and Shearer 2016). In maintaining what makes social media platforms social at all, offline friends are shown to be essential[1]. Communicatively, offline friends are more likely to reply and mention one another. Structurally, offline friends are shown to be more likely to reciprocate, have mutual friends, and consequentially form triads (Kim et al. 2016). The idiosyncrasies of offline friends, and their ability to reciprocate and form triads, have been used to predict the presence of offline friends on Twitter with great accuracy (Natali and Zhu 2016). Without a doubt, triad, which is known to be a precursor of close friendships in the offline world, is a precursor of offline friendships on Twitter.

These previous studies, however, do not explore subgraph formations beyond

[1]This study is not yet published. In this thesis, *we* refers to me and the chair of my committee who has been involved in the study.

triads. Yet, humans can cognitively socialize with 150-200 friends (Dunbar et al. 2015), most of which are likely to appear on Facebook or Twitter. A social community on Twitter with more than three members can easily form a clique on Twitter follow-networks beyond a closed triad. However, it is not only unclear how big these social media cliques could be, but also what shape they could take. But although online social interactions likely do not produce strong connections that elicit intense loyal-ties, they still foster connections critical to expanding networks (Best and Krueger 2006). Therefore, we wonder if social cliques consisting of more than three people are more frequent with online friends (strictly those that don't know each other in the real world).

This lack of understanding of social cliques on social media beyond the closed triad inspires us to conduct this study. Particularly, using Twitter as a case study, we want to better understand the types and shapes of social clique formation among offline friends on social media. (We call a clique on Twitter "social" if it mimics the cliques that exist in offline social networks.)

Our research questions are as follows:

- **RQ1:** What are clique structures that commonly exist in offline social networks beyond closed triads? How frequent do they exist? What shapes do they take?
- **RQ2:** Which of these structures exist on Twitter? Which ones exist in a significant manner?
- **RQ3:** Which of these follow-network structures on Twitter follow-networks have a higher probability of occurring among offline friends as opposed to online friends?

For brevity moving forward, we will refer to online friends who do not know one another offline simply as online friends. We answer the first question by using the Louvain algorithm (Blondel et al. 2008) to find cliques of various sizes in a selection

of offline networks. We answer the second and third questions by performing the Louvain algorithm iteratively.

We are the first to investigate social clique formations on Twitter beyond triads. The investigation includes three aspects, namely size, shape, and the type of friends in which these social cliques occur.

4.2 Background

In this section, we explain the background study related to our research questions. This is so readers will understand and become familiar with previous research that has inspired our research questions.

4.2.1 Sociology of Social Media

The two words that make up "social media" provide insight into its dual nature. On the one hand, social media has a communal side – this indicates the existence of human relationships that are in part defined by companionship and intimacy.

On the other hand, "media" is also a major component, and influences how users interact with specific content: they read it, they watch it, and they use it (Kietzmann et al. 2011).

As the Internet is utilized not just to expend content, but also to create, modify, and discuss it (Kietzmann et al. 2011), the "social" and "media" aspects are brought together. When merged, they feed off one another to increase their utilization: While the influence of socializing motivates news sharing (Lee and Ma 2012), news sharing also brings about social movements (Gleason 2013).

With social media platforms playing an increasingly pivotal role in supporting news production and diffusion (Gottfried and Shearer 2016; Lee and Ma 2012), questions arise as to whether social media sites such as Twitter have supplanted traditional media outlets and become truly new media platforms themselves (Kwak et al. 2010). But, does the sociological aspect of social media still persist? Recent

studies have proven that it does. Dunbar et al. (2015) have demonstrated that the communication layers that exist on Twitter are precisely the same as those that can be found among relationships in offline social networks. Meanwhile, a human's cognitive limit of maintaining stable relationships online is the same as the one that exists offline (Gonçalves et al. 2011).

4.2.2 Offline Friends Role in Maintaining the Sociology of Social Media

Offline friends' role in maintaining the sociology of social media has previously been researched. Kim et al. (2016) discovered that offline friends communicated more on Twitter. Additionally, Dunbar et al. (2015) determined that communication layers on Facebook and Twitter mimicked those in the offline world.

Beyond communication networks, other research has shown that offline friends also form distinguishable follow-network patterns on Twitter, which allow them to be found simply by observing users follow-networks (Natali and Zhu 2016). These distinguishable network patterns are highly reciprocal, and made up of closed triads. Interestingly, triads have also been known to be a balancer of social relationships in offline social networks (Granovetter 1973). Appearing among both offline friends on Twitter follow-networks and close friends in offline social networks, a closed triad indicates a social group offline and online. In this study, we investigate whether there are other social patterns (i.e. patterns that resemble social relationships in offline social networks) beyond closed triads that also exist on Twitter follow-networks.

4.2.3 Social Clique

There are two definitions of a social clique. The first is a narrow definition first presented by Luce and Perry (1949). Theirs presents a clique a maximal complete subgraph. The definition follows Granovetter's theorem (Granovetter 1973) which

states that members of a social clique were likely to be strong, and any two strong ties would create a closed triad among friends who were involved in said triad. If his theory of balance is strictly applied, a social clique of any size will have all members connected to one another.

Nevertheless, some researchers have argued that the definition is quite stingy because many friendship groups might be densely connected but not complete (Alba 1973). It is also possible that a member of a social clique does not like all members of the group that they frequently associate with. Not all members of a social clique will refer to one another reciprocally as best friends (Bagwell et al. 2000).

In this study, we use the loose definition of a social clique – that is, a cohesive subgroup of individuals in which relations among members of the subgroup are more important than relations between subgroups (Alba 1973). However, not all members need to be adjacent. In other words, social cliques are simply close communities, and some works such as (Bernard et al. 1980) use the community-finding algorithm CONCOR to discover them. In this study, we define a social clique as a cohesive subgroup in offline social networks. A cohesive subgroup on Twitter follow-networks may not necessarily be "social" since Twitter is a social information network in practice.

4.3 Datasets

To answer our research questions there are two datasets that are necessary. The first dataset comes from Twitter. The second dataset comes from the offline world. We will describe both datasets in this section.

4.3.1 Twitter Dataset

Twitter is the most popular microblogging service. It allows users to share posts, dubbed tweets, which cannot be longer than 140 words each. Besides, users are also allowed to share images or video links. A user can follow other users on Twitter and

by doing so be exposed to all the tweets that they publish. Unless a user specifies otherwise, a Twitter account is by default public and any web users can access the posts. Following a public account does not need an approval from the account holder. However, if an account is specified as private, only approved followers can see the tweets published by the account. Besides the fact that it has the most publicly available data than many social media, some other reasons make Twitter a platform of interest in our study.

- The informative nature of Twitter is most obvious in comparison with other social media. Loose social relationships on Twitter due to one-way following that generally does not require consent, have made Twitter one of the earliest social information network (Kwak et al. 2010). Information diffusion has been strongly coupled with social aspects of Twitter that the line between a social and information network becomes blurred (Shi et al. 2014). The news-like nature of Twitter makes it one of the best case studies to study information aspect of a social media.

- The platform was released in 2006, and therefore it has a large and mature community of users.

Survey data from 98 Twitter users in 2011 were obtained (Xie et al. 2012). It is the same dataset as the one described in Section 3.2. Therefore, we are not going to describe the dataset any more here.

Although the data were collected six years ago, they are still relevant to answer our research questions because Twitter utilities as both information and social network have not changed from the year the data were collected until the present. Since 2008, Twitter has already been known for its informational value and news use (Kwak et al. 2010) besides being started off as a pure social utility in 2006. Despite its continued development as an information media, Dunbar et al. (2015) have shown that social communities on Twitter are still thriving. Therefore, from 2008

Table 4.1: Offline Dataset

Category	Code	Network	#Nodes	#Edges	Dens.	Trans.	#CC	$\overline{D}$	$\overline{SP}$
Close Fr.	A1	Pupils	50	122	0.05	0.48	7	3.00	1.40
	A2	Dining	26	52	0.08	0.13	1	6.00	2.81
	A3	Prison	67	182	0.04	0.28	1	7.00	3.35
Office	B1	Wood Proc. Facility	24	76	0.14	0.35	1	6.00	2.99
	B2	Enterprise	35	168	0.14	0.40	1	10.00	3.23
	B3	Thurman Office	15	66	0.31	0.52	1	3.00	1.88
Int. Group	C1	Research Group	34	350	0.31	0.48	1	4.00	1.81
	C2	Flying Teams	48	340	0.15	0.36	1	5.00	2.40
	C3	Karate	34	154	0.14	0.25	1	5.00	2.41
Terrorist	D1	Al-Qaeda	271	1512	0.02	0.62	12	2.75	1.71
	D2	Bali	27	204	0.29	0.55	1	4.00	1.88
	D3	Greek	18	92	0.30	0.50	1	2.00	1.70

Col I. **Close Fr.** Close Friendship, **Int. Group** Interest Group
Col II. **Proc.** Processing
Col III-END. ***Dens.** Density, **Trans.** Transitivity, **CC.** Connected Component, $\overline{D}$ Average Diameter, $\overline{SP}$ Average Shortest Path

until recently, the utility of Twitter has not changed as an information and a social network at the same time.

4.3.2 Offline Dataset

Social relationships first existed in the offline world in various forms, mainly family relationships and close friendships. A network representing these social relationships is called a social network. Therefore, we define a clique on Twitter follow-networks as social if it closely resembles a clique of social relationships observed in the offline world. We will investigate cliques formed by various social relationships in the offline world (see Table 4.1). On Table 4.1 we convert any undirected graph to a symmetrical directed graph before quantifying its network measures. Some of the networks investigated are not connected but separated into several components, such as the Pupils network and Al-Qaeda network. For such networks, we quantify the average diameter and the average shortest path length by first breaking them into connected components before averaging the measures of all the connected components.

Close Friendship Network

The close friendship network is a network of close friends. In our dataset, we

have three close friendship networks, namely, Pupils network (Pearson and Michell 2000), Dining network (Moreno 1960), and Prison network (MacRae 1960). Close friendship networks are networks made up by having users name up to the x-th of their best friends.

Except for the Pupils network, close friendship networks have a larger average shortest path length showing that dining friends and prison friends rarely have mutual friends. As such, Dining network and Prison network also have lower transitivity. Therefore, although the closest friends may have many mutual acquaintances or ordinary friends, they may not have many mutual friends with whom they share close intimacy.

The reason that most intimate friends do not have many mutual friends who are very intimate could be what Simmel (1950) had stated, "In the dyad, the sociological process remains, in principle, within personal interdependence and does not result in a structure that grows beyond its elements. This also is the basis of *intimacy*.".

Office Network

The office network reflects the networks of people interacting in the office. In our dataset, we have three networks representing the office network, namely Wood Processing Facility network (Michael 1997), Small Enterprise network (Rogers and Kincaid 1981), and Thurman Office network (Thurman 1980).

In comparison to the close friendship network, the office network has as high average shortest path lengths, higher diameters, but higher transitivities and densities as well. This network is still loosely connected compared to the interest group and terrorist network.

Interest Group Network

The interest group network is a network of people who have similar interests. The values of network measures of the interest group are very similar to those of the office network, though with a slightly lower diameter and average shortest path length. In this study, we include the following interest group networks: Research Group network (Killworth and Bernard 1976), Flying Teams network (Moreno

1960), and Karate Club network (Zachary 1977).

Terrorist Network

The terrorist network is the densest network among all types of network. The Al-Qaeda network has a low density because the network is not fully connected. It breaks into 12 components. However, the transitivity of the Al-Qaeda network is extremely high showing that each component has a high density. The terrorist network also has a lower average diameter and a lower average shortest path length.The relations involved in a terrorist network vary. They are: (a) acquaintances and distant family ties, (b) friends and moderately close family ties, and (c) close friends/family. Therefore, the terrorist network resembles closely a network of social relationships of all degree of closeness, from the closest family members to acquaintances.

In our dataset, we include three terrorist networks. The Al-Qaeda network is the terrorist network responsible for over 10 attacks deployed by Al-Qaeda over a decade. The Bali network depicts the relations of individuals associated with the 2005 Bali bombing by Jemaah Islamiyah. The Greek network represents the relations of individuals associated with 17 November Revolutionary Organization, a Marxist urban guerrilla organization operating in Greece. All terrorist networks are sourced from Transnational Terrorism Database by John Jay and ARTIS[2].

4.4 Discovering Social Cliques in the Offline Social Networks

In this section, we answer our first research question: What are social clique structures that commonly exist in the offline world? How frequent do they exist? What are their shapes?

As explained in Section 4.2, social clique is none other than a cohesive subgroup of individuals in which relations among members of a subgroup are more

[2]http://doitapps.jjay.cuny.edu/jjatt/attributes.php

important than relations between subgroups (Alba 1973). Therefore, we use the Louvain algorithm (Blondel et al. 2008) to discover social cliques in offline social networks. The Louvain method is a heuristic algorithm that optimizes modularity of the communities discovered. Modularity measures the density of edges inside communities compared with the links between communities. Therefore, modularity is a quantitative representation of social cliques. The higher the modularity value, the greater the density of edges inside communities in comparison to those between communities, and thus, the more desirable these discovered communities are.

The social cliques that our study focuses on are of sizes larger than triads (4) and smaller than the size of a sympathy group (15). Previous studies have explored the occurrence of closed triads on the Twitter follow-network (Natali and Zhu 2016). Besides triads, there are two key social groups in social science, namely the support clique (individuals from whom one seeks assistance and support), and the sympathy group (individuals that one contacts at least once a month). The support clique typically consists of four to seven individuals, whereas the sympathy group typically consists of twelve to fifteen individuals (Dunbar and Spoors 1995). By running the Louvain algorithm, we will discover the shapes and frequencies of cliques that represent these two key social groups (cliques of size four to fifteen) in offline social networks.

4.4.1 Results

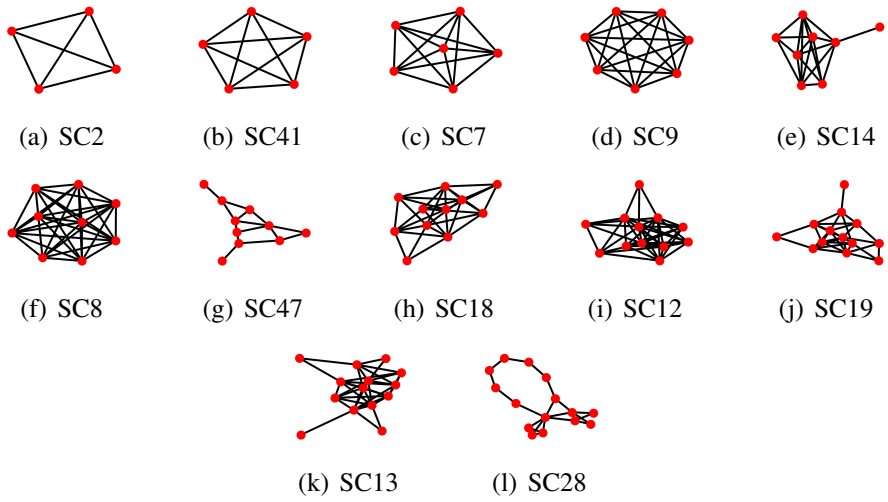
Using the Louvain algorithm, we discover social cliques of sizes 4 to 15. In total there are 50 social cliques discovered. Table 4.2 covers the densest social clique of each size. The table and figure that covers all social cliques can be viewed in Appendix A.2.1.

Our results show that the densest social cliques that are complete cliques (all nodes are adjacent to each other), such as SC2, SC7, SC8, and SC9, are mostly found in terrorist networks, particularly Al-Qaeda. Terrorist networks are the dens-

Table 4.2: Social Cliques in the Offline Networks

Code	$\|E\|$	$\|N\|$	freq.	networks
SC2	6	4	3	alqaeda,wood,bali
SC41	10	5	1	sawmill
SC7	15	6	1	alqaeda
SC9	21	7	1	alqaeda
SC14	20	8	1	research
SC8	36	9	1	alqaeda

Code	$\|E\|$	$\|N\|$	freq.	networks
SC47	15	10	1	wood
SC18	32	11	1	flying teams
SC12	43	12	1	research
SC19	33	13	1	flying teams
SC13	37	14	1	research
SC28	22	15	1	prison



(a) SC2 (b) SC41 (c) SC7 (d) SC9 (e) SC14

(f) SC8 (g) SC47 (h) SC18 (i) SC12 (j) SC19

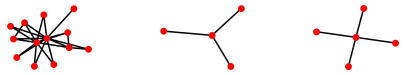
(k) SC13 (l) SC28

Figure 4.1: Densest social clique structures of sizes four to fifteen in offline networks.

(a) SC28 (b) SC32 (c) SC39

Figure 4.2: Chain-like structures are mostly found in close friendship networks

est networks among all offline networks. Meanwhile, chain-like structures are mostly found in close friendship networks, such as SC28, SC32, SC39 (see Figure 4.2), indicating that these close friends are very intimate therefore they are not well connected. Backstrom and Kleinberg (2014) have shown that mutual friends of intimate friends are not well-connected. Star-like social cliques, such as SC26, SC44, and SC46 (see Figure 4.3) are mostly found in interest group and office networks.



(a) SC26 (b) SC44 (c) SC46

Figure 4.3: Star-like structures are mostly found in interest group and office networks

4.5 Discovering Social Cliques on Twitter Follow-Networks

In this section, we investigate whether the social cliques discovered in Section 4.4 occur on Twitter follow-networks. We can do so by trying out all combinations of nodes on our Twitter follow-networks and test whether their structures are isomorphic to the social cliques that have been discovered. This method is inefficient and expensive. On the other hand, we can iteratively search for communities that are well clustered locally to determine whether they are isomorphic to the social cliques. This method is more efficient. Moreover, we will ensure that the structures discovered are meaningful community structures, because they are well-clustered locally. Therefore, we harness the iterative Louvain algorithm to investigate whether these social cliques occur on Twitter follow-networks.

In the iterative Louvain, we recursively perform the Louvain algorithm on all the communities discovered until the modularity of each community cannot be improved further by breaking the community down into smaller components. In other words, we break the communities into their smallest possible clusters. We will compare all the communities discovered with the social cliques that we have. We check how many of them are isomorphic. Isomorphism indicates that the social clique structure exists on the Twitter network.

4.5.1 Test of Significance

Not only do we want to know which of the social cliques exist on Twitter. We also want to know which of the structures significantly exist. The most common

approach to do so is by comparing the empirical network and its corresponding random network for the presence of these structures.

One of the most popular models of this approach is the exponential random graph models (ERGM), which is also called the p^* model (Frank and Strauss 1986). This is the most common approach that is used to test whether various network configurations, such as reciprocity and triads, have a higher probability of occurring in a network, an exemplary study is the one by Contractor et al. (2006). The process is analogous to regressing an equation to a set of data so that one finds which class of networks, resulting from a given objective function, best fits a set of data (Abbas et al. 2013). In a way, ERGM performs a null model comparison by trying out various models in which the resulting coefficients best fits the empirical networks (the ground truth data). The objective function can include the desired network configurations such as triads. It assumes that a network is represented by an exponential model that is solely explained by the variables in the objective function.

Although ERGM is widely used to test hypotheses on the significance of various network configurations in a network, it is not appropriate for our studies for several reasons. First, ERGM packages available are usually intended for social cliques of a limited size, the largest one being five (Yaveroğlu et al. 2015). Our study requires an analysis on social cliques of size up to fifteen. Second, the ERGM model is extremely slow to run especially for number of cliques above four. Therefore, it is commonly used for examining several hypotheses in a network. On the other hand, we want to test whether each of the social structures is likely to occur in many networks to discover which social clique structure is the best representation of social cliques among friends on Twitter follow-networks. Each social clique structure will represent a hypothesis. We need to test each hypothesis in 98 ground-truth networks that we have. ERGM requires a very long time to do so.

Therefore, we harness the approach of comparing empirical networks to null models, without using the ERGM model. We do so by the following steps.

For each of the empirical networks, we create a corresponding random network.

We want the corresponding random network to have density, size, percentage of offline friends, and degree distribution that are as similar as possible to those of the empirical network. Therefore we use the configuration model (Bender and Canfield 1978) to build our random network. The configuration model has been used to confirm the role of a given set of constraints in the presence of some empirically observed structural features (Tabourier et al. 2011). In this way, we do not consider all random models, but the closest random models to the empirical networks. Therefore, we get an upper bound of the significance level. If the structures significantly exist in comparison to the configuration model, we can be sure that they also significantly exist in comparison to other random models. Moreover, by comparing with the configuration model, we will ensure that the social cliques exist not because they are an artefact of the graph's inherent structural properties, but because the cliques formation is a property unique to Twitter users.

Because the empirical network has several constraints that the configuration model does not fulfil, we have to modify the resulting random network. First, the empirical network is an ego network. Thus, we make sure that ego users in the random network are connected to everyone else in the network. If they are not, then we will create a connection. Whether the connection is going to be from or to the ego user, is randomly decided. Next, the empirical network does not have self-loops and parallel edges, but the configuration model creates a random network that has these edges. Therefore, we remove all self-loops and parallel edges.

Next, we extract the offline and online networks from the empirical network and the corresponding random network. The offline network is a network that includes ego users, their offline friends, and all the connections among them. The online network is a network that includes an ego user, his online friends, and all connections among them.

Then, we count how many social cliques that have been discovered in the offline world appear in the offline, online, and overall network of the empirical network and its corresponding random network. We use the iterative Louvain algorithm to

find as many as possible on the Twitter, network clusters of nodes of the size of the social clique we are investigating. We then test whether they are isomorphic to the social clique.

Finally, we compare the quantities obtained in the previous steps across all empirical networks using paired t-test.

4.5.2 Results

After running the iterative Louvain, we present the social cliques that exist on the Twitter graphs. These results are presented on Table 4.3. We discover that the social clique structures on Table 4.3, are the structures in offline networks that also exist on Twitter networks.

Table 4.3: Average Number of Social Cliques on the Empirical Twitter Networks and the Random Twitter Networks

	All Graphs			Offline Graphs			Online Graphs		
Code	Emp.	Rand.	Sig.	Emp.	Rand.	Sig.	Emp.	Rand.	Sig.
SC2	**0.78**	0.76		0.60	**0.69**		0.19	**0.24**	
SC7	0.02	0.02		0.04	**0.06**		0.00	0.00	
SC9	**0.02**	0.00		0.01	**0.03**		0.00	**0.02**	
SC15	0.01	0.01		0.00	0.00		**0.02**	0.01	
SC20	**0.79**	0.33	**	**0.36**	0.11	*	**0.33**	0.12	**.**
SC21	**0.20**	0.19		0.02	**0.11**	**.**	0.05	**0.07**	
SC23	0.49	**0.63**		0.32	**0.41**		**0.15**	0.14	
SC25	0.34	0.34		0.03	**0.09**	**.**	**0.26**	0.09	**
SC27	1.24	**3.38**	***	0.49	**1.12**	***	0.70	**1.14**	**.**
SC34	**0.76**	0.32	**	**0.66**	0.51		**0.30**	0.06	*
SC37	0.00	**0.09**	**	0.00	**0.01**		0.02	0.02	
SC39	1.02	**3.03**	***	0.14	**0.89**	***	0.80	**2.09**	***
SC41	**0.22**	0.09	*	**0.22**	0.14		**0.04**	0.03	
SC44	3.72	**4.73**	*	**0.69**	0.60		2.64	**3.09**	
SC45	0.00	**0.01**		**0.03**	0.01		**0.02**	0.00	
SC46	**1.68**	1.43		**0.35**	0.15	*	**1.30**	0.92	*
SC49	3.81	**5.69**	***	2.16	**3.59**	***	1.58	**2.11**	

Emp. Average number of cliques on empirical networks

Rand. Average number of cliques on corresponding random networks

Bold Numbers Higher average values

Significance level: $^{***}p \leq 0.001$, $^{**}p \leq 0.01$, $^{*}p \leq 0.05$, . $p \leq 0.1$

The conjoined three closed triad (SC20) appears significantly on Twitter (see

(a) SC20 (b) SC46

Figure 4.4: Conjoined three closed triads (SC20) and four-star structures appear significantly on Twitter.

