Studying Social Influence on Twitter

By

Yan Mei

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

Department of Computing
Faculty of Science and Engineering
Macquarie University
Supervisor: Prof. Jian Yang

August 2017

Declaration

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university.

_____

Yan Mei

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor Prof. Jian Yang. This dissertation cannot be completed without her continuous guidance and tremendous help. During the past four years, I have learnt a lot from her, and her rigorous attitude and enormous passion for research will always motivate me to pursue excellence. I am also truly grateful to Dr. Weiliang Zhao for his generous time and valuable feedback on the dissertation. My thoughts and writing style are deeply inspired by him. In a way, he has been my unofficial co-supervisor.

I would like to thank other smart and hard-working Ph.D. students in our research group Robertus Nugroho, Zizhu Zhang, Lei Han and Pengbo Xiu. Discussion with them gives me great inspiration on my research. It has been a lot of fun and honor to work with them. The last four years being with them are filled with good memories.

Many thanks are given to the administration team of the department. Kind support and wonderful environment they have provided are important for me to complete the dissertation.

Last, but most importantly, I am greatly indebted to my family and especially my beloved wife Minjun Ma. My Ph.D. has been a rather long journey since I decided to pursue an academic career. My parents, parents-in-law, and my wife have been very understanding and supportive all the years. Nothing would have been possible without their love, support and sacrifice.

Abstract

The study of social influence has a long history in research areas of sociology and marketing. In recent years, with the rapid growth of Online Scoial Networks (OSNs), online influence has received lots of attention from both academic community and industry. For the work in this dissertation, we consider the Twitter platform and aim to address two problems: 1) feature selection for measuring social influence; and 2) influence maximization for marketing campaigns.

While many researchers focus on measuring social influence on Twitter, there is still lacking of a comprehensive analysis of feature selection. Most existing studies directly utilize their own pre-defined features to build the model without evaluation and judgment for these selected features. In order to find principal features for measuring user influence on Twitter, we select manifest features based on sociology knowledge. Besides principal manifest features, we identify hidden features and map them to the attributes of influencers in the research area of social science. Furthermore, we propose a hybrid feature selection method for predicting user influence. After evaluating

the quality of features by utilizing a filter method, a reduced feature subset is obtained. Following the principles of wrapper methods, we assess the feature subset at each searching step. Finally, an optimal feature set with a high degree of accuracy for predicting user influence is obtained .

Influence maximization is the most fundamental and important problem when studying social influence. In this work, we identify a specific influence maximization problem as selecting a set of seed users to maximize the effectiveness of advertising campaigns on Twitter. When studying influence maximization problem, we develop our solution with new ideas focusing on : 1) the definition of influence; 2) the influence probability model; 3) the influence diffusion model; and 4) the seed nodes selection algorithm. The proposed influence maximization approach has taken into consideration of social ties, user interactions, and the characteristics of advertising information propagation on Twitter. Our work provides a solid generic solution for promoting products or services in online social networks like Twitter.

List of Publications

[1] Yan Mei, Youliang Zhong, and Jian Yang. Finding and analyzing principal features for measuring user influence on Twitter. In *International Conference on Big Data Computing Service and Applications (BigDataService)*, pages 478-486, IEEE, 2015.

[2] Yan Mei, Zizhu Zhang, Weiliang Zhao, Jian Yang, and Robertus Nugroho. A hybrid feature selection method for predicting user influence on Twitter. In *International Conference on Web Information Systems Engineering*, pages 478-492, Springer, 2015.

[3] Robertus Nugroho, Weiliang Zhao, Jian Yang, Cecile Paris, Surya Nepal, and Yan Mei. Time-sensitive topic derivation in Twitter. In *International Conference on Web Information Systems Engineering*, pages 138-152, Springer, 2015.

[4] Yan Mei, Weiliang Zhao, and Jian Yang. Influence maximization on Twitter: a mechanism for effective marketing campaign. In *International Conference on Communications (ICC)*, pages 1-6, IEEE, 2017.

[5] Yan Mei, Weiliang Zhao, and Jian Yang. Maximizing the effectiveness of advertising campaigns on Twitter. In *International Congress on Big Data*, pages 73-80, IEEE, 2017.

Contents

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

1

Introduction

A few decades ago, no one could ever imagine that such a thing as the Internet would be invented. Nowadays, it is hard to imagine a life without the Internet. The Internet provides an enormous amount of information. People can find out the information they are interested in almost any topic. It is widely believed that the Internet was one of the greatest inventions of the 20th century that changed the world [1, 2].