(a) SC27 (b) SC39 (c) SC49

Figure 4.5: Half-stingray structure (SC27), Chain structure (SC39) and 3-full-clique (SC49) are not likely to appear on Twitter networks given the graphical properties of these networks.

Figure 4.4). Therefore, the structure is a special structure among Twitter users that is unlikely formed among users of other networks. Meanwhile, the four-star structure (SC46) appears significantly among offline and online friends on Twitter. The structure is unique among offline friends and online friends on Twitter, and unlikely to appear among users of other networks. In offline networks, four-star structure mostly appear in interest-group networks and office networks, indicating hierarchical relationships in these environments. Meanwhile, on Twitter network, a four-star structure represents a centralized information structure. However, this structure does not appear significantly among all types of friends on Twitter networks, which means that a star structure is a natural configuration given the distribution of edges among offline friends or online friends, but not between them.

Meanwhile, the half-stingray structure, chain structure and closed triad are not likely to appear on Twitter networks given the graphical properties of these networks (see Figure 4.5). In random networks with the same density and degree distribution, these structures are more likely to exist. Although previous studies have discovered that offline friends are more likely to form triads than online friends are (Natali and Zhu 2016), these triads are actually not common considering the graphical properties of the Twitter networks.

4.6 Discovering Social Cliques among Offline and Online Friends on Twitter Follow-Networks

In this section, we are going to statistically test whether these social cliques are more likely to happen among offline or online friends. The explanation of these results in Table 4.4 divides all social cliques into groups based on their shapes.

Table 4.4: Average Number of Social Cliques among Offline and Online Friends on the Twitter Follow-Network

Code	Offline	Online	Sig.
SC2	**0.60**	0.19	**
SC7	**0.04**	0.00	*
SC9	**0.01**	0.00	
SC15	0.00	**0.02**	
SC20	**0.36**	0.33	
SC21	0.02	**0.05**	
SC23	**0.32**	0.15	
SC25	0.03	**0.26**	**
SC27	0.49	**0.70**	
SC34	**0.66**	0.30	
SC37	0.00	**0.02**	
SC39	0.14	**0.80**	***
SC41	**0.22**	0.04	**
SC44	0.69	**2.64**	***
SC45	**0.03**	0.02	
SC46	0.35	**1.30**	***
SC49	**2.16**	1.58	
Bold Numbers Higher average values Significance level: $^{***}p \leq 0.001$, $^{**}p \leq 0.01$, $^{*}p \leq 0.05$, . $p \leq 0.1$			

4.6.1 Complete Clique Structure

As expected, complete clique structures (Figure 4.6) exist more among offline friends, although not all are significant. The only significant complete cliques are the complete cliques of sizes four and five. The complete clique of size three, although has been proved to improve prediction of offline friends (Natali and Zhu 2016) and is higher in number among offline friends, is not significantly higher in

(a) SC2 (b) SC7 (c) SC9 (d) SC41 (e) SC49

Figure 4.6: Complete clique structures.

(a) SC15 (b) SC21 (c) SC25 (d) SC27

Figure 4.7: Stingray structures.

number. When we count the percentage of reciprocal edges in all cliques, we discover that the percentage of reciprocal edges among offline friends is approximately 50%, whereas it is only approximately 40% among online friends. The number of hubs (users who are the source of information to all nodes) is also higher in the cliques among offline friends (2.4) in comparison to those among online friends (1.1). The results show that offline friends are more likely to form a group in which all nodes are adjacent.

4.6.2 Stingray Structure

Generally, there is no difference between the number of stingray structures among offline friends and online friends (Figure 4.7), except for SC25, which is the stingray structure with no closed triads. This structure is more likely to exist among online friends than among offline friends.

4.6.3 Chain Structure

There are two types of chain structures that happen on Twitter (Figure 3.12). The chain with a bump (SC37) and the chain (SC39). The chain with a bump (SC37) only happens among online friends, and the total number is also small: only two (0.02×98). Meanwhile, there are 14 chain structures (SC39) that exist among offline friends and 78 chain structures that exist among online friends. The existence

(a) SC37 (b) SC39

Figure 4.8: Chain structures.

(a) SC20 (b) SC23 (c) SC34 (d) SC45

Figure 4.9: The incomplete clique social cliques.

of such structures among online friends is significantly higher than that among offline friends.

The chain structure (SC39) can create a chain of information. Each chain structure has a potential of creating two information chains: one that flows to the left and one that flows to the right. On average, each offline chain structure creates 0.79 information chain out of possible two (approximately 40% of the time, there is an information chain in a chain structure). Meanwhile, on average, each online chain structure creates only 0.38 information chain out of possible two (approximately 20% of the time, there is an information chain in a chain structure). Therefore, although the chain structure significantly exists more among online friends, it has a higher potential to relay information from one end to another among offline friends.

4.6.4 Incomplete Clique

An incomplete clique is a clique larger than three where all nodes are not adjacent (Figure 3.13). Overall, this clique happens as often among offline friends as among online friends. Therefore, although complete clique structures happen mostly among offline friends, incomplete clique structures do not. However, there are more reciprocal edges in incomplete cliques formed among offline friends (60%) than the ones formed among online friends (40-50%).

(a) SC44 (b) SC46

Figure 4.10: The star social cliques.

4.6.5 Star Structure

Star structures appear more among online friends (Figure 4.10). However, the four-star structure is a specific structure among both offline friends and online friends on Twitter in comparison to other networks with the same graphical properties (see Table 4.3).

Among offline friends, approximately 70% of the edges in the star are reciprocal. Meanwhile, among online friends only approximately 20% of the edges in the star are reciprocal.

Both offline and online friends have an equal percentage of edges that go from a centre node (80%), but offline friends have a much higher percentage of edges that go to a centre node (90%) than online friends (40%). The results indicate that star structures are equally used by both offline and online friends to receive information. However, they are more likely to be used by offline friends to spread information.

4.7 Conclusion

In this study, we investigate follow-network formations among offline friends on Twitter. In the first study, we have discovered that offline friends have fewer followers and are more likely to reciprocate and form triads. In the second study, we explore social structures larger than triads. Out of these structures, we discover that complete cliques of sizes four and six significantly exist on Twitter follow networks but only among offline friends. Meanwhile, star structures appear mostly among online friends. Chain structures appear mostly among online friends, but have more potential to relay information when they exist among offline friends.

Although this study has provided readers with the knowledge of substructures that commonly exist among offline friends and among online friends on Twitter, it requires further analysis to assess the practical implications of these results. For example, do the structures that exist significantly compared to the configuration model are specific only to Twitter or social information network? Furthermore, do structures that exist significantly more among offline friends can be used to predict offline friends? Hence, further studies can be directed to figure out the practical implications of these substructures.

Chapter 5

Essay 2A: Investigating the Role of Reciprocal Ties for Information Diffusion of Various Topics on Twitter

5.1 Introduction

It is impossible to know whether all ties involved in a full retweet chain are offline or online. At most, we are only able to know whether ties involved in a retweet chain in an ego network are offline or online. Therefore, one way to analyse the role of offline friends in a full retweet chain is to assume other ties as offline friends. In the previous chapter, we have discovered that reciprocated follow links can predict offline friends on average at a 73% precision, and 65% recall. In this study, we are going to investigate the role of reciprocated friends for information diffusion of various topics on Twitter.

Before the emergence of the online social network, the study of information diffusion was limited and difficult[1]. There are several reasons why it was so. First, the

[1]This study is an extension of a published work *The Role of Different Tie Strength in Disseminat-*

scale of information broadcast offline is limited. Second, the difficulty in obtaining the path of information spread offline impedes further research effort. However, the popularity of the online social networks today changes the situation and attracts researchers to study information diffusion on the online social networks (Guille et al. 2013). In the beginning, the study of information diffusion treats all ties as homogeneous (Li et al. 2014). However, as the degree to which individuals relate to one another differ, it only makes sense to treat ties as heterogeneous. Some research studies have exerted effort to do so. For example, Peng et al. (2011) discovered that social relationship is a good predictor of whether a user would retweet his friend. Zhao et al. (2012) found that strong ties were more favorable to information diffusion on Facebook than weak ties were. Meanwhile, a contrary finding was discovered by Shi et al. (2014) who empirically learnt that weak ties were more likely to engage in a retweet.

So far, none of these studies consider the choice of ties in the information diffusion of different topics. Yet, previous studies have repeatedly shown the interdependence between choice of ties and topic when two people communicate offline. Friedkin (1982) showed that weak ties were more important than strong ties in promoting information flow about activities outside an organizational subsystem. On the other hand, as Krackhardt (1992) discovered, information about organizational changes that challenged the employees' status quo was more likely to flow through strong ties. Moreover, Straits (1991) showed that strong ties were crucial in spreading political influence.

This paper presents an analysis of user's choices of ties under different tweet topics. There are two choices of ties that a user has: reciprocated ties, and unreciprocated ties. In this paper, reciprocated tie is a tie in which two users follow one another. Meanwhile, unreciprocated tie is a tie in which one user follows another user. To estimate the probability of diffusion, we use the system dynamics model.

ing Different Topics on a Microblog (Natali et al. 2017). In this essay, *we* refers to the authors of the published study.

In the system dynamics model, we assume that all strong ties behave in the same way, and all weak ties do too.

A system dynamics model for information diffusion is typically a modification of epidemiological model for information diffusion. An epidemiological model for information diffusion uses the mathematical framework for understanding the spread of diseases to understand the spread of information. The model that we choose to harness in this study is the SEIZ model invented by Jin et al. (2013) because this model has been shown to model both rumours and news on Twitter well, whereas other epidemiological models only focus on the modelling of news. This model originally treats all ties homogeneously. We modify the model to incorporate two types of users: one that retweets due to the influence from reciprocated ties and one that retweets due to the influence from unreciprocated ties. Consequently, the parameters that represent the probability of transition from one type of user (the potential retweeter) to another type of user (the retweeter) also split into two types, one for the reciprocated ties and one for the unreciprocated ties.

As for the case studies, we choose thirty tweets of two categories, controversial and uncontroversial category. Since the previous studies show that political information and information that challenges the status-quo are more likely to flow through the strong ties (Straits 1991; Krackhardt 1992), we compare the diffusion of controversial topic and uncontroversial topic. First, controversial topic is often political in nature. Second, it normally challenges a person's status-quo that it creates so much controversy, for e.g. the issue of immigration and healthcare. The tweets under uncontroversial topic can be further divided into four topics, namely general news, personal, entertainment, and rewards.

From October 12, 2016 to December 2, 2016 we crawl tweets that contain some pre-determined hashtags such as: *trump*, *clinton*, *hurricane*, *sports*, *kids*, *win*, etc. These hashtags presumably appear often in one of the five topics we have chosen. These tweets are retweeted 4,161-225,496 times (average 35,318 times). Within the time range specified, we manage to crawl 33-99% (average 80%) of the retweeted

tweets. These tweets represent the retweet cascade within 1.1-51.4 days (average 17.5 days). There are 599,800 retweeters in all the retweet cascades.

A first goal of the study is to investigate whether assuming a heterogeneous choice of ties (diffusion depends on whether ties are strong or weak) is better than assuming a homogeneous choice of ties (diffusion does not depend on whether ties are strong or weak). Previous works either assume a heterogeneous choice of ties or a homogeneous choice of ties, but do not compare the two choices. We discover that in the SEIZ model, 56% of tweets spread better when not assuming a heterogeneous choice of ties. Meanwhile, 44% of tweets spread better assuming a heterogeneous choice of ties. A second goal of the study is to investigate whether the choice of ties depends on topics. We find that when we assume a heterogeneous choice of ties, 69% of tweets are more likely to flow through strong ties. The rest of the tweets that flow mainly through weak ties are 88% non-controversial. Meanwhile, all controversial tweets flow through strong ties.

Our studies contribute to the study of tweets diffusion by investigating unexplored question about tweets diffusion. To the best of our knowledge, ours is the first study that tries to model the interdependence between tie strength and topic in tweets diffusion.

5.2 Related Work

Various retweet models have been developed to understand the information diffusion process on Twitter. They can be divided into two categories: the system dynamics model, and the agent-based model.

The system dynamics model usually makes use of the epidemiological model. In other words, it uses the mathematical framework for modelling the spread of disease to model the spread of information on Twitter.

Other models are usually the agent-based model. While the system dynamics model assumes that a group of people behave in the same way, the agent-based

model assumes that each individual has a unique behaviour that depends on his attributes.

In our literature survey we will cover some examples of epidemiological models for information diffusion. We will also cover some examples of the agent-based models. However, because there are so many agent-based models for tweets diffusion, we only concentrate on those that are pertinent to our research question. These models are the diffusion models that involve tie strength and the diffusion models that incorporate topic.

5.2.1 Epidemiological Model for Retweet

Epidemiological models are the classical approach to model how information diffuses. These models divide the total population into several groups. An individual from one group can transit into another group. Well known models are SI, SIR, and SIS. These models are named based on the group to which a population can belong. The SI model has two states, susceptible (S) and infected (I). In the SIR model, there is an additional transition a user can take, that is from infected (I) to recovered (R). Meanwhile, in the SIS model, a user can go back to being susceptible again after being infected. A user transitions into another group either by self-transition or by getting into contact with another user.

In order to adapt the epidemiological model to better mimic the diffusion of information, some variations of the models are proposed. The earliest rumor model was the Daley-Kendall (DK) model (Daley and Kendall 1964). The model divided the population into three groups: ignorant, spreader, and stifler. Ignorant had never heard the information. Spreader spread the information. Stifler knew but refused to spread. The three groups were similar to the susceptible, infected, and recovered group in the SIR model. At a later time, a more widely used model named Maki-Thompson (MT) rumor model was introduced (Svensson 1993). The model assumed that only the user who initiated the contact changed his state when meet-

ing with another user. Meng et al. (2014) developed the SISR model to study tweets propagation on Weibo, a Chinese microblog. The distinguished feature of the SISR model was the shortcut from susceptible (S) to recovered (R). Thus, a user's transition from susceptible to infected was not deterministic. A user may transition into an infected, or into a recovered user. The model was shown to fit the Weibo data better than the MT model. Meanwhile, Xiong et al. (2012) investigated the characteristics of information diffusion by SCIR model, but did not fit the model to the real data. SCIR model had an additional possible contacted state (C). When in the contacted state, an agent was exposed to the information, but did not immediately make the choice of whether to spread the information.

In our study, we used the SEIZ (susceptible, exposed, infected, skeptic) model to fit our Twitter data. The reason we choose SEIZ is that the model has been used to model both rumors and news on Twitter (Jin et al. 2013), whereas other models usually only focus on news. We also modified the SEIZ model to incorporate strong and weak ties, something that has never been previously done for epidemiological models for information diffusion.

5.2.2 Topic-based Retweet Model

Some of the retweet models that have been developed consider topic as one of the main features in the models. TwitterRank applied topic-based PageRank to rank influential users. An adjacency matrix was set up, and each user had a probability to transit to another user depending on the number of tweets about a topic he published in comparison to the number of tweets about the topic his friends published (Weng et al. 2010). Each topic had a unique transition matrix. The model had been shown to predict influential users better than the previous models.

On the other hand, Macskassy and Michelson (2011) incorporated into their retweet model the probability that a user retweeted a friend given the similarity of the topic between a user and his friend. On the contrary to the result found by

Weng et al. (2010), the model showed that the model fit well when the probability of retweeting decreases as the topic between two users' content grew more similar.

However, unlike in our study, these models did not question whether the likelihood of a topic to flow through the strong ties was different from the likelihood of a topic to flow through the weak ties.

5.2.3 Tie-strength-based Retweet Model

Arnaboldi et al. (2014) incorporated tie strength into their retweet model and discovered that by doing so they could predict the depth of the retweet cascade. Although, they used the real networks and seed nodes to run their simulations, they did not validate whether the cascades produced is the same as the ones in the real world. Moreover, the model did not consider different topics.

Another study on tie strength for retweet was conducted by Peng et al. (2011). Using the conditional random fields that maximized the probability of retweet given users' characteristics, the study discovered that the number of reciprocated mentions between two users was one of the most significant predictors of a retweet. Meanwhile, Zhao et al. (2012) assumed five steps of retweet on various friendship networks. Their method provided flexibilities in controlling the preferences and the channels for information propagation. From simulating the model on friendship networks, they discovered that although compared to weak ties, strong ties were more favourable to information diffusion, random selection strategy was more efficient than selecting strong ties for information propagation.

While the three studies above encourage the selection of strong ties for information diffusion, Shi et al. (2014) opined differently. Using two stage model for maximum likelihood estimation, they showed that weak ties were more likely to diffuse tweets than strong ties.

5.3 Methodology

To analyse the choice of ties for the diffusion of different tweet topics, we have to first handle the missing retweet path before estimating the parameters using the SEIZ model. We choose the SEIZ model because it has been shown to model both news and rumors on Twitter well (Jin et al. 2013), whereas other system dynamics models focus only on news.

5.3.1 Assumption about the Retweet Path

Twitter only allows us to know the original tweet publisher that a user retweets, but does not let us know from whom the user retweets. Therefore, we need to make an assumption of who a user retweets. We assume that a user retweets his followee who last publishes or retweets the tweet before the user does.

5.3.2 Steady Infusion of the Susceptible in SEIZ Model

We modify the SEIZ model to capture the probability of transition of strong ties and weak ties. We will first explain the original SEIZ model by (Jin et al., 2013) before describing our modification.

The Original SEIZ model

The original SEIZ model by Jin et al. (2013) can be viewed on Figure 5.1. In the SEIZ model, the population of users are divided into four subpopulations. First is **S** the susceptible. The susceptible subpopulation consists of users who are exposed to the tweets we investigate. **I** stands for the infected. The infected subpopulation is users who retweet the tweets. **Z** is the skeptic. They are those who decide not to retweet the tweet because of the influence from those who also do not retweet the tweet. Last but not least, **E** is the exposed. The exposed subpopulation consists of users who get exposed to the news either because of meeting those who retweet the

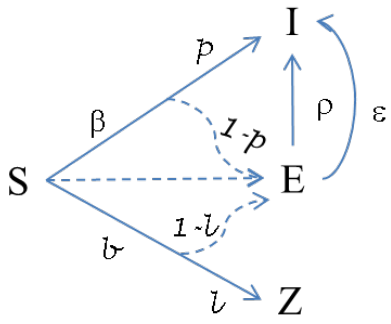


Figure 5.1: The original SEIZ model.

tweet (the infected), or those who do not retweet the tweet (the skeptic). Although the skeptic does not retweet the tweet, they may retweet other tweets that could divert the attention away from the target tweet, or tweets that debunk the target tweet.

The retweet process is described below:

1. The susceptible meets with the infected at the rate β. Some of them get infected with the probability p. Meanwhile, the rest $1 - p$ of them become exposed to the tweet.

2. The susceptible meets with the skeptic at the rate b. Some of them decide not to retweet the tweet with the probability l. Meanwhile, the rest $1 - l$ of them become exposed to the tweet.

3. The exposed subpopulation can meet again with the infected and thereby get infected at the rate ρ. Meanwhile, the exposed can also become infected at the rate ϵ due to the outside influence, or simply due to a change of mind after a certain time.

The Modifed SEIZ Model

Previously, in (Natali et al. 2017), we have made a modification to the SEIZ model that subdivides the **I**, **E**, and **Z** subpopulations into two, those that come from the strong ties, and those that come from the weak ties. We make further modifications

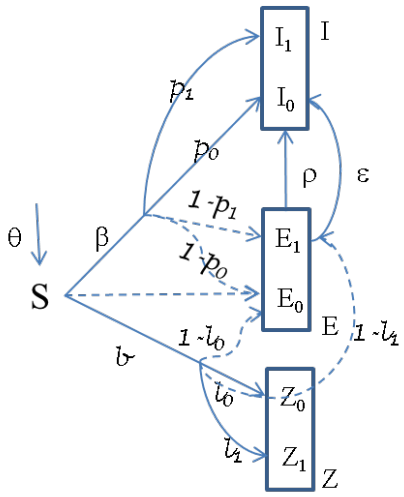


Figure 5.2: The modified SEIZ model.

to the model. The new model can be seen on Figure 5.2. The following describes the retweet process.

1. At each time step except the first time step when the value of the susceptible is estimated, there is an infusion θ that represents new susceptible users entering the system. The numbers of the susceptible entering the system at time $(t+1)$ is equal to the number of followers of the newly infected users at time t.

2. The susceptible meets with the infected at the rate β. If the susceptible and the infected users who meet are weak ties, the susceptible users get infected with the probability p_0. Meanwhile, the rest $1 - p_0$ of them become exposed to the tweet. If the susceptible and the infected users who meet are strong ties, the susceptible users get infected with the probability p_1. Meanwhile, the rest $1 - p_1$ of them become exposed to the tweet.

3. The susceptible meets with the skeptic at the rate b. If the susceptible and the skeptic users who meet are weak ties, the susceptible users become skeptic with the probability l_0. Meanwhile, the rest $1 - l_0$ of them become exposed to the tweet. If the susceptible and the skeptic users who meet are strong ties, the susceptible users become skeptic with the probability l_1. Meanwhile, the rest $1 - l_1$ of them become exposed to the tweet.

4. The exposed subpopulation can meet again with the infected and thereby get infected at the rate ρ. Meanwhile, the exposed can also become infected at the rate ϵ due to the outside influence, or simply due to a change of mind after a certain time.

Our SEIZ model is mathematically represented by the following systems of ordinary differential equations.

$$\frac{dS}{dt} = \theta - \beta S\frac{(I_0 + I_1)}{N} - bS\frac{(Z_0 + Z_1)}{N} \tag{5.1a}$$

$$\frac{dE_x}{dt} = (1 - p_x)\beta S\frac{I_x}{N} + (1 - l_x)bS\frac{Z_x}{N} - \rho E_x\frac{I_x}{N_x} - \epsilon E_x \tag{5.1b}$$

$$\frac{dI_x}{dt} = p\beta S\frac{I_x}{N} + \rho E_x\frac{I_x}{N} + \epsilon E_x \tag{5.1c}$$

$$\frac{dZ_x}{dt} = l_x bS\frac{Z_x}{N} \tag{5.1d}$$

In the equations above x represents the parameters for and subpopulations coming from different types of ties ($x = 1$ represents strong ties, $x = 0$ represents weak ties). Meanwhile, the function that we are trying to minimize is:

$$f = \sum_t \sum_{x=0}^{1} \frac{|I_t^x - rt_t^x|}{I_t^x} + \frac{N_{est}}{N} \tag{5.2a}$$

$$N_{est} = S + \sum_{x=0}^{1} I_{t_l}^x + \sum_{x=0}^{1} E_{t_l}^x + \sum_{x=0}^{1} Z_{t_l}^x \tag{5.2b}$$

In the equations above, x represents the type of tie, t the time step, rt_t^x the number of retweets at time t coming from tie type x, N the real total population, t_l the last time step, and N_{est} the estimate numbers of the total population.

We know the value of I_1^x, rt_t^x and N. The parameters that we are estimating in the models are β, p_0, p_1, b, l_0, l_1, ρ, ϵ, and the numbers of the subpopulation S, I_0, I_1, E_0, E_1 at time step 1. The parameters are estimated using the python function *lsqnonlin* that minimizes the error function 5.2a given other parameters. We first

assume the values of these parameters to be 0.5. We then perform the *forward Euler method* making use of the ordinary differential equations 5.1 to get the value of error function 5.2a that has to be minimized.

In summary, our model makes the following modifications to the original model:

1. We subdivide the population of **I**, **E**, and **Z** into two groups, one for the users that transition from strong ties, and another for the users that transition from weak ties.

2. We add a steady infusion of the susceptible into the model. One of the most unrealistic aspect of all the epidemiological models for information diffusion out there is the steady number of population inside the system. However, as more people retweet (get infected), a new batch of followers (susceptible) comes in. Therefore in this model, at each time step we add a new batch of susceptible that is equal to the number of followers of the new infected users.

3. We normalize all the parameters to 0-1. In the code for the original SEIZ model (Jin et al., 2013) that you can find on the author's website, the lower bound and the upper bound of the parameters differ, making the interpretation of the results difficult.

4. We do not estimate the original value of the infected any more. Instead, we make use of the number of the infected at the first time step as an input to the model. When we crawl the tweets, the retweet process may have already started for some time. As such, we do not know the number of susceptible, exposed, or skeptic at the first time step. Moreover, there is no concrete way on Twitter to categorize a user as susceptible, exposed, or skeptic according to our definition of these groups. However, we have the number of the infected at the first time step. Since the first time step can be at any time after the original tweet was published, we hope that by using the number of infected at the first time step as an input, we can get better estimates of the population of

the susceptible, the exposed, and the skeptic at the first time step.

5. We add the relative total population into the error function that we are minimizing. We get the relative total population by dividing the number of total population predicted by the model with the real total population. We assume that the real total population is the total number of followers of the people who retweet the tweets because they are the susceptible who can transition into other groups. By matching the total population to its real value, we hope to get better estimates of all the parameters.

5.3.3 Handling of Missing Data

As has been explained in Section 5.3.1, we assume the most recent retweeter of the tweet before a user retweets as the source that the user retweets. However, as some accounts are private, or blocked, and some retweet paths are missing, there are cases where we cannot find the source from which a retweet comes. In other words, no followee of the retweeter retweets before the retweeter does. In this case, we will assume that the retweet comes from a weak tie, because as we have seen from the existing data, there are more infected users coming from the weak ties than from the strong ties. It is important to note that these numbers do not reflect the percentage of users that transform into retweeters after receiving a tweet.

Parameter Identification

We identify the parameters exactly in the same way that Jin et al. (2013) did them on their paper. Here's how. The set of parameter values chosen are those that minimize $|\mathbf{I}_0(t) - \mathbf{tweets}_0(t)| + |\mathbf{I}_1(t) - \mathbf{tweets}_1(t)| + (N - (S(t_l) + E(t_l) + I(t_l) + Z(t_l)))$. $\mathbf{I}_0(t)$ is the total number of predicted tweets retweeted by weak ties at each time step t, and $\mathbf{I}_1(t)$ is the total number of predicted tweets retweeted by strong ties at each time step t. $S(t_l) + E(t_l) + I(t_l) + Z(t_l)$ is the final population as predicted by the model. Meanwhile, N is the true population that is achieved by adding all

Table 5.1: Input and latent parameters of the modified SEIZ model.

Input parameters	Latent parameters
tweets$_0(t)$, **tweets**$_1(t)$, N, θ	$\mathbf{I}_0(t)$, $\mathbf{I}_1(t)$, $S(t)$, $E(t)$, $I(t)$, $Z(t)$, β, p_0, p_1, b, l_0, l_1, ρ, ϵ

the followers of infected users (all the susceptible people). Initial population $S(t_0)$, $E(t_0)$, $I(t_0)$, $Z(t_0)$, are considered as unknowns and treated as parameters. The *lsqnonlin* function performed the least squares fit, while the ODE systems were solved with a forward Euler function. All the parameters are initialized.

Lsqnonlin is a matlab function that tries to find the values of x (in this study, all the transition probabilities and contact rates) so that the sum of squares of the values of $|I_x(t) - tweets_x(t)|$, is minimum. x can be 0 or 1. At each iteration of *lsqnonlin*, the forward Euler method estimates the other parameter values (the number of population in each compartment) through solving the ODE system. The forward Euler method states that the value of $I_1(t+1)$ is equal to the value of $I_1(t)$, added with a constant h times $f(I_1(t))$. In our case, $f()$ is the ODE system we have derived. We set h to be 0.1. So, the total step would be (end timestep - start timestep)/0.1. The original code by Jin et al. (2013) can be found on her homepage. We modify this original code for our studies.

Table 5.1 summarizes the input parameters and latent parameters in our model. Input parameters are observable, whereas latent parameters are estimated by the model.

5.4 Case Studies

Several studies have shown that information diffuses through strong or weak ties depending on the topic. Political discussion is more likely to flow among strong ties (Straits 1991), and so is information that challenges one's status-quo (Krackhardt 1992). On the other hand, inter-departmental information is more likely to be promoted by weak ties (Friedkin 1982). There is a category of topic that describes both

of the topics found to be mostly promoted by strong ties in the offline world, that is, a controversial topic. A controversial topic usually challenges one's status-quo that it creates a controversy in the first place. Moreover, many controversial topics are also political. Therefore, from October 12, 2016 to December 2, 2016 we crawl tweets that contain both controversial and non-controversial hashtags. The following sections will describe concisely the topics we consider in our study. Meanwhile, all the tweets analysed can be viewed in Appendix B.1.1.

5.4.1 Controversial Topics

The end of the year 2016 proves to be a divisive year in America. The presidential election brought up so many issues that became the sore points of many different stakeholders. Moreover, a controversial decision by the Britain to get out of the European Union popularly dubbed as *brexit* also just happened. Therefore, we do not lack for hashtags to represent controversial topics. Some of the hashtags we use are: *trump*, *clinton*, *brexit*, *immigrants*, and *BlackLivesMatter*. Overall we collect nine controversial tweets. Below are the examples of some controversial tweets:

Tweet 1: *Well there you have it. A highly intelligent experienced woman just debated a giant orange Twitter egg. Your move America. #debate*

Tweet 2: *Time to #DrainTheSwamp in Washington D.C. and VOTE #TrumpPence16 on 11/8/2016. Together we will MAKE AMERICA SAFE.*

Tweet 3: *Retweet if you are: -A woman -An immigrant -LGBT+ -Muslim -African American -Latino/Latina -In any other way completely terrified right now*

5.4.2 Non-controversial Topics

There are many non-controversial topics on Twitter. We crawl three categories of them.

General News

Twitter has been a popular medium to spread news (Kwak et al. 2010). There-

fore, we analyse whether the preference of ties applies to the diffusion of news. Overall we collect three tweets related to news. Two news are about natural disaster, and one is about the recent visit of little kids to the White House to showcase their scientific ability.

Tweet 1: *We were out here praying for Florida to stay safe from hurricane Matthew. Little did we know. Hurricane Matthew was...*

Tweet 2: *hurricane chris really 10 steps ahead of us all*

Tweet 3: *Check out my newest science advisors! These kids are fearless in using science to tackle our toughest problems.*

Personal

There are two types of tweets that we consider as personal on Twitter. The first type is the tweets about personal life on Twitter. Some people share about their daily lives and activities on Twitter. Although these tweets rarely get retweeted, in some cases they do. The second type is the tweets about the issues that become a genuine concern for some people although it may not directly impact their life. For example, some people tweet about natural disaster encouraging donation. Although the natural disaster is in no way affecting the retweeter, he is so concerned that he encourages others to donate (most likely he also donates himself). Overall we have five tweets under the personal topic. Some examples can be seen below.

Tweet 1: *Florida just got hit by a category 5 Hurricane! Please donate.*

Tweet 2: *It takes 3.2sec to retweet and help find missing Isabella Gonzalez she went missing from #Vegas #usa a year ago today*

Tweet 3: *I remember always running around the house in my underwear and playing games with Ashton and Brandon*

Entertainment

Many entertainers advertise their activities on Twitter. We include tweets about music and sports as the tweets under entertainment topic. Overall we have nine tweets under entertainment. Some examples are:

Tweet 1: *All Weekend Long: Official Music Video*

Table 5.2: Tweets for Case Studies

Category	#Tweets	#Retweets Crawled		#Retweets		First Retweet* (in s)		#Retweeters	
		Mean	Sd	Mean	Sd	Mean	Sd	Mean	Sd
Controversial	9	41,305	33,879	72,776	69,221	111	136	41,106	33,880
News	3	17,063	9,005	22,951	15,672	38	36	17,012	9,011
Personal	5	12,610	7,016	18,194	11,846	77,276	154,086	12,464	6,854
Entertainment	8	26,445	19,477	23,285	14,892	7	4	19,176	12,015
Rewards	5	10,872	6,280	12,515	7,433	241	471	10,809	6,258
*The number of seconds that have passed after the original tweet is published.									

Tweet 2: *English football's most successful clubs are showing why they are meeting on a Monday night in Champions League week*

Tweet 3: *Cubs win! We take a 3-2 #NLCS lead! Final: #Cubs 8 #Dodgers 4. #FlyTheW*

Rewards

There are also tweets that promise rewards to increase circulation. These tweets usually spread for marketing purpose. We want to find out whether there is a preference for strong or weak ties in circulating such a tweet. Overall we have five such tweets. Some examples are shown below.

Tweet 1: *Wrigley Field will be loud tomorrow. RT this for your chance to win two tickets to #NLCS Game 6! #FlyTheW*

Tweet 2: *RT TO WIN: OYSTER BRUSH ROLL FROM SPECTRUM ($100+) ? (must be following me & @SpectrumBrushes so we can dm winner)*

Tweet 3: *RT TO WIN: ABH GLOW KIT OF CHOICE ? (must be following me to win)*

5.5 Analysis and Results

Table 5.2 reports some statistics about the tweets collected. Thirty tweets of five different topics are collected. Controversial tweets get retweeted most, whereas rewards and personal tweets get retweeted least. It is therefore, unsurprising to see that around $40 - 50\%$ of the controversial retweets are missing in the crawling process, because twitter crawler can only crawl a percentage of all streaming tweets.

The rest of the categories on average has more than 70% of their retweets crawled.

The retweet of the tweets under the entertainment category happens faster than the retweet of other categories. Within seven seconds after an entertainment tweet is published, the first retweet happens (See Table 5.2). Meanwhile, personal tweets get retweeted slower. Only after 21 hours a personal tweet has been published, the first retweet happens. Meanwhile, the other categories have their first retweet on average few minutes after the tweet is published.

By default, Twitter only allows a user to retweet a tweet once. However, as you can see on the Table 5.2, the number of retweets crawled are always slightly higher than the number of retweeters. It means, that there are cases where users retweet, undo the retweet, and retweet again. Such cases happen mostly in the diffusion of entertainment tweets. Re-retweeting can grow influence by increasing the likelihood of a tweet to be on the top of the followers' timeline. It appears that some retweeters of the entertainment tweets are particularly eager to increase the tweets' influence and advertisement through a quick first retweet, and re-retweeting.

5.5.1 Homogeneous vs. Heterogeneous Ties

The goal is to assess whether assuming homogeneous or heterogeneous ties gives a better result when implementing the SEIZ model. To do so, we compare the error that happens when the original SEIZ model is used, and when the modified SEIZ model is used. The error of the model is calculated by the following equation formulated by Jin et al. (2013).

$$e = \frac{\sqrt{\sum_t (I_t - rt_t)^2}}{\sqrt{\sum_t rt_t^2}} \tag{5.3}$$

The equation above calculates the Ecludian norm of the errors at all time steps normalized by the Ecludian norm of the true value of the retweets at all time steps.

In the modified SEIZ model we have two error estimates, one for strong ties who are infected, and another for weak ties who are infected. To calculate the average

Table 5.3: SEIZ Model Results

Topic	id*	SEIZ err	mod err**	p_0	p_1	l_0	l_1	Cluster
Controversial	11072	**0.27**	**0.01**	0.00	1.00	1.00	0.00	0
	12512	**0.5357**	**0.14**	0.47	1.00	1.00	1.00	2
	94368	0.01	3.66	0.00	1.00	1.00	0.99	2
	67680	0.06	2.05	0.49	1.00	1.00	0.16	2
	65409	0.01	1.15	0.00	1.00	1.00	0.00	2
	90400	0.01	0.01	0.42	1.00	1.00	1.00	0
	21408	0.02	89.80	0.00	0.00	0.27	1.00	0
	99648	**0.60**	**0.29**	0.00	0.00	1.00	0.22	3
	24992	**0.04**	**0.01**	0.29	1.00	0.96	1.00	2
News	27456	**0.23**	**0.02**	1.00	0.00	1.00	1.00	2
	07328	**1.00**	**0.96**	1.0000	1.0000	1.0000	1.0000	1
	35588	0.01	0.14	1.00	1.00	1.00	0.48	2
Personal	08385	**0.01**	**0.01**	0.34	0.60	1.00	1.00	0
	96992	0.05	0.26	1.00	1.00	1.00	0.00	0
	58560	**1.00**	**0.84**	0.00	0.82	1.00	0.58	1
	22432	0.10	0.21	0.71	1.00	1.00	0.40	3
	46528	**0.29**	**0.28**	0.00	1.00	1.00	1.00	2
Entertainment	55136	**0.07**	**0.01**	0.00	1.00	1.00	1.00	2
	38880	0.01	0.24	1.00	1.00	1.00	1.00	2
	39264	**0.23**	**0.09**	0.13	0.00	0.97	1.00	2
	36069	**0.31**	**0.09**	0.66	0.62	1.00	1.00	0
	79424	0.07	29.58	1.00	0.00	0.65	1.00	2
	46656	0.00	0.08	0.00	0.48	0.93	0.00	2
	54821	0.30	15.43	0.00	0.00	0.98	1.00	2
	27680	0.01	198.21	0.86	0.00	0.60	0.00	3
Rewards	97216	0.03	0.03	0.00	0.00	1.00	1.00	2
	86144	**0.16**	**0.09**	0.64	1.00	0.98	1.00	2
	13504	0.01	0.02	1.00	0.54	1.00	0.87	2
	65888	0.02	0.21	1.00	1.00	1.00	1.00	2
	30240	0.00	1.58	0.00	0.59	0.88	0.00	0

*The last 5 digits of tweet id.

**mod error means the error of modified SEIZ.

error we use the following formula.

$$e = \frac{\sqrt{\sum_t ((I_t^0 + I_t^1) - rt_t)^2}}{\sqrt{\sum_t rt_t^2}} \tag{5.4}$$

On Table 5.3 we show the comparison of the error in the basic model and in the modified model. In 17 out of 30 cases, the basic model performs better. The modified model predicts $\frac{4}{9}$ controversial tweets, $\frac{2}{3}$ general news, $\frac{3}{5}$ personal tweets, $\frac{3}{8}$ entertainment tweets, $\frac{1}{5}$ rewards tweets better than the basic model does (See the numbers in bold). Overall, 56% tweets diffuse assuming homogeneous ties. A user does not care who retweets, as long as the tweets are interesting the user will retweet. Topic-wisely, separating strong and weak ties works best for general news, and personal tweets. The modified model performs almost as well as the basic model for entertainment and controversial tweets. Meanwhile, rewards tweets generally do not discriminate ties. Everyone likes rewards either the rewards are offered by close friends or strangers. Therefore, the discovery that rewards tweets do not discriminate ties is not surprising.

5.5.2 Strong Ties vs. Weak Ties

Table 5.3 shows the likelihood of a tweet to flow through strong ties or weak ties. There are four parameters that indicate this likelihood, p_0, the likelihood of retweeting given that a user meets a retweeter and the tie is weak, p_1, the likelihood of retweeting given that a user meets a retweeter and the tie is strong, l_0 the likelihood of not retweeting given that a user meets a non-retweeter the tie is weak, and l_1 the likelihood of not retweeting given that a user meets a non-retweeter and the tie is strong. As has been explained in our model (see Section 5.3.2), these probabilities are by no means complementary. The complement of these probabilities are the probabilities of getting exposed: $1 - p_0$ is the likelihood of getting exposed given that a user meets a retweeter and the tie is weak, $1 - p_1$ the likelihood of getting exposed given that a user meets a retweeter and the tie is strong, $1 - l_0$ is the likeli-

hood of getting exposed given that a user meets a non-retweeter and the tie is weak, $1 - l_1$ is the likelihood of getting exposed given that a user meets a non-retweeter and the tie is strong.

A tweet is more likely to flow through strong ties if $p_1 > p_0$ or $l_0 > l_1$ (See numbers in shade). If $p_1 > p_0$ and $l_1 > l_0$, or $p_1 < p_0$ and $l_1 < l_0$, we assume that a user does not have a preference whether to retweet through strong or weak ties. From this definition you can see that most tweets flow through strong ties. To be precise, $\frac{19}{30}$ tweets are more likely to flow through strong ties. Out of them, there are $\frac{8}{9}$ of controversial tweets, $\frac{1}{3}$ general news, $\frac{5}{5}$ personal tweets, $\frac{2}{8}$ entertainment tweets, and $\frac{2}{5}$ rewards tweets. If we only consider the tweets that in Section 5.5.1 are modelled better by the modified SEIZ, then $\frac{9}{13}$ tweets are more likely to flow through strong ties. Out of them, there are $\frac{4}{4}$ controversial tweets, $\frac{0}{2}$ general news, $\frac{3}{3}$ personal tweets, $\frac{1}{3}$ entertainment tweets, and $\frac{1}{1}$ rewards tweets.

The results show that strong ties play an important role in the diffusion of controversial and personal tweets. On the other hand, the diffusion of general news and entertainment tweets depend more on the weak ties, or happen without discriminating ties. Meanwhile, rewards tweets do not generally distinguish ties.

5.5.3 Conclusion

Overall, we have seen that (a) half tweets do not discriminate strong ties and weak ties when diffusing, (b) if they do, strong ties are the dominant diffuser of tweets, precisely in 69% of the cases. SEIZ model shows that weak ties are a likely diffuser of entertainment tweets and general news, but strong ties are the likely diffuser of controversial and personal tweets.

Although our experiment gives interesting and intuitive results, the size of the dataset limits the validity of the conclusion. To increase the size of the dataset, we require a good algorithm that can discover tweets in various categories with high accuracy.

Second, we can try several other reliable models to test the validity of results across models. There are many retweet models. Each model can produce different conclusions. For example, a retweet model by Peng et al. (2011) and Zhao et al. (2011) produce a conclusion that strong ties are more likely to diffuse information whereas a retweet model by Shi et al. (2014) concludes that it's the other way around.

Future works can be directed at improving the validity of our results by the two ways above.

Chapter 6

Essay 2B: Offline versus Online: A Paradigm for Meaningful Categorization of Ties for Retweets

6.1 Introduction

Retweets have long been an important research topic in the social media sphere [1]. With the emergence over the last decade of online social network platforms like Facebook and Twitter, online interactions have produced large volumes of data, offering researchers the opportunity to examine the information users have shared. As a result, information dissemination has become a prominent area of study in the field of social media analysis. But since social media sites gather and spread information in different ways, the methods they use to disseminate it must be considered. Consequently, Twitter's "retweet" function has become a hot topic of study among certain researchers. Retweeting is one of the most popular ways of disseminating information on Twitter, a social media and microblogging site that is widely used to circulate news and other media (Kwak et al. 2010), as well as more personal notes

[1] This study is not yet published. In this thesis, *we* refers to me and the chair of my committee who has been involved in the study

to friends and family. A retweet is a re-posting of a tweet on your feed, and so the feature allows you and others to share selected tweets with your followers. You can retweet your own tweets or tweets from someone else [2].

Understanding retweets is important since they are used for various practical purposes such as sharing news, promoting political views, marketing products, and tracking real time events. Java et al. (2007) attributed the high volume of tweets mostly to daily chatter, although tweets still usually contained a fair amount of news items. Enli and Skogerbø (2013) explored Twitter and Facebook as arenas for political communication. Thomases (2009), meanwhile, wrote a guide book about how to create a successful Twitter marketing campaign.

Therefore, if the drivers of retweets were understood properly, then harnessing them would bring immense benefits to marketing campaigns and public policy interventions. Boyd et al. (2010) compiled a comprehensive list of the motivations behind retweets. It included making new audiences aware of certain tweets and simply increasing a listener's visibility. In addition to these internal reasons, a number of external attributes also influence retweets, such as URLs and hashtags, and also Twitter accounts' age and follower count (Suh et al. 2010). The study by Kupavskii et al. (2012) determined that influential users with high scores on PageRank – a measure of a website page's importance applied to Twitter follow networks – received more retweets.

In addition to user-based attributes, tie-based attributes also drive retweets. Past research has looked into how different ties bring about retweets. Most determined that strong ties drove retweets (Peng et al. 2011; Zhao et al. 2012), although some concluded that weak ties did (Shi et al. 2014). Meanwhile, Natali et al. (2017) analysed how different ties resulted in different topics getting retweeted. In an extended study of this study (Chapter 5) that utilized a more extensive data, they discovered that Twitter users did not consider ties when retweeting any topic half of the time –

[2]Retweet FAQs `https://help.twitter.com/en/using-twitter/retweet-faqs`

though when they did pay attention to them, the results were largely similar to the previous study. Personal tweets were more likely to be disseminated through strong ties, whereas entertainment and news tweets were more likely to be disseminated through weak ties. These past studies, however, defined strong ties differently. Zhao et al. (2012) used the overlap of neighbours as the indicator of strong ties, while Peng et al. (2011) used mutual mentions, mutual retweets, mutual followers and mutual followees as the indicators of strong ties. Natali et al. (2017) and Shi et al. (2014), meanwhile, used reciprocity of follow ties to define strong ties.

In this study, we focus on different categories of ties, namely offline versus online. We aim to find out if offline and online ties can be used in place of other tie categories that were previously utilised in studies that analysed retweets. These categories of ties are reciprocated and unreciprocated. We discover that offline versus online are indeed better tie categories because they can be distinguished more easily by their retweet patterns. Our study is the first to reveal the retweet patterns of offline friends compared to online friends, and offer another promising way for Twitter users to increase the amount of retweets their tweets receive. They also highlight the importance of the offline-online paradigm when discussing retweets, and demonstrate that this paradigm cannot be replaced by another paradigm, that is the reciprocated-unreciprocated paradigm.

6.2 Background

This section lays out the necessary background on strong ties and how they are defined. The categories analysed in this study are determined by these definitions.

Strong Ties. Granovetter (1973) first introduced the concept of strong ties in his seminal work *The Strength of Weak Ties*. In the study, Granovetter described interpersonal ties as "a (probably) linear combination of the amount of time, the emotional intensity, the intimacy (or mutual confiding), and the reciprocal services which characterize each tie". In addition to this formula, Granovetter emphasized

the uniqueness of strong ties, that was that they had more overlapping friends compared to two individuals selected arbitrarily. Therefore, Granovetter concluded that information that circulated among close friends is usually stale and old.

Measuring Strong Ties on Offline and Online Social Network. There are several ways to measure a tie's strength. The first study to do so is the study by Marsden and Campbell (1984). They discovered that the question of how close a person to another was the best indicator of closeness. Their study applied to the offline setting.

In the online setting, Gilbert and Karahalios (2009) authored the most extensive study on the measurement of strong ties. They particularly studied Facebook. They made use of 74 Facebook variables in order to predict strength of ties. Their method achieved a good accuracy. Meanwhile, Backstrom and Kleinberg (2014) revealed that mutual friends of very intimate friends were rarely unconnected. Their study offered the distance of mutual friends as a potential measure of how intimate two friends are.

Reciprocated versus Unreciprocated. Reciprocated ties have often been used as an easy gauge of strong ties when studying retweets (Shi et al. 2014; Natali et al. 2017). On Twitter, a reciprocated tie appears in a situation where a user follows another user, and he or she is also followed back. On the other hand, an unreciprocated tie appears in a situation where a user follows another user, but he or she is not followed back. When someone follows another person on Twitter, he subscribes to the updates published by that person's account. In this study, the analysis of how reciprocated versus unreciprocated ties retweet will be the baseline for assessing how different the tweet novelty and topic of offline and online ties are.

Offline versus Online. Offline ties are not exactly the same as reciprocated ties, although reciprocated ties can predict offline ties with 73% precision and 65% recall. No one has previously studied how offline versus online friends retweet. In this study, we define *offline friends* as connections on Twitter who have met outside of the internet. The connections include both reciprocated and unreciprocated connections. Meanwhile, *online friends* are connections on Twitter who have never met

outside of the internet.

6.3 Dataset: Two-Hop Retweet Data

Determining whether each tie involved in a full retweet chain is offline or online is impossible; however we can find out if the ties in a retweet chain in an ego network are offline or online. In this study we use a dataset gathered in Section 3.2. The first dataset is the dataset of 98 Twitter users in 2011, and the second dataset is the dataset of 41 Twitter users in 2015.

The Twitter users that are involved in the 2015 dataset are chosen to be specifically active in tweeting. They should tweet at least once in a week. However, unlike in the previous dataset where only 14% of the users are private, in the new dataset, 37% of the users are private accounts. We combine the old dataset and the new dataset to get the overall dataset for our analysis.

We crawl the tweets of all the users in our dataset on March 2018. Additionally, we also crawl the latest follow-edges among these users.