In recent years, with the rapid growth of Online Social Networks (OSNs) including Twitter, Facebook, Google+, and LinkedIn, etc. there has been a revolutionary change in the way people communicate with each other. In OSNs, people from all over the world can stay in touch, share experience, publish information, exchange opinions, or join discussions. Fig. 1.1 shows the facts and statistics for six of the largest social media sites in the world [1]. These websites not only provide individual users platforms to share information and keep in touch

[1] http://marketingstrategyx.com/social-media-stats-infographic-2017/

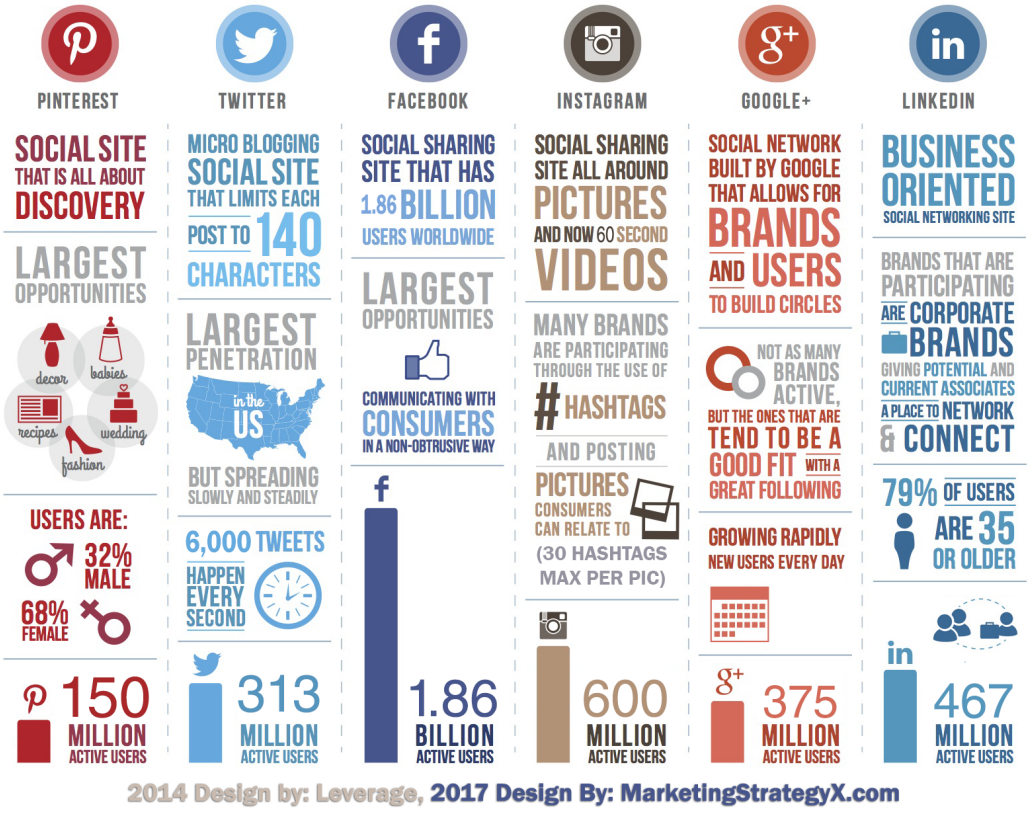

Figure 1.1: Social Media Stats Infographic 2017

with their friends, but also become important marketing channels for companies and organizations.

Social media advertising has made a great progress in a relatively short period of time. When Facebook launched its first advertising option in May 2005, no one could have predicted that social media advertising revenue reached 17.08 billion US dollars in 2015 [1], only ten years later.

Influencer marketing is the process of identifying, researching, engaging and supporting the people who create high-impact conversations with customers about

[1] http://www.statista.com/statistics/271258/facebooks-advertising-revenue-worldwide/

your brand, products or services. Influencer marketing offers brands the potential to unify their marketing, sales, product, digital marketing, and social media through powerful and relevant relationship-based communication. Both the ROI (Return On Investment) and marketing potential of influencer marketing are immense.

As online customer conversations increase drastically, influencers are playing a critical role in breaking online clutter, creating relevant customer dialogue and bringing trust to the table for brands and marketers alike. It should be no surprise that an increasingly digital landscape is changing the way to do business, the way for customers to access information and make decisions.

For the visionary marketers, the rise of the social media influencers creates a lot of new possibilities. It opens up a new channel for brands to connect with consumers more directly, more organically, and at a large scale. With the help of social media influencers, brands can amplify their message while seducing their target audience. Traditionally, consumers made purchasing decisions based on the advertisements that they saw or heard. Today, it is easier to connect with other consumers via social media and make better purchasing decisions by learning about their experiences with products or services.

Critics of the online marketing approach argue that only researching online sources misses critical influential individuals and inputs [3]. They note that much influential exchange of information occurs in the offline world, and is not captured

in online media. Indeed, the majority of consumer exchanges occurs face-to-face, not in an online environment, as evidenced by Carl [4]. However, as the world has shifted to social media, social networks provide a great opportunity to promote new products or ideas because of the large number of users and the high frequency of communication. More and more companies or organizations are paying attention to how their brands are discussed online and recent academic research has focused on online WOM (Word of Mouth).

2016 Social Media Marketing Industry Report published by Social Media Examiner [1] is generated based on surveys of more than 5,000 marketers from all around the world. Fig. 1.2 shows some results in this report. A significant 90% of marketers said that social media was important to their businesses. A significant 89% of all marketers indicated that their social media efforts had generated more exposure for their businesses. As for time commitment for social media marketing, a significant 63% of marketers were using social media for 6 hours or more and 39% for 11 or more hours weekly.

1.1 Motivation

Social influence is defined as the change in a person's cognition, attitude, or behavior, which has its origin in another person or group [5]. Social influence theory has been studied extensively in sociology and psychology [6, 7, 8, 9, 10, 11], since the mid-20th century.

[1] http://www.socialmediaexaminer.com/

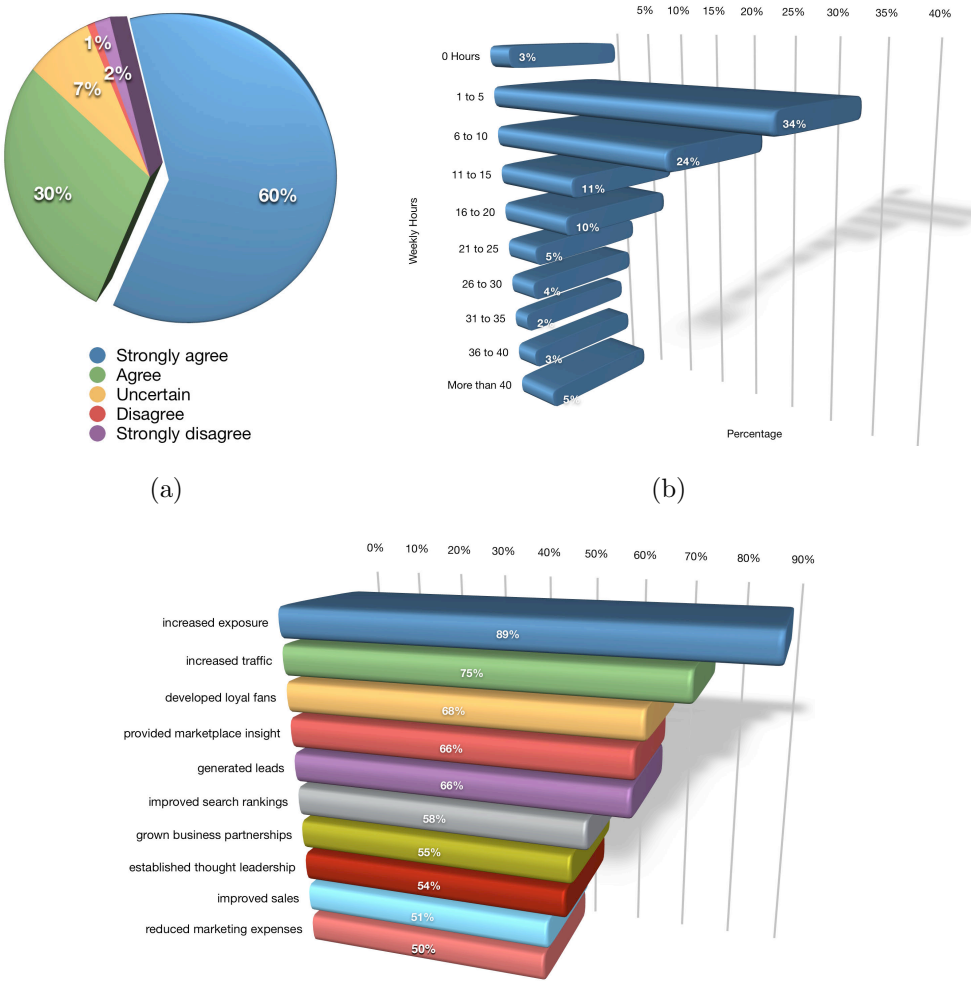

Figure 1.2: 2016 Social Media Marketing Industry Report. (a)Social media is important for my business; (b)Weekly time commitment for social media marketing; (c)Benefits of social media marketing.