6.4 Methodology: Calculating Retweets Depth and Quantifying Retweets Topic

Before proceeding to the methodology, we will recap the issues our research focuses on. In this study, we want to reveal the retweet patterns of offline and online friends on Twitter. Specifically, we want to know the difference in the tweet novelty and retweet topic of offline and online friends. We also want to know whether this difference is greater than the difference between the retweet patterns of reciprocal and unreciprocal friend categories

However, due to the limitation of the dataset explained in Section 6.3, we cannot analyse the whole retweet chain. Therefore the analyses performed will have the following limitations:

1. We can only analyse retweet patterns that happen among Twitter users in an ego network.

2. We can only analyse retweet patterns that go through public accounts, since their edges cannot be crawled otherwise.

3. Only when a retweet passes from or to an ego user, can we know whether a retweet passes through an edge that represents an offline or an online friendship. If the retweet does not come from or go to an ego user, we will only know whether the retweet passes from an ego user's offline or online friend, to another offline or online friend (See Figure 3.1).

Given these limitations, there are seven categories of ties that we analyse in this study.

1. *Offline ties* that represent connections on Twitter who know one another offline.

2. *Online ties* that represent connections on Twitter who do not know one another offline.

3. *Offline-to-offline ties* that represent connections on Twitter between an ego user's offline friend and another offline friend.

4. *Online-to-offline ties* that represent connections on Twitter between an ego user's offline friend and an ego user's online friend.

5. *Online-to-online ties* that represent connections on Twitter between an ego user's online friend and another online friend.

6. *Reciprocated ties* that represent connections on Twitter between two users in which the users follow one another.

7. *Unreciprocated ties* that represent connections on Twitter between two users in which only one user follows another.

As Twitter only reveals the original source of a tweet, and not from whom a retweeter retweets, we must make several assumptions to construct a retweet chain. We use these three:

1. *Latest timing*. In this assumption, the followee of a user who retweets something just before the user retweets, is assumed to be the source of the retweet. If there are no retweeters in the ego network who retweet before the user retweets, the original source of the retweeted tweet is considered. If the original source is a followee, he is considered as the source of retweet. Otherwise, the source of the retweet is unknown.

2. *Earliest timing*. We assume that a user's followee who retweets first is the source of the retweet. Therefore, if the original source of the retweeted tweet is a followee, then the original source is always the source of the retweet. If no followees are retweeters or an original source, then the source of the retweet is unknown.

3. *Most popular*. We assume that a user's followee who tweets or retweets before the user retweets a tweet, and has the most followers is the source of the retweet. When there are no followees who tweet or retweet before the user does, then the source of the retweet is unknown.

Figure 6.1 is used to illustrate these three assumptions. In the figure, each level represents the time a tweet is retweeted, with t_0 representing the time when the tweet first originates. Therefore, User B is the original source of tweet. The edges are the follow edges that exist among the nodes. Assuming that there are no other follow edges among the nodes outside the system, User C is the most popular. Based on this configuration, the source of retweet for User D is User A based on the latest timing assumption, User B based on the earliest timing assumption, and User C based on the most popular assumption.

In our analysis, we are concerned only with the retweet chain in an ego network. Therefore, all the analyses are based on the assumption that **a retweeter's source of**

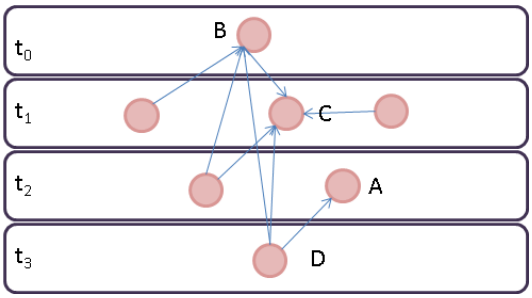


Figure 6.1: Illustration of different assumptions for constructing a retweet chain.

a retweet can only come from the ego network being analysed. The reason we make such an assumption is because, we do not know the category of friendship that exists between the source of the retweet outside an ego network and the retweeter, that is, whether it is offline or online. By applying this assumption, we may not get the user who is the true source of the retweet, but we will get the user in an ego network who has the highest likelihood of being the source of the retweet.

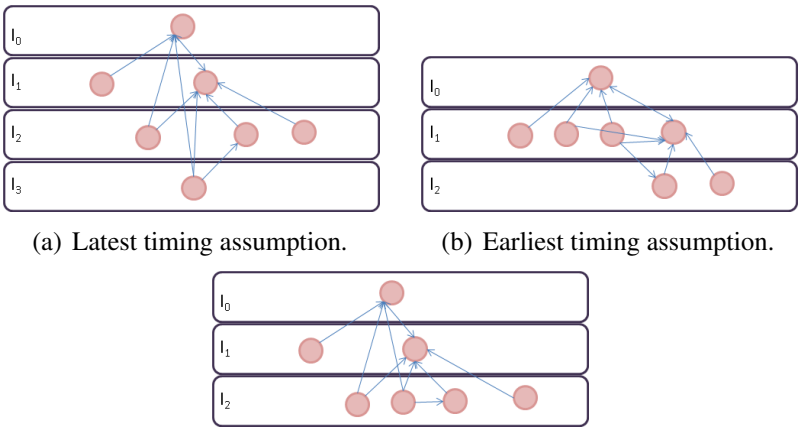
In this study, we need to calculate the depth of retweet chains and quantify retweet topics. Now, we will explain how to do these both sequentially.

6.4.1 Calculating The Depth of Retweet Chains

The depth of a retweet chain refers to the deepest level of a retweet chain. Each level represents not the time of a retweet, but the sequence of one. The value can change depending on the assumption that we make. If we stack nodes in Figure 6.1 by depth level and, not by the time of a retweet, we will come up with Figure 6.2. Figure 6.2 shows the depth level of different assumptions. The depth of the retweet chain is three if we use the latest timing assumption, and two if use other assumptions.

The depth of a retweet chain represents the greatest degree of separation that can be reached by the source of a tweet. The depth of the retweet chain represents tweet novelty. The deeper the level at which a user retweets, the longer the tweet has circulated among friends who are directly or indirectly connected to the user.

In this study, we calculate the frequency of different tie categories at each level



(a) Latest timing assumption. (b) Earliest timing assumption.

(c) Most popular assumption.

Figure 6.2: Levels of depth given different assumptions.

of depth for each assumption. We symbolize this frequency as f_l^c, where l represents the level of depth, a value that can range from one to infinity and c represents the frequency of ties that belong to the category c.

To ensure that the difference in the frequency of ties used for retweets is not due to the difference in the frequency of ties in the networks, we will normalize the frequency by N_c – the frequency of ties that belong to the category c in the networks. We symbolize the normalized f_l^c as $\hat{f}_l^c$ (See Equation 6.1). f_l^c represents the proportion of ties in those networks that belong to category c and are used for retweets.

$$\hat{f}_l^c = \frac{f_l^c}{N_c} \tag{6.1}$$

6.4.2 Quantifying Retweet Topics

In this study, we also want to find out how well different tie categories can be distinguished by topics. Therefore, we apply Twitter-LDA (Zhao et al. 2011) to extract topics from the tweets that are retweeted by various tie categories. From implementing Twitter-LDA to process the tweets, we get out 15 topics that are listed in Table 6.1. For the complete list of words in each topic, please refer to Appendix B.2.1.

In this study, we also want to know how well different categories of ties can be distinguished by topics. Therefore, we apply Twitter-LDA to extract topics from the tweets that are retweeted by various categories of ties. We come up with 15 topics that are listed in Table 6.1.

Table 6.1: Extracted topics from tweets.

Code	Topic	Sample Words
P0	Sexually explicit words	girl, love, baby, hot, fuck
P1	Shows and videos	live, tonight, youtube, video
P2	Global news	new york, trump, people, news
P3	Singapore politics	singapore, lee, pm, pap
P4	Sports	team, great, chicago, race
P5	Singapore news	people, police, singapore, man
P6	Education and Jobs	students, education, school, work
P7	Global politics	trump, president, obama, india
P8	Stocks	latest, price, bitcoin, usd
P9	Traffic and weather	singapore, time, weather, rain
P10	Fun and socialize	song, tonight, happy, guys
P11	Technology	apple, iphone, app, google
P12	Friends and daily life	people, happy, life, day
P13	Social media	tech, social, google, online
P14	Family and finance	money, day, food, children

In addition to churning these 15 topics out, Twitter LDA also produces the distribution of these tweet topics for each set of tweets retweeted by different tie categories.

6.5 Results: Categorizing Ties for Retweet

In this Section, we will discuss the results of calculating the depth of the retweet chains and quantifying retweet topics of tweets that belong to different tie categories.

6.5.1 "Offline versus Online" as the Category of Ties by Tweet Novelty

Table 6.2 calculates the normalized frequency of ties that belong to category c at depth level l ($\hat{f}_l^c$) in terms of percentage. c can be offline, online, offline-to-offline, online-to-offline, or online-to-online. Therefore, the value 28.33 in the first cell means that 28.33 % of offline ties are used to retweet at depth level 1. A user who retweets at depth level 1 is the first to retweet the followee that fulfils the latest timing assumption. The followee can be the source of the tweet (as shown in Figure 6.2) or not, as long as it is the start of a retweet chain.

The results show that there are more depth levels produced when the latest timing assumption is used. Some of the users who retweet at the higher level when the earliest timing assumption or the most popular assumption is used, now retweet at the lower level. This means that the users at the lower levels have more followees who retweet at the upper levels.

The results also demonstrates that a greater percentage of offline ties are used to retweet compared to online ties. Meanwhile, the greatest percentage of ties that are used to retweet are the online-to-online ties. However, when the latest timing assumption is used, offline-to-offline ties have the greatest percentage of retweeting ties compared to other ties at the lower depth levels (depth level ≥ 5). Such results indicate that offline-to-offline ties are more likely to retweet older news that has been retweeted by their friends at earlier times. We previously concluded that the difference in depth levels of different assumptions means that the users who retweet at the lower depth levels have more friends at upper levels.

A previous study by Natali and Zhu (2016) discovered that a user's offline friends were more highly connected on Twitter than a user's online friends. Therefore, we can conclude that friends who are likely to be offline (offline-to-offline ties) are more likely to retweet older news. Meanwhile, although a Twitter user's online friends are not as connected as their offline friends (Natali and Zhu 2016), they are

Table 6.2: Normalized frequency of ties that belong to the offline-online categories at depth level l ($\hat{f}_l^c$) in terms of percentage.

Depth Level	Latest timing assumption				
	off	on	off-to-off	on-to-off	on-to-on
1	28.33	17.57	22.93	28.19	58.29
2	2.17	0.55	0.94	0.63	1.98
3	0.09	0.02	0.16	0.06	0.31
4	0.02	0.01	0.06	0.01	0.06
5	0.00	0.00	0.03	0.00	0.01
6	0.00	0.00	0.02	0.00	0.00
7	0.00	0.00	0.01	0.00	0.00
8	0.00	0.00	0.01	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00
>= 11	0.00	0.00	0.01	0.00	0.00
	Earliest timing assumption				
1	28.64	17.94	23.49	28.83	60.08
2	1.37	0.26	0.31	0.20	0.57
3	0.04	0.02	0.01	0.00	0.02
4	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00
	Most popular assumption				
1	28.37	17.82	23.27	28.66	59.41
2	1.53	0.32	0.47	0.35	1.18
3	0.02	0.02	0.03	0.01	0.09
4	0.00	0.00	0.01	0.00	0.01
5	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00

the best circulator of information on Twitter networks at higher depth levels (depth level ≤ 4). These results support Granovetter's theory that strong ties confine information circulation within local clusters (Granovetter 1973). As such, novel news typically comes from weak ties.

These results also align with the study by Bakshy et al. (2012) that shows novel information is more likely to spread through weak ties but strong ties are generally better at diffusing information.

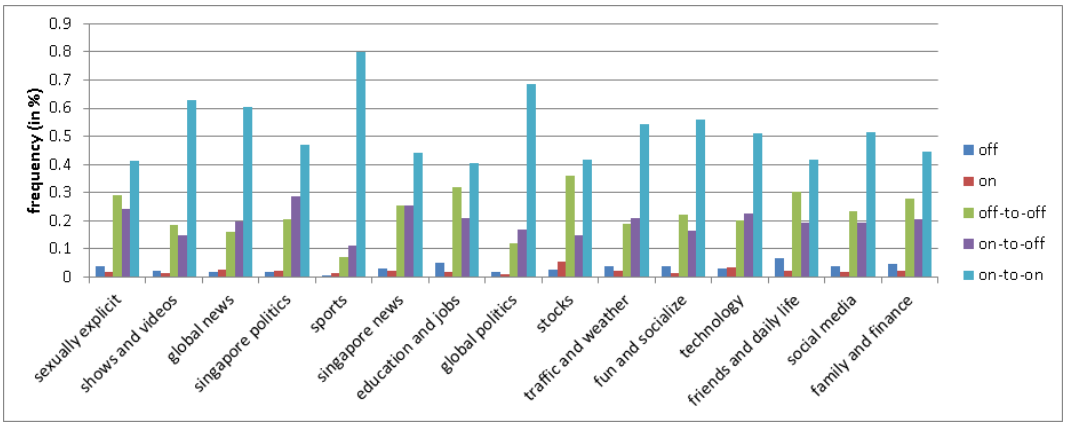


Figure 6.3: Frequency of tweets by offline-online categories.

6.5.2 "Offline versus Online" as the Category of Ties By Topic

We plot the topic distribution of tweets retweeted by ties that belong to the offline-online categories on Figure 6.3. Twitter-LDA gives us f_c^t, the frequency of tweets of topic t retweeted by ties belonging to category c. We normalize the frequency by f^t, the total frequency of tweets of topic t.

Across all topics, online-to-online ties dominate retweets, confirming the results in Section 6.5.1 that show these types of ties prompt the most retweets. The results reveal another pattern, demonstrating that a high frequency of offline ties usually indicates a high frequency of offline-to-offline ties, as well. This phenomenon appears in many topics, including "sexually explicit", "shows and videos", "education and jobs", "fun and socialize", "friends and daily life", "social media", and "family and finance". We conclude that these topics are more likely retweeted by offline ties, or the friends a user engages with outside of the internet.

Additionally, "global news", "Singapore politics", "sports", and "technology" are topics that are likely to be retweeted by online-to-offline ties or online ties. Meanwhile, other topics point to mixed results. Although the topics of "Singapore news", "global politics", and "traffic and weather" are more likely to be retweeted by offline ties than online ties, they are more likely to be retweeted by online-to-offline ties than offline-to-offline ties. Meanwhile, although the topic "stocks" is more likely to be retweeted by online ties than offline ties, it is more likely to be

retweeted by offline-to-offline ties than online-to-offline ties.

When we compare these results to the research work conducted by Natali et al. (2017), we can see some similarities as well as discrepancies. Natali et al. (2017) discovered that personal tweets were more likely to be disseminated through the stronger ties (reciprocated ties). In our study, personal tweets on topics such as "fun and socialize", "friends and daily life", and "family and finance", are also more likely to be disseminated through stronger ties (offline ties). However, while Natali et al. showed that entertainment tweets were more likely to be spread through weaker ties (unreciprocated ties), our study demonstrates that entertainment-focused topics ("shows and videos") are more likely to be circulated by stronger ties (offline ties). Yet, a different entertainment topic, "sports" is more likely to be disseminated by weaker ties (online ties).

6.5.3 "Reciprocated versus Unreciprocated" as the Category of Ties By Tweet Novelty

In order to discover how the different retweet patterns of "offline versus online" ties compare to those observed in "reciprocated versus unreciprocated" ties, we must analyse the retweet patterns of reciprocated and unreciprocated ties using the same dataset. Table 6.3 calculates the normalized frequency of ties that belong to category c at depth level l ($\hat{f}_l^c$) in terms of percentage. c can be reciprocated or unreciprocated.

The results show that at all depth levels a higher percentage of reciprocated ties are used to retweet when compared to unreciprocated ties. At level one, the percentage is even greater than one hundred, meaning that on average, each tie is used more than one time to retweet. It is also important to remember that the information that flows through reciprocated ties can go two ways, naturally increasing the likelihood of any information passing through. However, even if we increase the frequency of unreciprocated ties in Table 6.3 by a factor of two, the frequency of reciprocated

ties that is used to retweet is still higher at all depth levels.

Therefore, we cannot distinguish reciprocated-unreciprocated ties by tweet novelty, unlike how we can distinguish offline-online ties.

Table 6.3: Normalized frequency of ties that belong to the reciprocated-unreciprocated categories at depth level l ($\hat{f}_l^c$) in terms of percentage.

Depth Level	Latest timing assumption	
	reciprocated	unreciprocated
1	137.25	24.88
2	6.89	0.57
3	1.12	0.11
4	0.28	0.03
5	0.12	0.01
6	0.06	0.00
7	0.03	0.00
8	0.02	0.00
9	0.01	0.00
10	0.01	0.00
>= 11	0.02	0.00
	Earliest timing assumption	
1	143.69	25.34
2	2.15	0.18
3	0.09	0.01
4	0.01	0.00
5	0.00	0.00
6	0.00	0.00
	Most popular assumption	
1	141.41	25.23
2	4.13	0.31
3	0.25	0.04
4	0.03	0.00
5	0.01	0.00
6	0.00	0.00

6.5.4 "Reciprocated versus Unreciprocated" as the Category of Ties By Topic

We plot the topic distribution of tweets retweeted by ties that belong to the reciprocated-unreciprocated categories on Figure 6.4. Twitter-LDA gives us f_c^t, that is the frequency of tweets of topic t retweeted by ties that belong to category c. We

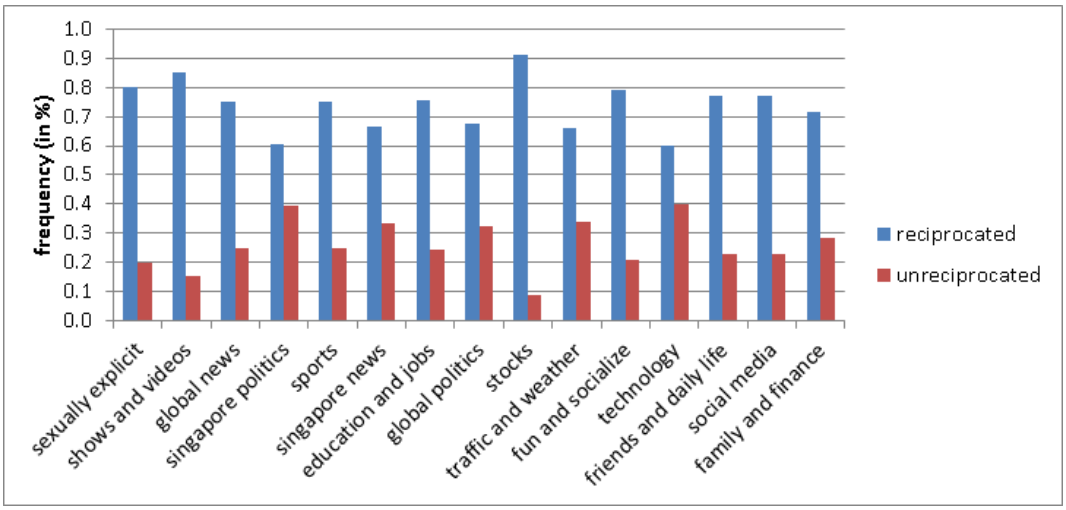


Figure 6.4: Frequency of tweets by reciprocated-unreciprocated categories.

normalize the frequency by f^t, the total frequency of tweets of topic t.

Across all topics, reciprocated ties are used more than unreciprocated ties to retweet. Although these results contradict the results of the research by Natali et al. (2017), they are not necessarily invalidated because the dataset used in this study is different than the one used by Natali et al. The contexts of the two studies are also different. In this study we examine the retweets in ego networks, whereas Natali et al. (2017) analysed the retweet chain across networks within a time period.

In conclusion, reciprocated-unreciprocated ties also cannot be distinguished by topics just as how they cannot be distinguished by tweet novelty. Meanwhile, offline-online ties can be distinguished by both criteria.

6.5.5 Putting It All Together: "Offline or Not and Reciprocated or Not" as Categories of Ties

We also want to know whether combinations of the above tie categories will improve the categorization of ties by making each category more distinguishable from one another.

Table 6.4 calculates the normalized frequency of ties that belong to category c at depth level l ($\hat{f}_l^c$) in terms of percentage. c can be any of the 10 categories made

by combining the offline-online categories and the reciprocated-unreciprocated categories.

Table 6.4: Normalized frequency of ties that belong to the combined categories at depth level l ($\hat{f}_l^c$) in terms of percentage.

Depth Level	Latest timing assumption									
	reciprocated					unreciprocated				
	off	on	off-to-off	off-to-on	on-to-on	off	on	off-to-off	off-to-on	on-to-on
1	26.96	8.03	23.05	19.97	53.06	45.85	32.80	22.17	44.54	70.24
2	2.29	0.49	1.00	0.67	2.27	0.62	0.64	0.59	0.55	1.31
3	0.09	0.01	0.17	0.07	0.34	0.21	0.05	0.06	0.05	0.23
4	0.02	0.01	0.06	0.01	0.07	0.07	0.01	0.03	0.01	0.05
5	0.00	0.00	0.04	0.00	0.01	0.00	0.00	0.00	0.00	0.01
6	0.01	0.00	0.02	0.00	0.00	0.00	0.01	0.01	0.00	0.00
7	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
>= 11	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Earliest timing assumption									
1	27.31	8.09	23.66	20.58	54.88	45.64	33.66	22.44	45.23	71.98
2	1.45	0.36	0.32	0.22	0.63	0.27	0.10	0.23	0.15	0.44
3	0.04	0.02	0.01	0.00	0.02	0.07	0.04	0.01	0.01	0.02
4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Most popular assumption									
1	27.07	8.01	23.41	20.39	53.99	45.02	33.50	22.34	45.11	71.81
2	1.63	0.37	0.50	0.37	1.31	0.34	0.25	0.32	0.29	0.89
3	0.03	0.02	0.04	0.01	0.08	0.00	0.03	0.02	0.00	0.13
4	0.00	0.01	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.02
5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

The results show that reciprocated-unreciprocated categories can help to explain the behaviours of ties that belong to offline-online categories in the context of retweets. At depth level one, a higher percentage of unreciprocated ties is used to retweet compared to reciprocated ties, regardless of which offline-online categories the ties belong to. Meanwhile, at the depth level two, a higher percentage of reciprocated ties is used to retweet. The results for level three and four are mixed.

For some categories a higher percentage of reciprocated ties retweets more, while for other categories a higher percentage of unreciprocated ties retweet more. At the level beyond five, reciprocated offline-to-offline ties are the ones mostly used for retweet.

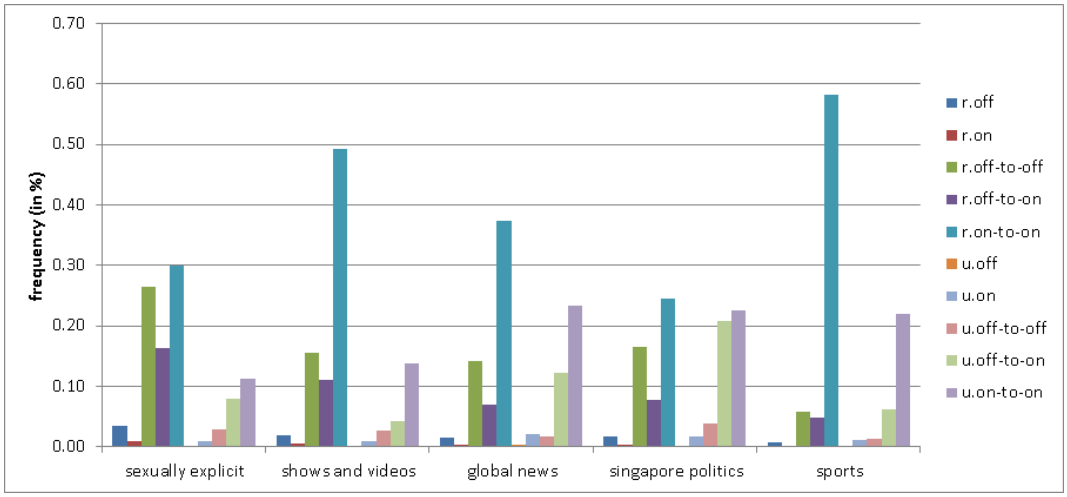
The results can be explained by the theory of weak ties (Granovetter 1973). At depth level one most tweets are novel, and therefore, the weaker ties of each offline-online category are used to retweet. Meanwhile, at depth level five and above, the tweets are old, and therefore, reciprocal offline-to-offline, the strongest category of ties is used to retweet. Although offline-online categories alone cannot distinguish retweet behaviour at depth level one, the combination of offline-online and reciprocal-unreciprocal categories can do so.

We also plot the topic distribution of tweets retweeted by ties that belong to the combined categories (See Figure 6.5). In the combined categories, offline-reciprocated and online-unreciprocated ties are more likely to be retweeted across all topics. We can conclude that combined categories cannot be distinguished by topics as well as the offline-online categories can be.

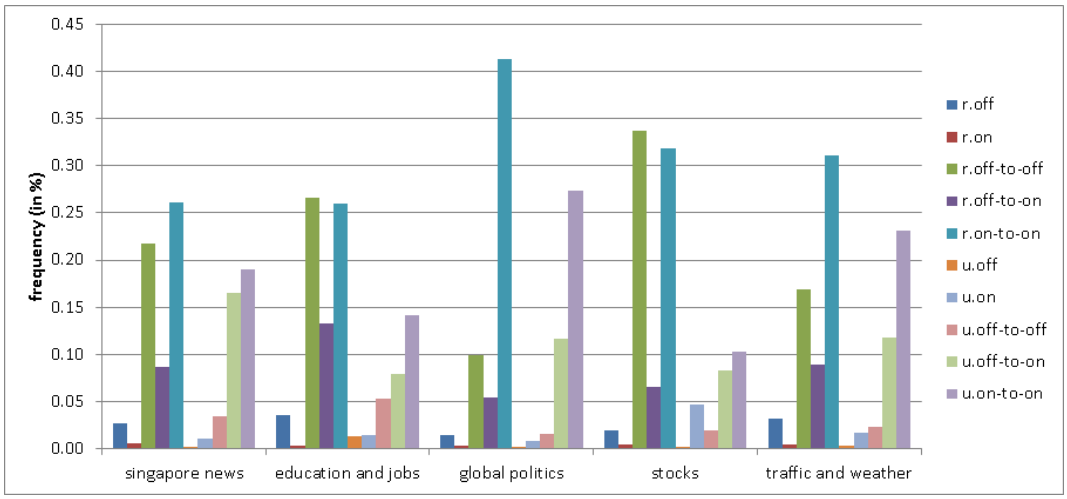
Moreover, the results are also inconsequential. In the combined categories, offline and reciprocal ties are more likely to be retweeted by all topics. At the same time, online and unreciprocal ties are more likely to be retweeted by all topics. In conclusion, combined categories cannot be distinguished by topics as well as the offline-online categories can be.

6.6 Conclusion

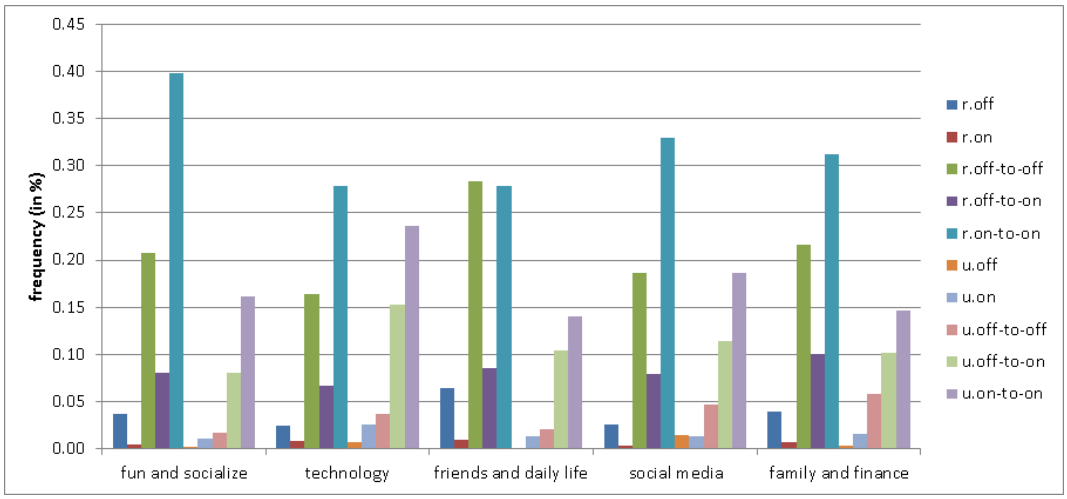
Overall, we have analysed the retweet patterns, specifically tweet novelty and tweet topics, of offline and online ties on Twitter ego networks. We compare our results with the analysis of retweet patterns of reciprocal and unreciprocal ties. We have shown that offline ties and friends who are likely to be offline (offline-to-offline ties) are the ones who tweet old news. However, online-to-online ties play the most im-



(a)



(b)



(c)

Figure 6.5: Frequency of tweets by the combined categories. r stands for reciprocated, u stands for unreciprocated.

portant role in circulating information on Twitter. Offline ties are also more likely to retweet about family and friends, while online ties are more likely to retweet news. On the other hand, reciprocated and unreciprocated ties show similar retweet patterns across tweet topics and tweet novelty. Hence, offline versus online is a more reliable tie category with regard to retweets than reciprocated versus unreciprocated. In terms of practical application, someone who wants to increase a tweet's shelf life and popularise personal tweets should focus more effort on targeting offline friends. Our study highlights the importance of the offline-online network paradigm for retweets that cannot be replaced easily, such as by the reciprocated-unreciprocated network paradigm.

We can further strengthen the importance of offline-online paradigm for retweet by comparing offline-online retweet pattern to other paradigms besides reciprocated-unreciprocated, such as reply-no reply. However, replying data is usually sparse. Therefore, if future works would like to do such analysis, more data may be required. Other efforts can be directed at predicting a tweet topic and novelty given its retweet pattern.

Chapter 7

Essay 3: Measuring Tie Strength Offline vs. Online: Is Redefinition of Tie Strength Necessary on a Social Information Network?

7.1 Introduction

Twitter provides users with a platform to both socialize and retrieve information. On Twitter, people can make friends with strangers, connect with close friends, share information, and exchange conversation. The myriad usages of this network have separated friendships into different spectrum based on their roles, namely those that are mainly used to socialize and those that are mainly used to acquire information. The offline vs. online spectrum has been shown to separate friendships based on their roles on Twitter. In a study by Kim et al. (2016), offline friends are discovered to have a higher number of reply on Twitter. In Chapter 3 and 4, we have seen how a user's offline friends build a denser follow network on Twitter than the user's online friends do. In Chapter 6, we discovered that offline friends retweet personal and controversial tweets, but online friends retweet entertainment tweets. Meanwhile,

offline friends are also more likely to repost old news whereas online friends are more likely to retweet novel news. The follow-up question that comes up after this observation is, does the spectrum that differentiate behaviour on a social information network extend beyond the offline vs. the online realm? If yes, how so? This question can be broken down into three research questions that we will describe in Section 7.2.3.

Our study answers this question and by doing so, sees whether a redefinition of tie strength in a social information network is necessary. Let us give an example how redefinition may happen according to the answers that we may find in our study. If people interact equally with strangers as they do with close friends on Twitter, then the definition of closeness that we know barely carries any meaning on Twitter as it does not manifest into different social behaviours online. However, if close friends interact more on Twitter, then the concept of closeness meaningfully defines how people are going to interact on Twitter.

Knowing the answer, we will figure out how one of the most influential sociological concept, that is tie strength, is redefined when a friendship transfers from the offline world to Twitter. Consequently, we know whether to expect different social behaviours online when we observe different social behaviours offline among friends.

7.2 Background and Research Questions

7.2.1 Strength of Tie: Definition and Measurement

The concept of tie strength was first introduced by Granovetter (1973). In his seminal work, "The Strength of Weak Ties", he stated that the most intuitive notion of the strength of an interpersonal tie should be satisfied by the following definition: A (likely linear) combination of the amount of time, emotional intensity, intimacy (mutual confiding), and reciprocal services that characterize each tie. Operational

measures of tie strength were not discussed in this paper; instead, the paper argued how weak ties (acquaintances) were beneficial in passing on novel information that helps someone to, for instance, find a job. Meanwhile, strong ties (close friends) usually circulated old, stale information and impeded the spread of information beyond their tight-knit group. The power of new information circulated by weak ties was the strength of having weak ties in one's friendship network. When the concept of "tie strength" was first introduced, social media did not exist as it does today. As such, the application of the concept was intended to deal with social relationships in the non-digital world.

Following Granovetter's seminal work, several studies sought to come up with operational measures of tie strength in the physical world. A study by Marsden and Campbell (1984) was the earliest to attempt this, and it concluded that a measure of closeness, or the emotional intensity of a relationship, was the best indicator of the concept of tie strength. In the study, closeness was measured as a trichotomy: (1) An acquaintance, (2) a good friend, or (3) a very close friend. On the other hand, duration and frequency of contact were not good indicators because they overestimated the strength of ties between neighbors and/or co-workers. Friedkin (1990) measured tie strength based on the stage a specific relationship was in: (1) simple awareness, (2) interaction, (3) provision of resources and assistance, and (4) affective attachment. Meanwhile, in a marketing context, (Shi et al., 2009) offered an approach in conceptualizing tie strength by using a tie's robustness or resilience – or how much a tie will adapt multiple media to the requirements of their relationship.

7.2.2 Tie Strength on the Online Social Network

When online social networks became a common phenomena, there were efforts to come up with new operational measures of tie strength specific to online social networks. Both abundance and reliability represent the advantages of having online operational measures for tie strength. In the past, surveys that relied solely on recall

had to be conducted to record interactions among users, whereas now the online footprints of users can be easily acquired. The data of online interactions does not rely on a person's memory – often incomplete and unreliable – but on a computer's perfect memory. Petróczi et al. (2007) developed a virtual tie strength scale based on 11 questions that they claimed provided a reliable measure of tie strength in virtual communities and were capable of distinguishing acquaintances and friendships. Gilbert and Karahalios (2009) built a predictive model on a dataset of over 2,000 social media ties on Facebook that could distinguish between strong and weak ties with over 85% accuracy. The tie strength was assessed through survey questions where participant responses were mapped on a continuous scale of 0-1. Sixty-seven predictive variables, which included various network and interactive measures, were considered. Among them, the top five were: days since last communication, days since first communication, wall words exchanged, mean strength of mutual friends, and educational difference. There are few other efforts in this area, and usually they harness simple statistical methods such as correlation or regression. However, they vary in terms of predictors.

7.2.3 Research Questions

It is unquestionable that offline friendships are stronger than online friendships because online friends are strangers or public figures in the offline world. Empirically speaking, the notion of offline friends being closer to one another than online friends has been demonstrated by Antheunis et al. (2012). However, is this difference in strength reflected in the different behavior observed on the social information network? Past studies have proven that it is.

A previous study investigated how the three principles of network formation, namely reciprocity, preferential attachment, and triadic closure, apply on the social information network to friends who are connected offline, and those who are not (Natali and Zhu 2016). The study confirmed that the social network forma-

tion principles that we see in real life also apply on Twitter – but mainly to friends who were connected offline. Meanwhile, Kim et al. (2016) determined that offline friends were more likely to reciprocate, reply to each other, and share similar friends on Twitter. In previous chapters, we have also shown that these two types of friendships produce different behaviour on the social information network. Offline friends network and interact more. Meanwhile, offline friends spread personal information while online friends spread novel news. All these studies have shown that offline friends and online friends behave differently on Twitter: In other words, offline friends are more social, and online friends more informative. Behaviour on the social information network reflects the tie strength of offline friends vs. online friends.

Figure 7.1(a) sketches how behaviours offline and on Twitter reflects tie strength. In Figure 7.1(a), the rectangle represents the offline world, the circle represents the Twitter environment, and the rectangle in the middle represents tie strength. Meanwhile, the colour represents frequency or strength. The deeper a colour is, the higher the interaction is, or the stronger the tie is. The left side of the bottom rectangle – the side that represents an offline network – is coloured blue, which slowly dissipates into white at the right side of the rectangle – the side that represents strangers. The colour degradation of the rectangle represents how the levels of interactions vary in the offline world, from close friends to strangers. On top of the figure, the offline circle is coloured blue, whereas the online circle is coloured white, reflecting their respective interactions on Twitter.

However, tie strength is not limited only to the spectrum of offline vs. online. In fact, the concept of offline vs. online is quite new considering it only came up after social media became prevalent around the world. In the past, the concept of tie strength applied only to offline friends. Tie strength sorts offline friends by several levels of closeness; categories include acquaintances, close friends, and very close friends (Marsden and Campbell 1984). What we would like to investigate in this study is whether Figure 7.1(b), which represents the increase of interactions

among friends in the offline world across closeness levels, also happens on the social information network. If it does, close friends interact more on the social information network than acquaintances do. Moreover, we are also curious as to whether the degradation of tie strength is only reflected on Twitter but for family members (See Figure 7.1(c)) since family members may not interact on the social information network. For example, parents do not tweet their children.

In a nutshell, our research question is the following:

Research Question. Do different levels of interactions on the social information network explain different levels of tie strength (beyond offline vs. online)? If yes, do they explain them in the same way that different levels of interactions in the offline world do?

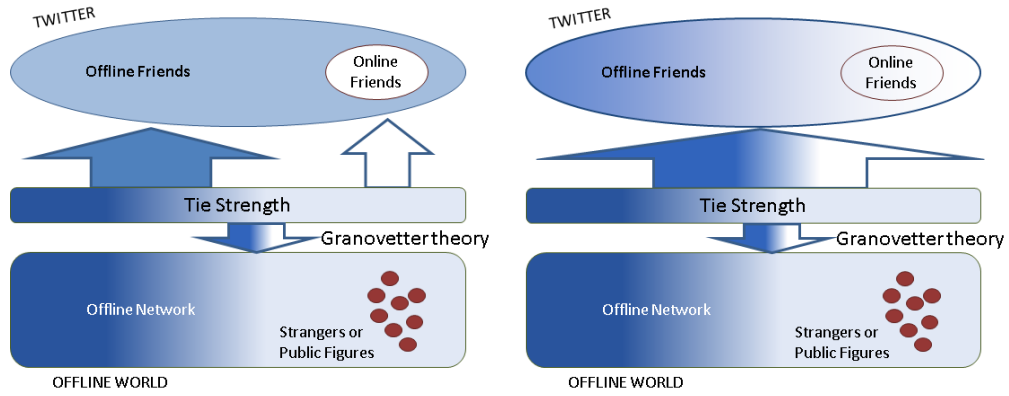
Although the study by Gilbert and Karahalios (2009) has proven that interactions online can explain closeness on Facebook (as shown in Figure 7.1(b)), we do not know whether a unit of increase of online interactions actually represents the same closeness level as a unit of increase of offline interactions.

7.3 Dataset

In 2015, we conducted a survey directed to a random sample of 41 Twitter users. The survey participants mainly come our university. We have gained IRB ruling to conduct the survey. Not all people who registered for our survey were considered in our study. We only considered those whose Twitter account fulfil the following criteria:

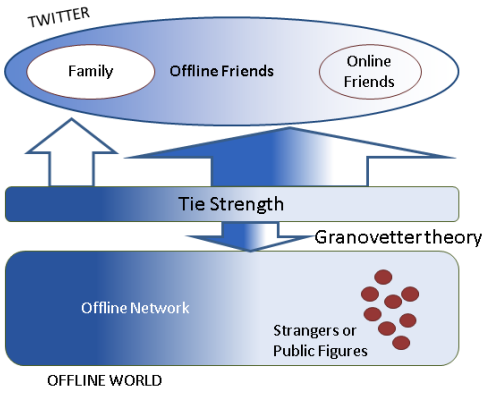
1. It has a number of followers ≤ 350 and the number of followees ≤ 400. This criteria is imposed so that we limit the size of the network so that we do not pick up social capitalists or marketing accounts, but people who use Twitter to socialize and acquire information.

2. It has a number of followers ≥ 20 and the number of followees ≥ 20. The



(a) Behaviour on Twitter reflects type of friendship (offline vs. online).

(b) Behaviour on Twitter reflects different levels of tie strength.



(c) Behaviour on Twitter reflects different levels of tie strength but not for family members.

Figure 7.1: How behaviour on Twitter reflects tie strength.

criteria is imposed so that we do not pick inactive users.

3. It posts tweets at least once a week in the past six months. In this way, we ensure that the accounts that we pick are active accounts who are not silent users.

4. Out of all the posts that they have posted in the past six months, 80% are in English. This criteria is imposed to ensure that we can analyse the content of the tweets easily.

5. It has been on Twitter for at least two years. By imposing this criteria, we ensure that the network of the Twitter users is already stable.

All the survey participants were asked to answer questions regarding their relationships with at most 100 of their followers and followees on Twitter. We limited the number of friends they need to label in order to avoid fatigue when filling up the survey. It was a way to ensuring the quality of the answers given. We obtained IRB ruling before we started our survey. We also performed trials by interviewing few of our colleagues who were asked to fill out the survey. We asked them whether the questions were clear. We modify our survey questions accordingly to make sure that our intended meaning is conveyed.

These participants were then asked to answer the following questions regarding each of their sampled Twitter followers/followees:

1. How close he or she is with the survey participant. The answer is given in five levels (1) not at all, (2) a little, (3) somewhat, (4) much, and (5) a great deal. On top of the page, we explain what each level means. Not at all means "I barely know him or her", a little means "I know him or her a little", somewhat means "I know him or her quite well", much means "we are quite close", and lastly a great deal means "we are very close, I know him inside out".

2. Whether he or she is a family, an offline friend, an offline acquaintance, purely an online friend.

3. How often he or she communicates with the participant offline. The choices are the following, (1) at least once a week, (2) at least once every two weeks, (3) at least once a month, (4) at least once every few months, (5) at least once a year, and (6) less than once a year. In our research questions, we want to compare whether different frequencies of interaction on the social information network explain tie strength in the same way that different frequencies of interaction in the offline world explain tie strength. Therefore, this question is necessary.

Our questions were designed according to the instruments we used in this study. Some of these instruments were developed on the basis of prior research. Some others were developed based on our own reasoning.

1. Closeness. Closeness has been used as a true indicator of tie strength in previous studies (Marsden and Campbell 1984; Gilbert and Karahalios 2009).

2. Number of interactions. Interactions are theoretically defined as one of the factors that build tie strength (Granovetter 1973). Moreover, unlike other factors such as amount of time spent together, emotional intimacy or intensity, they are easy to quantify and measure online.

3. Topic similarity. We add topic similarity as a comparison to how interactions online explain closeness. Since Twitter has been used as news media, content consumption and production are a type of interactions on Twitter that do not exist offline. Topic similarity measures how similar two persons are in content consumption and production.

4. Relationship types (family or online friends). We use this instrument as a control in our models because different interaction patterns may apply for these people. Online friend is a type of relationship that previously does not exist in a measurement of tie strength before the era of social media. Meanwhile family may be very close, but does not tweet to one another a lot.

Lastly, we crawled the ego network of the 41 participants. An ego network is a network that consists of a user, his followers/followees, and the edges among all of them. We also crawled the tweets of the participants and their followers/followees.

7.3.1 Handling of Missing Data

Out of the 41 participants who attend our survey, 21 are private accounts. We ask their consent to follow their accounts so that we can crawl their data. They have on average around 226 friends on Twitter (followers/followees). Fifty percent of these friends are private users, therefore we cannot crawl their networks or their tweets. Because our survey data is not much, we choose not to ignore edges where one of the user is a private user. Instead, we ignore the missing information when running our statistical model. However, we will take into account the missing information when we analyse our results.

7.4 Methodology

Before proceeding to the methods, we recap the research questions in our study. First we want to know whether different levels of tie strength (beyond offline vs. online) are explained by different behaviours on the social information network. Second, we question whether they are explained in the same way as how they are explained by different behaviours in the offline world.

7.4.1 Variables for the Constructs

Our equation is based upon the theory that defines tie strength as a combination of the amount of time, the emotional intensity, the intimacy, and reciprocal services (See Section 7.2.1). Therefore, tie strength, or what best measures it would be the dependent variable, and the behaviours in the offline and the online world would be the explanatory or independent variables.

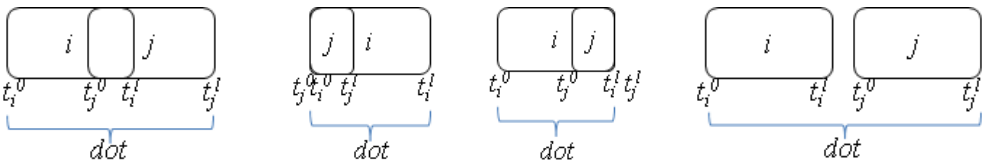


Figure 7.2: Duration of tweets (dot) of user i and user j in four scenarios.

In order to answer our research questions we need to come up with variables for all the constructs in our research questions. These variables should be quantifiable. The levels of tie strength is measured by closeness, as explained in Section 7.3. Meanwhile, we focus on several behaviours on the social information network that could explain closeness, namely interaction and topic similarity. In terms of interaction, we consider the number of @ *interaction* between a user and his follower/followee. On Twitter, @ *interaction* indicates **reply**, **mention**, and **retweet**. We use Twitter-LDA (Zhao et al. 2011) to create a topical representation of each user's tweets. A topical representation is a user's tweets distribution over 15 topics. The top words in each topic can be viewed on Appendix B.3.1. Topic similarity is achieved by calculating the cosine similarity between two topic distributions. Meanwhile the behaviour in the offline world that we use is the frequency of interaction among Twitter friends in the offline world that we have acquired from the survey answers. All the variables are summarized on Table 7.1.

The table shows that interactions between two users are divided into three: (1) the @ calls from user i to user j, (2) the @ calls from user j to user i, and (3) the reciprocated @ calls between user i and user j. What we mean by reciprocated @ calls are not necessarily tweets of the same conversation thread, for example user i mentions user j, and user j replies user i. Such reciprocated calls are rare. Instead, what we mean by reciprocated @ call is user i mentions user j, and user j may mention user i at another time about a completely different tweet. It simply indicates a reciprocated attention on another user's presence and tweets. Therefore, the formula for reciprocated @ call symbolized by $r_{\overline{ij}}$ is $\min(r_{ij}, r_{ji})$.

f_{ij} represents the number of offline interactions between user i and user j across

Table 7.1: List of Variables

Construct	Symbol	Formula	Description
Closeness	c_{ij}		The closeness level between user i and user j.
Interaction on Twitter	r_{ij}		Number of @ calls that user i makes to user j on Twitter.
	r_{ji}		Number of @ calls that user j makes to user i on Twitter.
	$r_{\overline{ij}}$	$\min(r_{ij}, r_{ji})$	Number of reciprocated @ calls between user i and user j on Twitter.
Topic similarity on Twitter.	s_{ij}		Topic similarity between user i tweets and user j tweets.
Duration of tweets	d_i	$t_i^l - t_i^j$	Duration of tweets of user i.
	t_i^l		Timestamp of the last tweet of user i.
	t_i^0		Timestamp of the first tweet of user i.
Interaction offline	f_{ij}	$f_{ij}/\text{day} \times (\min(t_i^0, t_j^0) - \max(t_i^l, t_j^l))$	Number of offline interactions between user i and user j across the duration of their tweets.
	f_{ij}/day		Number of offline interaction between user i and user j in a day.
Relationship	m_{ij}		Whether user i and user j is a family.
	n_{ij}		Whether user i and user j is purely an online friend.

the duration of their tweets. The duration of tweets of user i and user j can happen in one of the four scenarios (See Figure 7.2). The formula $(\min(t_i^0, t_j^0) - \max(t_i^l, t_j^l))$ will give the duration of tweets (dot) between user i and user j on Figure 7.2. Meanwhile, we have designed our survey questions to discover the frequency of offline interactions between a user and his friend (See Section 7.3). We transform the answers into a quantifiable measure of number of interactions per day f_{ij}/day. The transformation is such for each answer:

- At least once a week. We assume that the two users interact once a week, and therefore f_{ij}/day is $\frac{1}{7}$.
- At least once every two weeks. We assume that the two users interact once every two weeks, and therefore f_{ij}/day is $\frac{1}{14}$
- At least once a month. We assume that the two users interact once every

month, and therefore, f_{ij}/day is $\frac{1}{30}$.