Social influence can be exploited in many applications. The most widely recognized application is viral marketing (a.k.a. viral advertising). Consider an online social network where users can perform various actions. As an example, a user on Twitter just bought a new laptop and then posted a tweet sharing his experience. As another example, a user on Netflix rates a movie he just watched, which is also an action. As a third example, a user on Facebook invites his friends to join a coming event (e.g. a Fitness Show). These actions could bring reactions from his own social network. For instance, some of his friends might buy the laptop on which he made a positive comment, watch the movie which he gave a 5 star rating or go to the event he recommended. We consider this process as influence, and it can be propagated.

Such influence patterns can be of interest to companies. For example, if we know that there are some "leaders" who have the ability to lead the trend for various actions, then targeting them to adopt new products or technology, it could help companies to increase their profit. This kind of targeted advertising is called "Viral Marketing". The idea behind is to identify a small number of key influential individuals in a social network, whose comments or recommendations have enormous influence on others and finally might lead to a large number of adoptions of their recommendations.

In addition to viral marketing, social influence has been leveraged in other applications like recommender systems [12, 13, 14], events detection [15, 16],

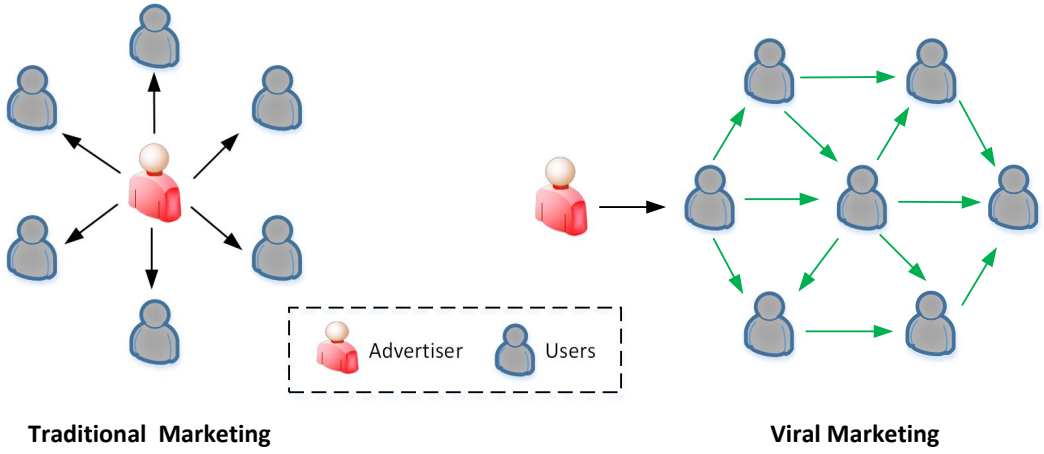

Figure 1.3: Viral Marketing v.s. Traditional Marketing

community detection [17, 18], expert finding [19, 20], link prediction [21, 22, 23], etc. Other interesting problems like outbreak detection [24, 25] and epidemics on networks [26, 27], are also related to the study of social influence.

One of the fundamental problems in the study of social influence is the problem of **Influence Maximization**, motivated by the application in viral marketing. The problem was originally defined by Kempe et al. [28] as follows. Given a directed graph $G = (V, E)$ in which vertices represent individuals in a social network and edges represent the links or relationships between individuals, as well as a positive integer k, the task of influence maximization is to find a seed set S of size k, such that by targeting them initially for early activation, the expected influence spread (defined as the expected number of activated nodes in the social network) is maximized, under a certain diffusion model. Note that it

is the diffusion model that governs how influence diffuses or propagates through the network.

Although considerable research has recently been done on influence maximization, there are still some limitations and open problems that need to be addressed.