- At least once every few months. We assume that the two users interact once every three months, and therefore f_{ij}/day is $\frac{1}{90}$.

- At least once a year. We assume that the two users interact once every year, and therefore f_{ij}/day is $\frac{1}{365}$.

- Less than once a year. We assume that the two users interact once every 1000 days (around three years), and therefore f_{ij}/day is $\frac{1}{1000}$.

We calculate the number of offline interactions across the duration of the tweets (dot) to normalize the number of offline interactions between a range that can be compared with the number of Twitter interactions across the same duration of tweets. By doing so, the coefficients of offline interactions and interactions on Twitter that we get from running the statistical model can be compared.

7.4.2 Statistical Model

In our study, the dependent variable c_{ij} is ordinal and not continuous. The closeness is separated into five levels where the higher the level is, the closer the two users are (See Section 7.3). In order to use the linear regression method, we need to assume that closeness is continuous. However, we do not want to do so because one of our research questions inquires whether the relationship between closeness and interactions changes when the level of closeness passes a certain level. In other words, we want to be able to analyse the relationship between closeness and interactions at each level to check out whether they are similar or different. Below are the two of the most common ways to handle dependent variable that's ordinal:

- Ordinal logistic regression (Harrell 2001). It estimates the same coefficient for each ordinal level. The assumption that the coefficients of the independent variables are the same at each ordinal level is called the proportional

odds assumption. The ordinal logistic regression should only be applied if the proportional odds assumption holds true.

- A series of binomial logistic regression (Hilbe 2011). Binomial logistic regression can be applied multiple times, one time at each ordinal level, to estimate the coefficients that best separate the data with dependent variable above or equal to an ordinal level K from the data with dependent variable below the ordinal level K.

In this study, we are going to perform the ordinal logistic regression if the proportional odds assumption holds true, otherwise we are going to perform a series of binomial logistic regression to estimate the coefficients at each ordinal level.

Proportional Odds Assumption

To test whether the proportional odds assumption holds true for the relationship between closeness and the independent variables, we can employ a graphical method. The values displayed on the graph are essentially (linear) predictions from a logit model that models the probability that y is greater than or equal to a given value, using one independent variable (x) at a time[1] (Harrell 2001). With little math, it is easy to see that this is none other than running a series of binary logistic regressions with varying cutpoints on the dependent variable and checking the equality of coefficients across cutpoints (See Equation 7.1).

$$\ln\left(\frac{p\left(c_{ij} \geq K\right)}{1 - p\left(c_{ij} \geq K\right)}\right) = \alpha + \beta x + \sigma_i \tag{7.1}$$

In the Equation 7.1 $p\left(c_{ij} \geq K\right)$ is the probability of the closeness between user i and user j is greater than or equal to K. Meanwhile, K represents each level of closeness from two to five. We will run the model for each value of K. Meanwhile, σ_i controls for the variation in the closeness scores of the survey participants (ego

[1]For its implementation using R, visit https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/

users). For example, an ego user tends to rate everyone higher in closeness, and another ego user tends to rate everyone lower. x is the dependent variable which is being tested for proportional odds assumption. It can be r_{ij}, r_{ji}, $r_{\overline{ij}}$, etc.

The shortcoming of this approach to test for proportional odds assumption is that there is no statistical significance given. We can only eyeball whether the coefficients are similar or different.

The second way to test the proportional odds assumption is to calculate a likelihood ratio to see whether the coefficients estimated by ordinal logistic regression model are statistically different from the ones estimated by the multinomial logistic regression model. The multinomial logistic regression model is typically used to model more than two unordered responses [2]. In this way, we can calculate whether the two sets of coefficients are different significantly.

After testing the proportional odds assumption, we discover that all variables (r_{ij}, r_{ji}, $r_{\overline{ij}}$, s_{ij}, f_{ij}) have coefficients of ordinal logistic regression model that are significantly different from the coefficients of multinomial logistic regression model. We conclude that the proportional odds assumption *does not hold*. Therefore, we will perform a series of binomial logistic regression to answer our research questions.

Statistical Model to Explain How Online Interactions Explain Closeness

In this study, we want to discover how online interactions explain closeness. Additionally, we want to know whether a certain type of relationship, such as family or online friend, increases or decreases interactions on Twitter necessary to explain closeness level K. To find this out, we use the following logistic model.

[2]For its implementation using R, visit http://data.library.virginia.edu/fitting-and-interpreting-a-proportional-odds-model/

$$\ln\left(\frac{p\left(c_{ij}\geq K\right)}{1-p\left(c_{ij}\geq K\right)}\right)=\alpha+\beta r_{ij}+\gamma_1 m_{ij}+\gamma_2 n_{ij} \\ +\theta_1\left(r_{ij}\times m_{ij}\right)+\theta_2\left(r_{ij}\times n_{ij}\right)+\sigma_i \tag{7.2a}$$

$$\ln\left(\frac{p\left(c_{ij}\geq K\right)}{1-p\left(c_{ij}\geq K\right)}\right)=\alpha+\beta r_{ji}+\gamma_1 m_{ij}+\gamma_2 n_{ij} \\ +\theta_1\left(r_{ji}\times m_{ij}\right)+\theta_2\left(r_{ji}\times n_{ij}\right)+\sigma_i \tag{7.2b}$$

$$\ln\left(\frac{p\left(c_{ij}\geq K\right)}{1-p\left(c_{ij}\geq K\right)}\right)=\alpha+\beta r_{\overline{ij}}+\gamma_1 m_{ij}+\gamma_2 n_{ij} \\ +\theta_1\left(r_{\overline{ij}}\times m_{ij}\right)+\theta_2\left(r_{\overline{ij}}\times n_{ij}\right)+\sigma_i \tag{7.2c}$$

The interaction variables θ_1 and θ_2 explain whether a unit of interaction means the same thing for family and online friends respectively. For example, if θ_1 is positive, a unit of interaction means a higher likelihood of family being close than friends being so.

Statistical Model to Explain How Topic Similarity Explains Closeness

Another behaviour on Twitter that we investigate besides interaction is topic similarity. We perform binomial logistic regression for each closeness level K to find out how topic similarity between two Twitter users explains their closeness.

$$\ln\left(\frac{p\left(c_{ij}\geq K\right)}{1-p\left(c_{ij}\geq K\right)}\right)=\alpha+\beta s_{ij}+\gamma_1 m_{ij}+\gamma_2 n_{ij} \\ +\theta_1\left(s_{ij}\times m_{ij}\right)+\theta_2\left(s_{ij}\times n_{ij}\right)+\sigma_i \tag{7.3}$$

Statistical Model to Explain How Offline Interactions Explain Closeness

To know whether interactions or topic similarity on Twitter explain closeness in the same way that interactions offline explain closeness, we first find out how offline interactions explain closeness. We use the following formula that also controls for the type of relationship (family, online friends, or neither). There is only one interaction variable $(f_{ij} \times m_{ij})$ because online friends do not interact offline. Interaction variable $(f_{ij} \times n_{ij})$ is always zero.

$$\ln\left(\frac{p\left(c_{ij} \geq K\right)}{1 - p\left(c_{ij} \geq K\right)}\right) = \alpha + \beta f_{ij} + \gamma_1 m_{ij} + \gamma_2 n_{ij} + \theta_1 \left(f_{ij} \times m_{ij}\right) + \sigma_i \tag{7.4}$$

Because the edges with topic similarity data are not exactly similar to the edges with online interactions data, we have to run Equation 7.4 two times. At the first time we run the equation using edges with online interactions data. The results are compared with the results from running Equation 7.2. At the second time we run the equation using edges with topic similarity data. The results compared are compared with the results from running Equation 7.3.

Statistical Model to Explain How Offline and Online Interactions Explain Closeness

Although we can compare the direction of the coefficient in Equation 7.2 and Equation 7.4, we cannot compare the magnitude of the coefficients unless they are in the same equation. Therefore, we use Equation 7.5 to find out how offline interactions compare to online interactions in magnitude when they are used together in explaining closeness. In doing so, we can also find out which one, offline or online interactions, is the best independent variable to estimate closeness. The variable x that gives a higher decrease in AIC when it is added into a logistic model employing all variables except the variable x, is the best variable in estimating closeness. In

Table 7.2: Average Interactions and Topic Similarity across Closeness Levels

Data	Variable	X				
		1	2	3	4	5
Interactions	$\overline{r_{ij}}$	6.14	10.67	14.87	25.09	47.55
	$\overline{r_{ji}}$	3.06	5.03	9.40	12.29	12.31
	$\overline{r_{ij}}$	0.62	1.56	2.63	3.55	6.18
	$\overline{f_{ij}}$	2.23	10.55	43.29	85.81	153.30
Topic Similarity	$\overline{s_{ij}}$	0.44	0.49	0.51	0.52	0.52
	$\overline{f_{ij}}$	2.05	11.56	41.98	84.97	148.20

other words, we compare Equation 7.5 with Equation 7.2 and Equation 7.4.

$$\begin{aligned} \ln\left(\frac{p\left(c_{ij}\geq K\right)}{1-p\left(c_{ij}\geq K\right)}\right) &= \alpha+\beta_1 r_{ij}+\beta_2 f_{ij}+\gamma_1 m_{ij}+\gamma_2 n_{ij} \\ &+\theta_1\left(r_{ij}\times m_{ij}\right)+\theta_2\left(r_{ij}\times n_{ij}\right)+\theta_3\left(f_{ij}\times m_{ij}\right)+\sigma_i \end{aligned} \tag{7.5}$$

7.5 Results

7.5.1 Basic Analysis

On Table 7.2, we provide the average of interactions (offline and online) and topic similarity, given each closeness level K.

The results show that interactions and topic similarity increase as closeness increases, proving that hypothesis shown in Figure 7.1(b) is true. However, the level of increase in interactions are different for the variables. For example, online interactions (r_{ij}) start at around 11 on closeness level $K = 2$, and so do offline interactions. However, offline interactions increase more than online interactions do when closeness level K reaches 3. To know how exactly a unit increase of each variable changes at each level of closeness, we will perform the logistic regressions described in Section 7.4.

7.5.2 How Online Interactions Explain Closeness

We perform logistic regression (Equation 7.2) on online interactions at each closeness level.

The results are shown on Table 7.3, 7.4, and 7.5. Positive βs indicate that online interactions increase as closeness level increases. Decreasing βs indicate that online interactions increase more from one level to another, explaining the average online interactions that we see on Table 7.2.

The results also show that β values when the level of closeness is 2 and when the level of closeness is 3 and above, are different. Section 7.3 defined level 2 as a level at which one knows someone a little, whereas level 3 as a level at which one knows someone quite well. If we call those who are at level 1 and 2 as acquaintances, and those who are at level 3 and above as friends, we may conclude that the frequency of interactions on Twitter increases *abruptly* from an acquaintance to a friend. However, it increases *steadily* from a friend to a very close friend (coefficients at level 3 to 5 are similar).

γ_1 is significant and positive for friends (those who score three to five). It indicates that family members do not interact as much on Twitter, even though they are quite close. Additionally θ_1 is significant and positive for very close friends who score five. It indicates that family members interact even less on Twitter when they are very close.

The results also reveal that r_{ij} is the best independent variable among the three types of online interactions we consider. The logistic model employing r_{ij} has the minimum AIC. One plausible reason is because missing r_{ij} values are fewer than missing r_{ji} values. Although we can crawl all tweets of user is (survey correspondents) because we get the permission to follow them on Twitter, we cannot crawl the tweets of user js who are private users. Moving forward, the analyses will only use r_{ij}.

Table 7.3: Logistic Regression on r_{ij} Controlling for Relationship Types

	K = 2		K = 3		K = 4		K = 5	
		Sig.		Sig.		Sig.		Sig.
α	-1.81	***	-1.96	***	-2.95	***	-4.08	***
$\beta(r_{ij})$	0.05	***	0.03	***	0.03	***	0.03	***
$\gamma_1(m_{ij})$	1.64		2.15	*	3.07	***	3.58	***
$\gamma_2(n_{ij})$	-4.45	***	-3.37	***	-3.54	***	-2.36	*
$\theta_1(r_{ij} \times m_{ij})$	12.82		1.26		0.05		0.13	*
$\theta_2(r_{ij} \times n_{ij})$	-0.03		0.00		0.02		0.02	
AIC	1940		2102		1696		992	
Significance level: $^{***}p \leq 0.001$, $^{**}p \leq 0.01$, $^{*}p \leq 0.05$, $.\ p \leq 0.1$								

Table 7.4: Logistic Regression on r_{ji} Controlling for Relationship Types

	K = 2		K = 3		K = 4		K = 5	
		Sig.		Sig.		Sig.		Sig.
α	-1.10	***	-1.42	***	-2.35	***	-3.37	***
$\beta(r_{ji})$	0.02	***	0.01	***	0.01	***	0.01	***
$\gamma_1(m_{ij})$	3.16	**	3.33	***	3.30	***	4.44	***
$\gamma_2(n_{ij})$	-4.87	***	-3.66	***	-3.32	***	-2.44	**
$\theta_1(r_{ij} \times m_{ij})$	0.09		0.14		0.03		-0.01	
$\theta_2(r_{ij} \times n_{ij})$	0.02		0.02		-0.01		-0.01	
AIC	2106		2289		1880		1160	
Significance level: $^{***}p \leq 0.001$, $^{**}p \leq 0.01$, $^{*}p \leq 0.05$, $.\ p \leq 0.1$								

Table 7.5: Logistic Regression on $r_{\overline{ij}}$ Controlling for Relationship Types

	K = 2		K = 3		K = 4		K = 5	
		Sig.		Sig.		Sig.		Sig.
α	-1.21	***	-1.46	***	-2.39	***	-3.42	***
$\beta(r_{\overline{ij}})$	0.07	***	0.04	***	0.03	***	0.02	***
$\gamma_1(m_{ij})$	3.11	**	3.24	***	3.33	***	4.28	***
$\gamma_2(n_{ij})$	-5.03	***	-3.71	***	-3.24	***	-2.31	**
$\theta_1(r_{ij} \times m_{ij})$	10.51		12.43		0.06		0.05	
$\theta_2(r_{ij} \times n_{ij})$	0.30	*	0.24	*	-9.65		-9.86	
AIC	2082		2284		1880		1154	
Significance level: $^{***}p \leq 0.001$, $^{**}p \leq 0.01$, $^{*}p \leq 0.05$, $.\ p \leq 0.1$								

Table 7.6: Logistic Regression on s_{ij} Controlling for Relationship Types

	K = 2		K = 3		K = 4		K = 5	
		Sig.		Sig.		Sig.		Sig.
$\beta_1(s_{ij})$	0.67	***	0.89	***	1.07	***	0.92	**
$\gamma_1(m_{ij})$	6.95	*	3.75	***	3.71	***	3.59	***
$\gamma_2(n_{ij})$	-5.31	***	-3.16	***	-2.85	***	-2.01	**
$\theta_1(s_{ij} \times m_{ij]})$	-4.59		-0.74		-0.63		1.12	
$\theta_2(s_{ij} \times n_{ij]})$	0.47		-1.93	.	-2.26		-3.47	
AIC	5299		4678		3313		1912	
Significance level: $^{***}p \leq 0.001$, $^{**}p \leq 0.01$, $^{*}p \leq 0.05$, . $p \leq 0.1$								

7.5.3 How Topic Similarity of Tweets Explains Closeness

Besides @ interactions, we also analyse how topic similarity as another behaviour on Twitter, explains closeness. Topic similarity describes the similarity of tweets that two users post on Twitter. We perform a series of binomial logistic regression controlling for relationship type (Equation 7.3). The results on Table 7.6 displays positive coefficients. They indicate that topic similarity increases as closeness level increases. However, increasing coefficients also indicate that unlike online interactions, topic similarity increases less from one level to another.

7.5.4 How Offline Interactions Explain Closeness

We run Equation 7.4. The results are shown on Table 7.7 and Table 7.8. Positive βs show that an increase in offline interactison at each level of closeness. However, the increase that is necessary to distinguish one type of closeness level from another differs.

Similar to the results in estimating online interactions, there is a sudden decrease in β when K increases from two to three. The decrease is higher for offline interactions than online interactions indicating that the increased offline interactions from closeness level two and below to closeness level three and above, is much higher than the increased online interactions. If we call those who are at level one and two as acquaintances and those who are at level three and above as friends, we may conclude that the increase of offline interactions from an acquaintance to a friend is

Table 7.7: Logistic Regression on f_{ij} Controlling for Relationship Types (for Interactions Data)

	X = 2		X = 3		X = 4		X = 5	
		Sig.		Sig.		Sig.		Sig.
α	-3.19	***	-2.31	***	-2.88	***	-4.09	***
$\beta(f_{ij})$	0.18	***	0.06	***	0.03	***	0.02	***
$\gamma_1(m_{ij})$	-3.64		3.64	**	2.77	***	1.87	.
$\gamma_2(n_{ij})$	-3.51	***	-2.59	***	-2.55	***	-1.22	
$\theta_1(f_{ij} \times m_{ij})$	1.42		0.00		0.00		0.05	*
AIC	1575		1615		2369		801	
Significance level: $^{***}p \leq 0.001$, $^{**}p \leq 0.01$, $^{*}p \leq 0.05$, . $p \leq 0.1$								

Table 7.8: Logistic Regression on f_{ij} Controlling for Relationship Types (for Topic Similarity Data)

	X = 2		X = 3		X = 4		X = 5	
		Sig.		Sig.		Sig.		Sig.
$\beta(f_{ij})$	0.07	***	0.04	***	0.02	***	0.02	***
$\gamma_1(m_{ij})$	-0.84		3.60	***	3.32	***	3.38	***
$\gamma_2(n_{ij})$	-4.60	***	-3.26	***	-3.00	***	-2.17	***
$\theta_1(s_{ij} \times m_{ij]})$	0.91		-0.01	**	-0.01	*	0.00	
AIC	4534		3620		1340		1328	
Significance level: $^{***}p \leq 0.001$, $^{**}p \leq 0.01$, $^{*}p \leq 0.05$, . $p \leq 0.1$								

steeper than the increase of online interactions.

There is also a considerable jump in β when K increases from three to four indicating that unlike online interactions that increase equally from friends who score three to friends who score five, offline interactions increase more in the same setting. Therefore, offline interactions are better than online interactions to separate friends at different closeness levels.

Meanwhile, positive γ_1 shows that a family member does not need to interact offline as much as a close friend does even when they are considered equally close.

7.5.5 How Offline Interactions and Online Interactions Explain Closeness

When both variables are used to estimate closeness, we discover that the βs of online interactions across closeness levels do not decrease as much as when online

Table 7.9: Logistic Regression on r_{ij} and f_{ij} Controlling for Relationship Types

	X = 2		X = 3		X = 4		X = 5	
		Sig.		Sig.		Sig.		Sig.
α	-3.72	***	-2.73	***	-3.30	***	-4.49	***
$\beta_1(r_{ij})$	0.03	***	0.03	***	0.02	***	0.02	***
$\beta_2(f_{ij})$	0.17	***	0.06	***	0.02	***	0.02	***
$\gamma_1(m_{ij})$	-3.59		3.01	*	2.94	***	1.96	.
$\gamma_2(n_{ij})$	-3.50	***	-2.67	***	-2.99	**	-1.45	
$\theta_1(r_{ij} \times m_{ij})$	11.76		0.66		-0.04		-0.05	
$\theta_2(r_{ij} \times n_{ij})$	0.00		0.01		0.04		0.04	
$\theta_3(f_{ij} \times m_{ij})$	1.34		0.00		0.01		0.06	*
AIC	1523		1550		1270		735	
Significance level: $^{***}p \leq 0.001$, $^{**}p \leq 0.01$, $^{*}p \leq 0.05$, . $p \leq 0.1$								

interactions are used alone. Meanwhile the βs of offline interactions across closeness levels still decrease.

The results show that the increase of online interactions across closeness levels is more stable. Meanwhile the increase of offline interactions across closeness levels is more uneven. These can be reflected on Figure 7.1(c), by having a more uneven colour gradation of the rectangle at the bottom (offline world) than colour gradation of the circle at the top (Twitter). The change in colour of the rectangle (number of interactions) should be most abrupt on the far left (earlier closeness levels).

Lastly, the decrease in AIC is the highest when the results are compared to the results in estimating closeness by online interactions alone. It implies that offline interactions decreases most AIC and therefore, is the better variable than online interactions to estimate closeness.

7.5.6 Offline Interactions vs. Topic Similarity in Explaining Closeness

We regress both topic similarity and offline interactions together on Table 7.10. By comparing the AIC of the regression results and the ones on Table 7.6 and Table 7.8, we find that offline interactions are the better variable to estimate closeness.

Table 7.10: Logistic Regression on s_{ij} and f_{ij} Controlling for Relationship Types

	X = 2		X = 3		X = 4		X = 5	
		Sig.		Sig.		Sig.		Sig.
α	-2.43	***	-2.74	***	-3.92	***	-5.41	***
$\beta_1(s_{ij})$	0.52	*	1.02	***	1.24	***	0.97	*
$\beta_1(f_{ij})$	0.07	***	0.04	***	0.03	***	0.02	***
$\gamma_1(m_{ij})$	-0.07		4.50	***	4.02	***	3.40	***
$\gamma_2(n_{ij})$	-4.77	***	-2.41	***	-2.01	**	-1.00	
$\theta_1(s_{ij} \times m_{ij})$	-4.68		-1.92		-1.50		0.15	
$\theta_2(s_{ij} \times n_{ij})$	0.60		-2.10	*	-2.47		-3.58	
$\theta_3(f_{ij} \times m_{ij})$	1.35		-0.01	**	-0.01	**	0.00	
AIC	4529		3607		2360		1327	
Significance level: $^{***}p \leq 0.001$, $^{**}p \leq 0.01$, $^{*}p \leq 0.05$, . $p \leq 0.1$								

7.6 Conclusion

Previous studies have shown that behaviour on Twitter reflects type of friendship (offline vs. online). In this study we want to know whether behaviour on Twitter reflects different levels of tie strength beyond the realm offline vs. online and whether it does so in a manner similar to how offline interactions reflect tie strength. Using @ interactions and topic similarity as the operational measures for behaviour on Twitter, we discover that behaviour on Twitter reflects different levels of tie strength but not in the same manner as how offline interactions do so. The increase of offline interactions is more abrupt as tie strength increases, especially when closeness level moves from two to three (from acquaintances to friends). Meanwhile, the increase of online interactions is more stable across closeness levels. Additionally, a family members do not need interact as much both on Twitter or offline even when they are considered very close.

Meanwhile, online interactions are more similar to offline interactions than topic similarity in reflecting tie strength. Both offline and online interactions increase more as closeness level increases. Meanwhile, topic similarity increases less as closeness level increases.

Lastly, offline interactions are better than Twitter interactions or topic similarity in explaining tie strength.

Our studies can still be improved by two ways. First, we can devise new measures to quantify other factors besides frequency of interactions that has been theoretically supported for explaining closeness. Second, we can improve the survey questions by devising a quantifiable measure to ensure the quality of survey answers.

Chapter 8

Summary and Discussions

In this thesis, we aim to explore the behaviour of offline friends on a social information network, using Twitter as a case study. Previous research works have mainly compared the behaviour offline vs. online, but rarely do they become a bridge between the offline and the online world. For a study to become a bridge between the offline and the online world, it has to understand how offline friends behave in the online world. Several studies have been conducted on this topic, but because of the lack of ground-truth data they mainly focus on the event based social networks.

We continue on their effort but focusing on a social information network, specifically Twitter. We rely most of our experiments on a dataset of 98 Twitter ego networks of which we have obtained the information on types of friends (offline or not) of the ego user.

8.1 Essay 1A and 1B

The first essay explores how offline friends form follow-network on Twitter. The three principles proposed by Schaefer, namely reciprocity, preferential attachment and triadic closure, are found to be crucial in predicting offline friends on Twitter.

Going beyond triads, in the second essay we want to discover other social structures that exist among offline friends on Twitter. Using configuration model, we

create a corresponding random network for each Twitter network that we have. Harnessing the Louvain algorithm, we perform iterative Louvain to find out whether these structures significantly exist on Twitter network. We discover that complete clique structures exist significantly more among offline friends. Meanwhile, chain and star-like structures exist significantly more among online friends.

8.1.1 How We Should Interpret Online Social Media Data

When online social media comes about, social science researchers rejoice because they now have abundant data for analysis. Social networks data was usually scarce, and its collection had to rely on humans imperfect memory. However, can we truly rely on on online networks data for our analysis?

Dunbar et al. (2015) showed that we can, for people who mutually reply one another. However, reply data is usually scarce. Meanwhile following data is usually abundant. Is following data a reliable substitute for offline social network data? Our study shows that the use of online social follow-networks to substitute for offline social network data should be performed with caution. Networks that can replace the offline friendship network should have high density, and high reciprocity. Meanwhile, office or interest group network in the offline world may better mimic online information network whose structure is usually star-like and more hierarchical.

8.2 Essay 2A and 2B

Our third essay investigates the effect of separating tie choices on epidemiological modelling of tweets. The purpose on doing so is to discover how strong ties benefit information diffusion in the system dynamics model. In our work, strong ties are represented by reciprocated tie. Meanwhile, the strong ties are also the representative of offline ties. In other words, by examining the role of strong ties in information diffusion, we hope to gain insights into the role of offline friends in information diffusion. Taking reciprocated ties to substitute for offline friends is not

too far a stretch, and that's the only way to get the closest to offline friendships in a long chain of information diffusion. Our analyses show that personal tweets are more likely to be relayed by reciprocated ties. Meanwhile, entertainment tweets are more likely to be relayed by unreciprocated ties.

In the fourth essay, we also investigate retweet chain of the ego networks for which we have the ground-truth data of who the ego user's offline friends are. Our analyses show that personal tweets are more likely to be relayed by offline ties and ties that are likely to be offline. Meanwhile, entertainment tweets are more likely to be relayed by online friends and ties that are likely to be online (ties between a user's offline friend and a user's online friend, and ties between a user's online friends). However, unlike in a full retweet chain, in a retweet chain of an ego network, reciprocated ties are more likely to diffuse tweets regardless of topics. Additionally offline ties are more likely to retweet old news.

8.2.1 The implication for future research on information diffusion

Besides an external influence such as the offline information diffusion, Bakshy et al. (2012) have shown that the role of online social networks in information diffusion is still extremely important. Meanwhile, we discover that the role of one of the sources of external influence (offline friends) on information diffusion on the online social networks is quite important. The results that come out of these studies have highlighted the interdependency between the offline world and the online world, and how important it is to consider these external influences when thinking about retweet.

On another spectrum, information spreaders can also be dissected by their offline and online relations. For example, Lotan et al. (2011) classifies types of information spreaders into several categories: mainstream media organizations, mainstream new media organizations, non-media organizations, journalists, bloggers,

activists, digerati, political actors, celebrities, researchers, bots and others. Some of these categories, such as journalists, bloggers, and activists, are individuals who may have more acquaintances outside of the internet on their online networks. Meanwhile, other categories such as mainstream media media organizations, political actors and celebrities are an organization or an extremely popular individual who have relatively fewer offline friends. Lotan et al. (2011) found that news on Twitter is co-constructed mainly by bloggers, activists, alongside journalists. Do they play such an important role because they have a higher presence of offline connections online?

As online social media use for news diffusion is increasing, the interdependency between these two worlds, offline and online, are going to increase, and the studies of information diffusion that consider these two worlds in analysis – either by removing the influence that one world has on another, or by studying the influence that one world has on another world – will continue to be a popular topic.

8.3 Essay 3

In our last essay, we explore how tie strength is reflected on the behaviour in offline world and on Twitter. We discover that increased tie strength, means increased interactions both offline and online, although it does not apply to family members. However, the way interactions increase in the offline world and on Twitter differs. The jump of increased interactions in the offline world is larger than the jump of increased interactions on Twitter, as tie strength increases. On the other hand, topic similarity reflects tie strength differently. It increases less as tie strength increases.

Overall we have provided readers with a comprehensive analysis on offline friends behaviour on Twitter including network formation, and information diffusion. Additionally, we also explore how tie strength is reflected offline and on Twitter. Although many people acknowledge the importance of our research question, submission often finds a stumbling block that comes in the form of misgivings on

the dataset. Our dataset is often deemed to be too small. We have obtained a new dataset of additional active Twitterers, but we discover that most active Twitter users in Singapore have private accounts that disable us to crawl their network data. In a sense, the additional dataset is also incomplete. Therefore, throughout my PhD candidacy, I keep improving on the method to circumvent the issue of validity given a small dataset. Future works can keep improving on the method to strengthen validity if the same dataset is to be used. Otherwise, if funding is not an issue, additional survey can include active Twitterers out of Singapore who usually have public accounts.

8.4 Are Bots a Concern?

The fact that bots account for possibly 62% of all messages in Twitter may influence our results. These bots have been mostly used for malicious tasks such as spreading false information (Morstatter et al. 2016). However, we argue that bots are not a concern in most of our studies.

In most of our essays, we use the dataset from real users who usually do not follow malicious bots (Morstatter et al. 2016). Therefore, we can expect that the results of Essay 1A, 1B, 2B, and 3 are barely affected by bot.

Meanwhile, the 30 tweets that we analyse in Essay 2A, are not a big dataset. We have scanned these tweets and discover that their sources are not likely to be bots, except for the topic rewards. Tweets on rewards promise a chance for receiving gifts.

Although the information sources are not likely to be bots, these information may be promoted by bots in the middle of the way. These bots are likely to be weak tie because they usually are not followed by who they follow (Morstatter et al. 2016). If these tweets are spread by bots who promote individuals and topics the differences in diffusion parameters across topics may be exaggerated. Moreover, the role of weak ties may be exaggerated.

Bibliography

Abbas, S. M. A., Alam, S. J., and Edmonds, B. (2013). Towards validating social network simulations. In *Advances in Social Simulation: Proceedings of the 9th of the European Social Simulation Association (ESSA'13)*, pp. 1–12. Springer-Verlag, Heidelberg.

Adler, P. A., Kless, S. J., and Adler, P. (1992). Socialization to gender roles: Popularity among elementary school boys and girls. *Sociology of Education 65*(3), 169–187.

Alba, R. D. (1973). A graph-theoretic definition of a sociometric clique. *Journal of Mathematical Sociology 3*, 113–126.

Antheunis, M. L., Valkenburg, P. M., and Peter, J. (2012). The quality of online, offline, and mixed-mode friendships among users of a social networking site. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace 6*(3).

Arnaboldi, V., Conti, M., Massimiliano, L. G., Passarella, A., and Pezzoni, F. (2014). Information diffusion in OSNs: The impact of nodes' sociality. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing (SAC'14)*, pp. 616–621. ACM Press, New York.

Backstrom, L. and Kleinberg, J. (2014). Romantic partnerships and the dispersion of social ties: A network analysis of relationship status on facebook. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW'14)*, pp. 831–841. ACM Press, New York.

Bagwell, C. L., Coie, J. D., Terry, R. A., and Lochman, J. E. (2000). Peer clique participation and social status in preadolescence. *Merrill-Palmer Quarterly 46*(2), 280–305.

Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web (WWW'12)*, pp. 519528. ACM.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science 286*(5439), 509–512.

Bender, E. A. and Canfield, R. (1978). The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory 24*(3), 296–307.

Bernard, H. R., Killworth, P. D., and Sailer, L. (1979-1980). Informant accuracy in social network data IV: A comparison of clique-level structure in behavioral and cognitive network data. *Social Networks 2*, 191–218.

Best, S. J. and Krueger, B. S. (2006). Online interactions and social capital: Distinguishing between new and existing ties. *Social Science Computer Review 24*(4), 395–410.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment 10*(P10008).

Boase, J. and Wellman, B. (2006). Personal relationships: On and off the internet. In A. L. Vangelisti and D. Perlman (Eds.), *The Cambridge Handbook of Personal Relationships*, pp. 709–723. Cambridge University Press, Cambridge.

Boyd, D., Golder, S., and Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *The 43th Hawaii International Conference on System Sciences (HICSS'10)*, pp. 1–10. IEEE Computer Society Press, Washington, DC.

Boyd, D. M. and Ellison, N. B. (2008). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication 13*, 210–230.

Chan, D. K. and Cheng, G. H. (2004). A comparison of offline and online friendship qualities at different stages of relationship development. *Journal of Social and Personal Relationships 21*(3), 305–320.

Cohen, Y. A. (1969). Social boundary systems. *Current Anthropology 10*(1), 103–126.

Coie, J. and Dodge, K. (1983). Continuities and changes in children's social status: A five-year longitudinal study. *Merrill-Palmer Quarterly 29*(3), 261–282.

Contractor, N. S., Wasserman, S., and Faust, K. (2006). Testing multitheoretical, multilevel hypotheses about organizational networks: A

n analytic framework and empirical example. *Academy of Management Review 31*(3), 681–703.

Daley, D. J. and Kendall, D. G. (1964). Epidemics and rumours. *Nature 204*(1118).

Dunbar, R. I., Arnaboldi, V., Contib, M., and Passarella, A. (2015). The structure of online social networks mirrors those in the offline world. *Social Networks 43*, 39–47.

Dunbar, R. I. and Spoors, M. (1995). Social networks, support cliques, and kinship. *Human Nature 6*(3), 273–290.

Ellison, N. B., Steinfield, C., and Lampe, C. (2007). The benefits of facebook "friends": Social capital and college students' use of online social network sites. *Journal of Computer-Mediated Communication 12*(4), 1143–1168.

Enli, G. S. and Skogerbø, E. (2013). Personalized campaigns in party-centred politics. *Information, Communication & Society 16*(5), 757–774.

Frank, O. and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association 81*(395), 832–842.

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

Friedkin, N. E. (1990). A guttman scale for the strength of an interpersonal tie. *Social Networks 12*(3), 239–252.

Ghosh, S., Viswanath, B., Kooti, F., Sharma, N. K., Korlam, G., Benevenuto, F., Ganguly, N., and Gummadi, K. P. (2012). Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st International Conference on World Wide Web (WWW'12)*, pp. 61–70. ACM Press, New York.

Gilbert, E. and Karahalios, K. (2009). Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'09)*, pp. 211–220. ACM Press, New York.

Gleason, B. (2013). #Occupy wall street: Exploring informal learning about a social movement on twitter. *American Behavioral Scientist 57*(7), 966–982.

Golder, S. A. and Yardi, S. (2010). Structural predictors of tie formation in twitter: Transitivity and mutuality. In *IEEE Second International Conference on Social Computing (SocialCom'10)*, pp. 88–95. IEEE Computer Society Press, Washington, DC.

Gonçalves, B., Perra, N., and Vespignani, A. (2011). Modeling users' activity on twitter networks: Validation of dunbar's number. *PLoS One 6*(8), 1–5.

Gottfried, J. and Shearer, E. (2016). News use across social media platforms 2016. In *Journalism and Media*. Pew Research Center, Washington, DC.

Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review 25*(2), 161–178.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology 78*(6), 1360–1380.

Guille, A., Hacid, H., Favre, C., and Zighed, D. A. (2013). Information diffusion in online social networks: A survey. *SIGMOD Record 42*(2), 17–28.

Harrell, F. (2001). Ordinal logistic regression. In *Regression Modeling Strategies*, pp. 331–343. Springer-Verlag, New York.

Hilbe, J. M. (2011). Logistic regression. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science*, pp. 755–758. Springer-Verlag, Heidelberg.

Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (KDD'07)*, pp. 59f–65. ACM Press, New York.

Jin, F., Dougherty, E., Saraf, P., Cao, Y., and Ramakrishnan, N. (2013). Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis (SNAKDD'13)*. ACM Press, New York.

Kietzmann, J. H., Hermkens, K., McCarthy, I. P., and Silvestre, B. S. (2011). Social media? Get serious! Understanding the functional building blocks of social media. *Business Horizons 54*(3), 241–251.

Killworth, P. and Bernard, H. (1976). Informant accuracy in social network data. *Human Organization 35*(3), 269–286.

Kim, Y., Natali, F., Zhu, F., and Lim, E.-P. (2016). Investigating the influence of offline friendship on twitter networking behaviors. In *The 49th Hawaii International Conference on System Sciences (HICSS'16)*, pp. 736–745. IEEE Computer Society Press, Washington, DC.

Krackhardt, D. (1992). The strength of strong ties: the importance of philos in organizations. In N. Nohria and R. G. Eccles (Eds.), *Networks and organizations: Structure, Form, and Action*, pp. 216–239. Harvard Business School Press, Boston.

Kupavskii, A., Ostroumova, L., Umnov, A., Usachev, S., Serdyukov, P., Gusev, G., and Kustarev, A. (2012). Prediction of retweet cascade size over time. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pp. 2335–2338. ACM Press, New York.

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

Laursen, B. and Hartup, W. W. (2002). The origins of reciprocity and social exchange in friendships. *New Directions for Child and Adolescent Development* (95), 27–40.

Lee, C. S. and Ma, L. (2012). News sharing in social media: The effect of gratifications and prior experience. *Computers in Human Behavior 28*, 331–339.

Leskovec, J., Backstrom, L., Kumar, R., and Tomkins, A. (2008). Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'08)*, pp. 462–470. ACM Press, New York.

Li, R., Wang, C., and Chang, K. C.-C. (2014). User profiling in an ego network: Co-profiling attributes and relationships. In *Proceedings of the 23rd International Conference on World Wide Web (WWW'14)*, pp. 819–830. ACM Press, New York.

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., and Boyd, D. (2011). The revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International Journal of Communication 5*, 1375–1405.

Luce, R. D. and Perry, A. D. (1949). A method of matrix analysis of group structure. *Psychometrika 14*(2), 95–116.

MacRae, D. (1960). Direct factor analysis of sociometric data. *Sociometry 23*(4), 360–371.

Macskassy, S. A. and Michelson, M. (2011). Why do people retweet? Anti-homophily wins the day! In *Proceedings of the Fifth Annual Conference on Weblogs and Social Media (ICWSM'11)*, pp. 209–216. ACM Press, New York.

Marsden, P. V. and Campbell, K. E. (1984). Measuring tie strength. *Social Forces 63*(2), 482–501.

Meng, D., Wan, L., and Zhang, L. (2014). A study of rumor spreading with epidemic model based on network topology. In *The 18th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'14)*, pp. 381–392. Springer International Publishing, Switzerland.

Menon, T. and Phillips, K. W. (2011). Getting even or being at odds? Cohesion in even- and odd-sized small groups. *Organization Science 22*(3), 738–753.

Mesch, G. and Talmud, I. (2006). The quality of online and offline relationships: The role of multiplexity and duration of social relationships. *The Information Society 22*(3), 137–148.

Michael, J. H. (1997). Labor dispute reconciliation in a forest products manufacturing facility. *Forest Products Journal 47*, 41–45.

Moreno, J. (1960). *The Sociometry Reader*. New York, NY, USA: Free Press.

Morstatter, F., Wu, L., Nazer, T. H., Carley, K. M., and Liu, H. (2016). A new approach to bot detection. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'16)*, pp. 203–207. ACM Press, New York.

Natali, F., Carley, K. M., Zhu, F., and Huang, B. (2017). The role of different tie strength in disseminating different topics on a microblog. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'17)*, pp. 203–207. ACM Press, New York.

Natali, F. and Zhu, F. (2016). A comparison of fundamental network formation principles between offline and online friends on twitter. In *Proceedings of the 12th International Conference and School on Network Science (NetSci-X '16)*, pp. 169–177. Springer-Verlag, New York.

Newman, M. E. (2003). Fast algorithm for detecting community structure in networks. *Physical Review E 69*(6), 066133.

Pearson, M. and Michell, L. (2000). Smoke rings: social network analysis of friendship groups, smoking and drug-taking. *Drugs: Education, Prevention and Policy 7*(1), 21–37.

Peng, H. K., Zhu, J., Piao, D., Yan, R., and Zhang, Y. (2011). Retweet modeling using conditional random fields. In *IEEE 11th International Conference on Data Mining Workshops (ICDMW'11)*, pp. 336–343. IEEE Computer Society, Washington, DC.

Petróczi, A., Nepusz, T., and Fülöp, B. (2007). Measuring tie-strength in virtual social networks. *Connections 27*(2), 39–52.

Price, D. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science 27*(5), 292–306.

Rogers, E. M. and Kincaid, D. L. (1981). *Communication Networks: Toward a New Paradigm for Research*. New York, NY, USA: Free Press.

Schaefer, D. R., Light, J. M., Fabes, R. A., Hanish, L. D., and Martin, C. L. (2010). Fundamental principles of network formation among preschool children. *Social Networks 32*, 61–71.

Shi, G., Shi, Y.-z., Chan, A. K., and Wang, Y. (2009). Relationship strength in service industries. *International Journal of Market Research 51*(5), 659–685.

Shi, Z., Rui, H., and Whinston, A. B. (2014). Content sharing in a social broadcasting environment: evidence from twitter. *MIS Quarterly 38*(1), 123–142.

Simmel, G. (1950). *The Sociology of Georg Simmel*. Glencoe, Illionis, USA: Free Press.

Snijders, T. A. B. (2001). The statistical evaluation of social network dynamics. *Sociological Methodology 31*(1), 361–395.

Snijders, T. A. B. (2011). Statistical models for social networks. *Annual Review of Sociology 37*, 131–153.

Straits, B. C. (1991). Bringing strong ties back in interpersonal gateways to political information and influence. *Oxford Journals 55*(3), 432–448.

Suh, B., Hong, L., Pirolli, P., and Chi, E. H. (2010). Want to be retweeted? Large scale analytics on factors impacting retweet in twitter network. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, pp. 177–184. IEEE Computer Society Press, Washington, DC.

Svensson, Å. (1993). On the duration of a maki-thompson epidemic. *Mathematical Biosciences 117*(1-2), 211–220.

Tabourier, L., Roth, C., and Cointet, J.-P. (2011). Generating constrained random graphs using multiple edge switches. *ACM Journal of Experimental Algorithmics 16*(1), 1.7:1–1.7:15.

Tang, W., Zhuang, H., and Tang, J. (2011). Learning to infer social ties in large networks. In *Proceedings of the 2011 European conference on Machine learning and knowledge discovery in databases (ECML PKDD'11)*, pp. 381–397. Springer-Verlag, Heidelberg.

Thomases, H. (2009). *Twitter Marketing: An Hour a Day*. New Jersey, NJ, USA: John Wiley & Sons.

Thurman, B. (1979-1980). In the office: Networks and coalitions. *Social Networks 2*(1), 47–63.

Vitak, J. (2012). The impact of context collapse and privacy on social network site disclosures. *Journal of Broadcasting & Electronic Media 56*(4), 451–470.

Vitak, J., Lampe, C., Gray, R., and Ellison, N. B. (2012). ẅhy won't you be my facebook friend?̈: Strategies for managing context collapse in the workplace. In *Proceedings of the 2012 iConference (iConference'12)*, pp. 555–557. ACM Press, New York.

Weng, J., Lim, E.-P., and He, Q. (2010). Twitterrank: Finding topic-sensitive influential twitterers. In *Third ACM International Conference on Web Search and Data Mining (WSDM'10)*, pp. 261–270. ACM Press, New York.

Xie, W., Li, C., Zhu, F., Lim, E.-P., and Gong, X. (2012). When a friend in twitter is a friend in life. In *The 4th ACM Web Science Conference (WebSci'12)*, pp. 344–347. ACM Press, New York.

Xiong, F., Liu, Y., Zhang, Z.-J., Zhu, J., and Zhang, Y. (2012). An information diffusion model based on retweeting mechanism for online social media. *Physics Letters A 376*(30-31), 2103–2108.

Yaveroğlu, Ö. N., Fitzhugh, S. M., Kurant, M., Markopoulou, A., Butts, C. T., and Nataša, P. (2015). Ergm.graphlets: A package for erg modeling based on graphlet statistics. *Journal of Statistical Software 65*(12), 1–29.

Yin, P., He, Q., Liu, X., and Lee, W.-C. (2014). It takes two to tango: Exploring social tie development with both online and offline interactions. In *"Proceedings of the 2014 SIAM International Conference on Data Mining (SIAM'14)*, pp. 334–342. ACM Press, New York.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research 33*, 452–473.

Zhao, J., Wu, J., Feng, X., Xiong, H., and Xu, K. (2012). Information propagation in online social networks: a tie-strength perspective. *Knowledge and Information Systems 32*(3), 589–608.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). Comparing twitter and traditional media using topic models. In *The 33rd European Conference on Information Retrieval (ECIR'11)*, pp. 338–349. Springer-Verlag, Heidelberg.

Zuo, X., Chin, A., Fan, X., Xu, B., Hong, D., Wang, Y., and Wang, X. (2012). Connecting people at a conference: A study of influence between offline and online using a mobile social application. In *IEEE International Conference on Green Computing and Communications (GreenCom'12)*, pp. 277284. IEEE Computer Society Press, Washington, DC.

Appendix A

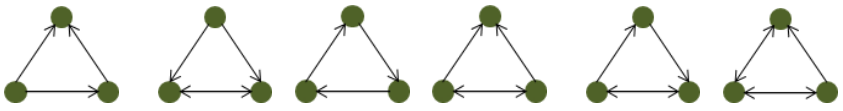
Network Formation among Offline Friends on the Social Information Network

A.1 Predicting Offline Friends on Twitter Using the Principles of Social Network Formation in the Offline World

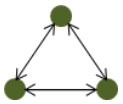
A.1.1 Triads Considered For Equation 3.1

(a) 021U (b) 012D (c) 012C (d) 111D (e) 111U (f) 201

Figure A.1: Open triads.



(a) 030T (b) 120D (c) 030C (d) 120U (e) 120C (f) 210



(g) 300

Figure A.2: Closed triads.

A.2 Going Beyond Triads: Discovering Social Cliques among Offline Friends on the Twitter Follow Network

A.2.1 Social Cliques in the Offline Network

Table A.1: Social Cliques in the Offline Networks

Code	$\|E\|$	$\|N\|$	freq.	networks
SC0	26	14	1	alqaeda
SC1	26	12	1	alqaeda
SC2	6	4	3	alqaeda,wood,bali
SC3	10	9	1	alqaeda
SC4	27	9	1	alqaeda
SC5	20	9	1	alqaeda
SC6	12	7	1	alqaeda
SC7	15	6	1	alqaeda
SC8	36	9	1	alqaeda
SC9	21	7	1	alqaeda
SC10	41	12	1	bali
SC11	31	11	1	bali
SC12	43	12	1	research
SC13	37	14	1	research
SC14	20	8	1	research
SC15	7	6	1	dining
SC16	12	9	1	dining
SC17	12	11	1	dining
SC18	32	11	1	flying teams
SC19	33	13	1	flying teams
SC20	7	5	1	flying teams
SC21	6	5	1	greek
SC22	8	6	1	greek
SC23	8	5	1	greek
SC24	24	12	1	karate
SC25	5	5	1	karate
SC26	24	13	1	karate
SC27	4	4	1	karate
SC28	22	15	1	prison
SC29	13	8	1	prison
SC30	17	11	1	prison
SC31	19	14	1	prison
SC32	12	7	1	prison
SC33	12	7	1	prison
SC34	9	5	1	prison
SC35	27	13	1	pupils
SC36	12	7	1	pupils
SC37	6	6	1	pupils
SC38	8	6	1	pupils
SC39	3	4	1	pupils
SC40	19	11	1	pupils
SC41	10	5	1	sawmill
SC42	35	12	1	sawmill
SC43	17	11	1	sawmill
SC44	3	4	1	sawmill
SC45	18	7	1	thurman
SC46	4	5	1	thurman
SC47	15	10	1	wood
SC48	9	7	1	wood

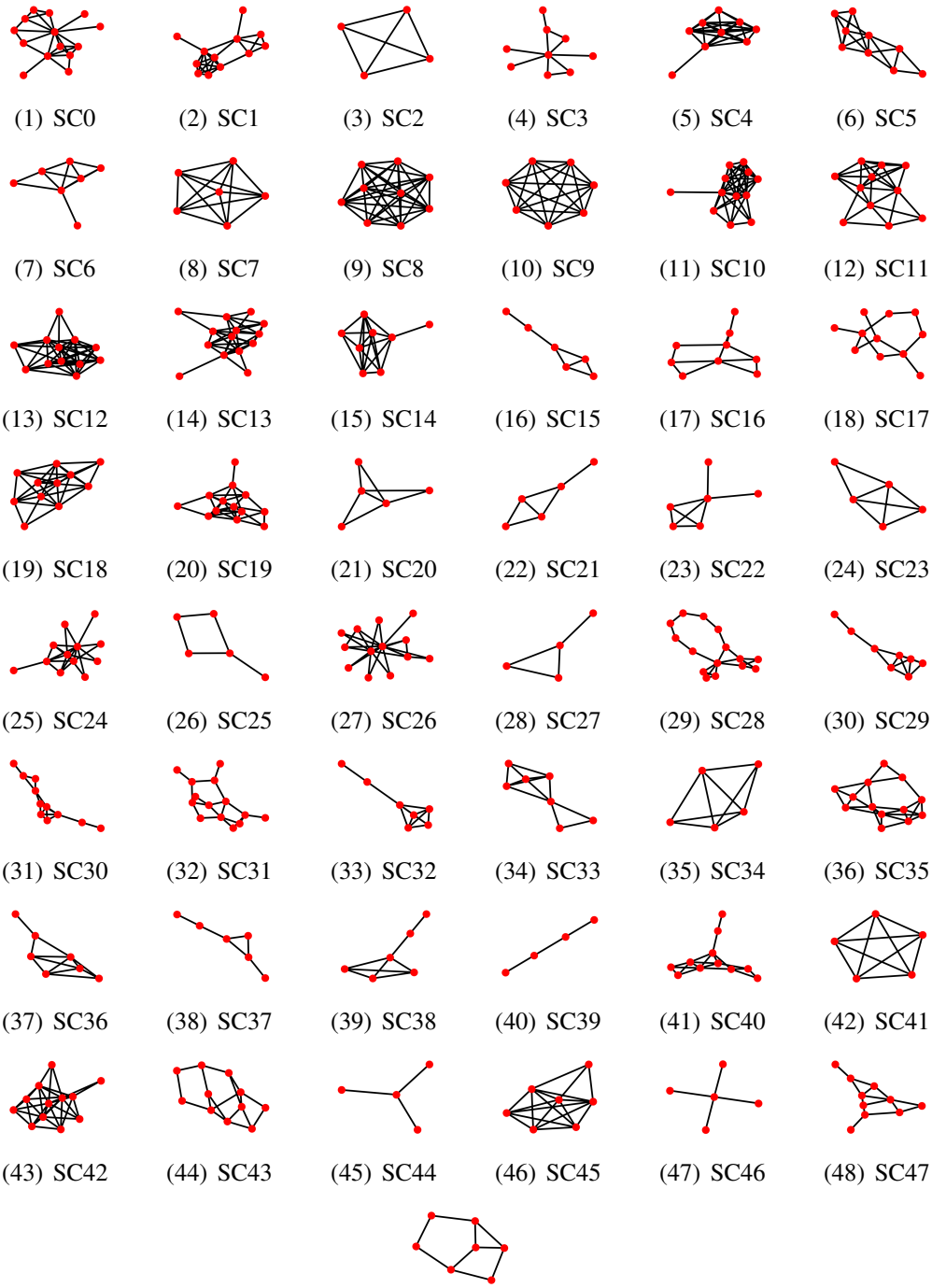


Figure A.3: Social Cliques in the Offline Network

Appendix B

The Role of Offline Friends in Information Diffusion

B.1 Investigating the Role of Reciprocal Ties for Information Diffusion of Various Topics on Twitter

B.1.1 Tweets Considered in the Study

Topic	id*	basic SEIZ err
Controversial	11072	Well there you have it. A highly intelligent experienced woman just debated a giant orange Twitter egg. Your move America. #debate
	12512	Time to #DrainTheSwamp in Washington D.C. and VOTE #TrumpPence16 on 11/8/2016. Together we will MAKE AMERICA SAFE https://t.co/rVcjXdWxzp
	94368	Voter fraud! Crooked Hillary Clinton even got the questions to a debate and nobody says a word. Can you imagine if I got the questions?
	67680	RT if youre proud of Hillary tonight. #DebateNight #SheWon https://t.co/H7CJep7APX
	65409	BRITAIN: Brexit is the stupidest most self-destructive act a country could undertake. USA: Hold my beer.
	90400	Retweet if you are: -A woman -An immigrant -LGBT+ -Muslim -African American -Latino/Latina -In any other way completely terrified right now
	21408	how can immigrants be lazy and stealing your jobs at the same time https://t.co/Iq6hicy7mt
	99648	#BlackLivesMatter https://t.co/y2yHoDuDJb
	24992	Looking at Air Force One @ MIA. Why is he campaigning instead of creating jobs & fixing Obamacare? Get back to work for the American people!
News	27456	WE WERE OUT HERE PRAYING FOR FLORIDA TO SAY SAFE FROM HURRICANE MATTHEW. LITTLE DID WE KNOW. HURRICANE MATTHEW WAS https://t.co/DRbKFRbkhv
	07328	”Florida just got hit by a category 5 Hurricane! Please donate.” Me: https://t.co/xYjALm72Gw
	35588	hurricane chris really 10 steps ahead of us all https://t.co/npnY4Nk2TW
	08385	”Florida just got hit by a category 5 Hurricane! Please donate.” Me: https://t.co/xYjALm72Gw
	96992	It takes 3.2sec to retweet and help find missing Isabella Gonzalez she went missing from #Vegas #usa a year ago tod https://t.co/iNq51MjeXe

Personal	58560	#Haiti We must provide safe water & hygiene items quickly to avoid spread of disease. You CAN help https://t.co/T5jIjIF2Ia
	22432	Welcome to California we just had our 3 days of fall/winter/spring and now it's back to summer for the rest of the year
	46528	I remember always running around the house in my underwear and playing games with Ashton and Brandon #tb
Entertainment	55136	All Weekend Long: Official Music Video https://t.co/VRvN60NU1v #AWLMusicVideo
	38880	#PurposeTour in #mycalvins https://t.co/FahXxb3JsL
	39264	See u at #iHeartJingleBall
	36069	New collection of #PurposeTour merch available now at @pacsun https://t.co/lT4kNEpAQ7 https://t.co/fsKAvWWsTJ
	79424	#AYO @TheChainsmokers maybe u guys'll like this 1 better #NowPlaying Apple: https://t.co/u0r8kOeLCr Spotify: https://t.co/iCwMjenJut
	46656	WE GO TO CHICAGO!!! #LALovesOctober https://t.co/NH7DGg2B3J
	54821	English football's most successful clubs are showing why they are meeting on a Monday night in Champions League week
	27680	Cubs win! We take a 3-2 #NLCS lead! Final: #Cubs 8 #Dodgers 4. #FlyTheW https://t.co/CID6ydaYec
Rewards	97216	Wrigley Field will be loud tomorrow. RT this for your chance to win two tickets to #NLCS Game 6! #FlyTheW https://t.co/L0mwAGmNSV
	86144	RT TO WIN: OYSTER BRUSH ROLL FROM SPECTRUM ($100+) ? (must be following me & @SpectrumBrushes so we can dm winner) https://t.co/0BeBHZweQ7
	13504	RT TO WIN: SLEEK MAKEUP 'PRECIOUS METALS' HIGHLIGHTING PALETTE ? (must be following me so I can dm winner) https://t.co/hYfSs3UPNb
	65888	RT TO WIN: ABH GLOW KIT OF CHOICE ? (must be following me to win) https://t.co/Jyfugt2v8s
	30240	RT TO WIN: Morphe 12-piece brush set ? (must be following me so I can dm winner) https://t.co/o0RTv9jlqy
*The last 5 digits of tweet id.		

B.2 Investigating the Role of Offline Friends on Two-Hop Information

B.2.1 Words Distribution of the Extracted Topics

Topic 0	don	0.011	girl	0.005	drink	0.003
	people	0.007	dog	0.004	lol	0.003
	shit	0.006	good	0.004	kids	0.003
	fuck	0.006	love	0.004	face	0.003
	eat	0.006	ukraine	0.004	hate	0.003
	time	0.005	cat	0.004	gonna	0.003
	man	0.005	fucking	0.004	hot	0.003
	guy	0.005	thing	0.004	sex	0.003
	day	0.005	baby	0.004	women	0.003
	food	0.005	wife	0.003	night	0.003
Topic 1	live	0.019	bit	0.007	ready	0.006
	tonight	0.016	album	0.007	excited	0.006
	watch	0.011	week	0.007	follow	0.005
	video	0.011	coming	0.006	sale	0.005

	tomorrow	0.011	free	0.006	youtube	0.005
	today	0.011	lt	0.006	guys	0.005
	check	0.010	music	0.006	friday	0.005
	day	0.009	tickets	0.006	tweet	0.005
	win	0.009	night	0.006	don	0.005
	join	0.008	episode	0.006	season	0.004
Topic 2	san	0.006	church	0.003	report	0.002
	people	0.006	francisco	0.003	york	0.002
	news	0.0048	board	0.003	president	0.002
	trump	0.005	police	0.003	attack	0.002
	wee	0.004	hurricane	0.003	man	0.002
	breaking	0.004	killed	0.003	today	0.002
	puerto	0.004	dead	0.003	years	0.002
	city	0.004	death	0.003	state	0.002
	chew	0.004	bonds	0.003	victims	0.002
	rico	0.004	time	0.003	case	0.002
Topic 3	singapore	0.021	vote	0.005	tony	0.004
	pap	0.015	day	0.005	mp	0.004
	lee	0.012	dr	0.004	president	0.004
	pm	0.008	people	0.004	national	0.003
	bit	0.007	singaporeans	0.004	breaking	0.003
	party	0.007	evans	0.004	st	0.003
	election	0.007	yew	0.004	votes	0.003
	ow	0.007	news	0.004	workers	0.003
	minister	0.007	helium	0.004	police	0.003
	parliament	0.006	rally	0.004	opposition	0.003
Topic 4	team	0.013	film	0.007	watch	0.004
	great	0.010	season	0.006	congratulations	0.004
	day	0.008	opening	0.006	good	0.004
	chicago	0.008	win	0.006	game	0.004
	race	0.008	week	0.006	closing	0.004
	today	0.007	weekend	0.005	india	0.004
	tonight	0.007	final	0.005	night	0.004
	theatre	0.007	congrats	0.005	recommended	0.004
	time	0.007	happy	0.005	fans	0.004
	world	0.007	year	0.005	amazing	0.003
Topic 5	people	0.010	china	0.003	kong	0.003
	police	0.008	bit	0.003	year	0.003
	singapore	0.008	don	0.003	death	0.003
	man	0.007	years	0.003	call	0.003
	car	0.005	missing	0.003	cars	0.002
	white	0.005	killed	0.003	water	0.002
	black	0.004	chinese	0.003	road	0.002
	news	0.004	hong	0.003	country	0.002
	bus	0.004	dead	0.003	media	0.002
	post	0.004	train	0.003	lost	0.002
Topic 6	students	0.009	stay	0.006	illinois	0.004
	great	0.009	downloads	0.005	job	0.004
	twitter	0.008	design	0.005	team	0.004
	bit	0.007	join	0.005	check	0.004
	school	0.007	read	0.005	singapore	0.004
	social	0.007	learn	0.004	learning	0.004
	post	0.007	today	0.004	people	0.004
	media	0.006	day	0.004	kids	0.003
	work	0.006	time	0.004	facebook	0.003
	blog	0.006	linger	0.004	education	0.003

Topic 7	trump	0.023	women	0.005	state	0.003
	president	0.020	health	0.004	states	0.003
	obama	0.015	americans	0.004	india	0.003
	people	0.012	tax	0.004	senate	0.003
	vote	0.007	congress	0.004	breaking	0.003
	house	0.006	america	0.004	election	0.003
	white	0.005	care	0.003	gop	0.003
	donald	0.005	don	0.003	today	0.003
	country	0.005	american	0.003	china	0.003
	bill	0.005	news	0.003	time	0.003
Topic 8	latest	0.066	php	0.004	est	0.003
	price	0.065	vitamins	0.004	contra	0.003
	bitcoin	0.065	gran	0.004	pas	0.002
	usd	0.060	fa	0.004	mi	0.002
	harry	0.009	nit	0.004	people	0.002
	te	0.005	benefits	0.003	primera	0.002
	potter	0.005	hem	0.003	gent	0.002
	posted	0.005	fins	0.003	blog	0.002
	prince	0.005	vols	0.003	son	0.002
	health	0.005	luis	0.003	tweet	0.002
Topic 9	singapore	0.009	track	0.004	massa	0.003
	time	0.008	result	0.004	psi	0.003
	weather	0.007	today	0.004	water	0.0033
	rain	0.006	gp	0.004	high	0.003
	morning	0.006	car	0.004	button	0.003
	day	0.006	people	0.004	bus	0.003
	train	0.005	bit	0.004	air	0.003
	hamilton	0.005	japan	0.004	race	0.003
	update	0.005	raikkonen	0.003	late	0.003
	hour	0.004	east	0.003	afternoon	0.003
Topic 10	love	0.011	day	0.005	amazing	0.004
	video	0.009	film	0.005	youtube	0.004
	song	0.008	happy	0.005	great	0.004
	tonight	0.007	guys	0.005	fun	0.004
	watch	0.007	today	0.005	ago	0.004
	music	0.006	time	0.004	watching	0.004
	years	0.006	awesome	0.004	star	0.003
	man	0.006	game	0.004	good	0.003
	super	0.006	movie	0.004	lol	0.003
	night	0.006	vote	0.004	birthday	0.003
Topic 11	apple	0.021	store	0.006	video	0.004
	iphone	0.019	apps	0.006	account	0.004
	app	0.015	free	0.005	windows	0.003
	bit	0.014	phone	0.005	singapore	0.003
	google	0.014	update	0.005	samsung	0.003
	twitter	0.011	post	0.005	check	0.003
	ios	0.010	users	0.004	pay	0.003
	android	0.090	mac	0.004	buy	0.003
	ipad	0.008	mobile	0.004	web	0.003
	facebook	0.008	blog	0.004	pro	0.003
Topic 12	people	0.022	birthday	0.009	thing	0.005
	love	0.022	twitter	0.007	great	0.005
	don	0.019	today	0.007	guys	0.004
	happy	0.017	feel	0.007	hope	0.004
	life	0.016	friends	0.006	man	0.004
	day	0.015	work	0.006	hard	0.004

	good	0.013	person	0.005	family	0.004
	god	0.011	heart	0.005	night	0.003
	time	0.010	bad	0.005	sleep	0.003
	things	0.010	world	0.005	real	0.003
Topic 13	tech	0.018	app	0.007	marketing	0.004
	social	0.012	facebook	0.006	online	0.004
	google	0.011	digital	0.006	platform	0.004
	mobile	0.010	business	0.006	post	0.004
	eu	0.010	twitter	0.005	cloud	0.004
	bit	0.010	web	0.005	future	0.004
	media	0.009	learning	0.005	great	0.004
	data	0.008	company	0.005	funding	0.003
	raises	0.008	free	0.004	asia	0.003
	ibm	0.007	technology	0.004	video	0.003
Topic 14	money	0.008	job	0.004	play	0.003
	day	0.007	year	0.004	jobs	0.003
	people	0.006	save	0.004	men	0.003
	school	0.005	bit	0.004	pay	0.003
	kids	0.005	work	0.004	cancer	0.003
	time	0.005	win	0.003	high	0.003
	christmas	0.005	food	0.003	world	0.003
	don	0.004	tips	0.003	free	0.003
	women	0.004	children	0.003	art	0.002
	life	0.004	prize	0.003	find	0.002

B.3 Measuring Tie Strength Offline vs. Online: Is Redefinition of Tie Strength Necessary on a Social Information Network?

B.3.1 Words Distribution of the Extracted Topics

Topic 0	check	0.0127	insurance	0.004	tips	0.003
	free	0.007	time	0.004	followers	0.003
	health	0.005	twitter	0.004	service	0.003
	pool	0.005	marketing	0.004	website	0.003
	work	0.005	personal	0.004	background	0.003
	business	0.005	people	0.004	today	0.003
	great	0.005	media	0.004	site	0.003
	online	0.004	weight	0.004	loss	0.003
	social	0.004	good	0.003	find	0.003
	money	0.004	facebook	0.003	day	0.003
Topic 1	http	0.006	town	0.003	love	0.002
	india	0.005	thefancy	0.003	ka	0.002
	hai	0.0048	morgandorr	0.003	ho	0.002
	kitty	0.004	boyslikegirls	0.003	ko	0.0019
	martinsays	0.004	narendramodi	0.002	johnblg	0.002
	kawaii	0.004	day	0.002	paulblg	0.002
	time	0.004	check	0.002	nie	0.002
	good	0.0033	https	0.002	sir	0.002
	hellokittykawaiitown	0.0032	happy	0.002	event	0.002
	hello	0.0032	bessemervp	0.002	indian	0.002
Topic 2	people	0.012	feel	0.004	find	0.003
	don	0.012	heart	0.004	dont	0.003

	love	0.0116	person	0.004	work	0.003
	life	0.0113	world	0.004	bad	0.002
	time	0.0076	thing	0.004	im	0.002
	things	0.0061	happy	0.003	hard	0.002
	day	0.0058	man	0.003	hate	0.002
	good	0.0057	friends	0.003	years	0.002
	today	0.0054	girl	0.003	shit	0.002
	god	0.0054	year	0.003	its	0.002
Topic 3	collection	0.009	today	0.005	coming	0.004
	shop	0.009	selling	0.005	items	0.004
	dress	0.009	http	0.005	singapore	0.004
	lt	0.008	arrivals	0.004	thecarousell	0.004
	sale	0.007	launched	0.004	bag	0.004
	check	0.007	preorder	0.004	visit	0.003
	free	0.006	love	0.004	carousell	0.003
	top	0.006	day	0.004	giveaway	0.003
	black	0.006	facebook	0.004	follow	0.003
	time	0.0052	win	0.004	sales	0.003
Topic 4	day	0.012	night	0.004	coffee	0.003
	singapore	0.009	ice	0.004	chocolate	0.003
	happy	0.0085	lunch	0.004	hair	0.003
	food	0.0074	cream	0.004	cake	0.003
	good	0.0064	year	0.004	free	0.003
	time	0.0061	christmas	0.003	awesome	0.003
	today	0.0058	chicken	0.003	craving	0.003
	love	0.0055	great	0.003	finally	0.003
	dinner	0.0054	morning	0.003	tea	0.002
	birthday	0.0048	eat	0.003	breakfast	0.002
Topic 5	les	0.016	dans	0.006	par	0.003
	en	0.015	il	0.005	ne	0.003
	est	0.0120	qui	0.005	qu	0.003
	pour	0.012	avec	0.005	mais	0.003
	des	0.012	ce	0.005	paris	0.003
	du	0.010	vous	0.004	dr	0.003
	sur	0.008	blog	0.003	se	0.002
	une	0.008	notre	0.003	editionsjentayu	0.002
	pas	0.007	nous	0.003	aux	0.002
	au	0.006	france	0.003	satyapal	0.002
Topic 6	singapore	0.009	happy	0.004	sg	0.003
	today	0.009	tomorrow	0.004	boys	0.003
	day	0.008	music	0.004	don	0.003
	time	0.006	win	0.004	watch	0.003
	love	0.006	night	0.003	check	0.003
	good	0.006	week	0.003	people	0.003
	year	0.006	coming	0.003	live	0.003
	libra	0.005	school	0.003	hope	0.003
	guys	0.004	tonight	0.003	days	0.003
	great	0.004	song	0.003	hey	0.003
Topic 7	man	0.007	united	0.005	world	0.003
	game	0.006	team	0.004	people	0.003
	person	0.006	peopleschoice	0.004	unfollowed	0.003
	checked	0.006	breakoutartist	0.004	chelsea	0.003
	automatically	0.006	season	0.004	football	0.003
	time	0.005	play	0.004	lol	0.003
	wanted	0.0051	today	0.003	manutd	0.003
	win	0.0050	match	0.003	don	0.003
	arsenal	0.0050	league	0.003	manchester	0.003

	good	0.0050	goal	0.003	liverpool	0.003
Topic 8	time	0.012	school	0.005	night	0.004
	lol	0.011	life	0.005	feeling	0.004
	day	0.0106	days	0.005	finally	0.003
	don	0.0093	damn	0.005	tired	0.003
	good	0.0078	week	0.005	wanna	0.003
	today	0.0073	gonna	0.004	wait	0.003
	omg	0.0067	people	0.004	man	0.003
	sleep	0.0062	long	0.004	tmr	0.003
	work	0.0059	bad	0.004	fuck	0.003
	feel	0.0055	shit	0.004	year	0.003
Topic 9	lol	0.017	sia	0.004	long	0.003
	don	0.007	love	0.004	watch	0.003
	omg	0.007	yeah	0.004	sleep	0.003
	time	0.007	lt	0.004	alr	0.003
	good	0.007	nice	0.003	birthday	0.003
	happy	0.005	man	0.003	bad	0.002
	day	0.005	dont	0.003	pls	0.002
	hahah	0.004	today	0.003	guys	0.002
	damn	0.004	tmr	0.003	hahahah	0.002
	ur	0.004	wanna	0.003	gonna	0.002
Topic 10	photo	0.024	video	0.005	youtube	0.003
	posted	0.023	year	0.005	today	0.003
	facebook	0.018	https	0.005	superhero	0.003
	bts	0.012	wait	0.005	wall	0.003
	twt	0.006	lovedrunk	0.005	birthday	0.003
	album	0.006	happy	0.004	indian	0.003
	love	0.006	lol	0.004	time	0.003
	omg	0.006	long	0.004	top	0.003
	photos	0.006	lt	0.003	good	0.003
	day	0.005	song	0.003	krrish3firstlook	0.003
Topic 11	el	0.031	con	0.006	pero	0.003
	en	0.023	si	0.006	telecogresca	0.003
	es	0.013	las	0.005	esta	0.003
	del	0.0098	te	0.005	mi	0.003
	una	0.0087	les	0.005	ser	0.002
	los	0.0084	els	0.004	ja	0.002
	se	0.0077	s	0.004	este	0.002
	al	0.0073	amb	0.004	su	0.002
	por	0.0067	como	0.003	ms	0.002
	para	0.0066	ms	0.003	va	0.002
Topic 12	https	0.017	ada	0.003	apa	0.002
	aku	0.007	bulu	0.003	lol	0.002
	tak	0.006	dia	0.003	activities	0.002
	nak	0.005	stay	0.003	events	0.002
	mola	0.005	snail	0.003	exciting	0.002
	kau	0.004	trecru	0.003	taxi	0.002
	yg	0.004	transponder	0.003	buat	0.002
	yang	0.004	ini	0.002	hari	0.002
	lt	0.003	dan	0.002	driver	0.002
	gwddqd41ox	0.003	tuned	0.002	ctwr2x7888	0.002
Topic 13	youtube	0.015	whatsapp	0.004	don	0.003
	video	0.015	music	0.004	uae	0.003
	blog	0.013	sms	0.003	gaga	0.003
	post	0.009	hope	0.003	feed	0.003
	women	0.005	love	0.003	lt	0.003

	dubai	0.004	good	0.003	playlist	0.003
	massage	0.004	check	0.003	smallzy	0.003
	web	0.004	time	0.003	week	0.002
	abu	0.004	couples	0.003	man	0.002
	lol	0.004	dhabi	0.003	watch	0.002
Topic 14	today	0.004	good	0.003	work	0.002
	mobile	0.004	data	0.002	tech	0.002
	singapore	0.004	iphone	0.002	kuala	0.002
	world	0.003	google	0.002	stcom	0.002
	apple	0.003	app	0.002	lumpur	0.002
	news	0.003	day	0.002	facebook	0.002
	time	0.003	asia	0.002	read	0.002
	people	0.003	china	0.002	years	0.002
	great	0.003	trump	0.002	android	0.002
	year	0.003	india	0.002	free	0.002