Firstly, it is still a problem how to obtain or calculate the influence probabilities. The framework employed in existing solutions, including that reported by Kempe et al. [28] requires two types of input data, a directed graph and an assignment of probabilities (or weights) to the edges of the graph, representing degrees of influence between users. In real online social networks, the graph representing the network structure is often explicit, but the probabilities associated with edges are not. Kempe et al. [28] assign each edge of the co-authorship graph a uniform probability (1% or 10%) in their experiments. Chen et al. [29] use three pre-determined propagation probabilities of $p = 0.01$, 0.02 and 0.05 in their work. Another popular model is referred to as "weighted cascade" [28, 29] in which the probability of user u influencing v is assigned with the value $1/d_v$, where d_v denotes how many people user v follows. Although some recent studies develop their methods to learn the influence probabilities from historical data [30, 31], it is still lacking of enough research with details to consider the specific features of a social network when modelling influence probabilities.

Secondly, since influence maximization problem was formally proposed by

Kempe et al. [28], the inefficiency of the solution algorithms has drawn much attention and become a research hot spot. Kempe et al. [28] presented a simple greedy algorithm which repeatedly picks the node with the maximum marginal gain and adds it to the seed set, until the budget k is reached. This algorithm is known to work out the best possible seed set (in terms of influence spread), so is commonly used as a benchmark algorithm. However, computing the expected influence spread under both the independent cascade and linear threshold models is $\#P$-hard [32, 33]. The Monte Carlo simulation that runs a large number of times to obtain an accurate estimate of the influence spread is very time consuming. The simple greedy algorithm makes $O(nk)$ calls to the spread estimation function where n is the number of users in the network and k is the seed set size (a.k.a. budget). Thus, the greedy algorithm cannot be applied to deal with large real world social networks.

Thirdly, it is crucial to choose an appropriate diffusion model in the influence maximization problem. The diffusion model reflects how influence propagates in the social network. The expected influence spread and the selection of seed users are directly related to the diffusion model. There are classic diffusion models, such as Linear Threshold model and Independent Cascade model [28], as well as many extension models based on the classic models, such as [34, 35, 36, 37, 38, 39, 40, 41, 42], etc. A well-defined diffusion model should capture the major characteristics of influence propagation in a specific social network.

Fourthly, the optimization criteria of influence maximization might need adjustment when solving specific problems. The objective of classical influence maximization problem defined by Kempe et al. [28] is to identify a small number of key individuals in a social network such that by targeting them, a large number of users get influenced eventually. There is a limited budget (i.e. the size of seed set) and the optimization goal is to maximize the expected number of activated users at the end of propagation. This definition has its limitation and cannot cover various characteristics of a realistic problem. For example, there could be some requirements on timeliness, which means only the propagation during a limited time period should be considered. Besides, the primary motivation of the influence maximization problem is viral marketing, but an activated or influenced user may not necessarily adopt an advertising product or recommend the product to his/her friends. There is a gap between influence and product adoption. Thus, it is desirable to formulate and study alternative optimization criteria in order to bridge this gap.

Suppose a company has released a new product or an online service. This company would like to promote or advertise its new product on Twitter. Since the company's Twitter account only has a small amount of followers (seems not influential), the company needs to select a set of users (i.e. seed users) to help the propagation of the marketing information on Twitter. The company expects that these seed users will influence their followers, and then these followers will

influence their own followers as well. As a result, a large number of users on Twitter could receive this marketing information through the online word-of-mouth effect. Due to the constraints of the budget or relevant resources, it is necessary to find a set of seed users to maximize the expected coverage.

1.2 Research Problems

The research presented in this dissertation focuses on two problems as: feature selection for measuring social influence, and influence maximization for marketing campaigns on Twitter.

While many researchers focus on the measurement of social influence on Twitter, there is a lack of comprehensive analysis regarding the effectiveness of the principal features for measuring user influence. Most existing studies directly utilize their own pre-defined features to build the model. We believe that identifying important features which are crucial for influence measurement is the first step towards influence model construction. Our research aims to find the most effective features for measuring user influence.

One of the fundamental problems in the field of studying social influence is the problem of influence maximization, primarily motivated by the application of viral marketing. The problem was first proposed by Kempe et al. [28] and has received lots of research interests from both academic community and industry. There are several key questions we need to address when solving a specific influence

maximization problem.

1. What is the definition of influence? Influence is an abstract concept which could show in various ways in different contexts. The first question is what kind of influence should be maximized.

2. How to calculate influence probabilities? This information cannot be directly obtained from the social networks. A proper model should be designed to calculate the influence probability between users.

3. How is the influence diffused or propagated? The diffusion model determines how influence propagates in the social network. It should capture the characteristics in the dynamics of influence diffusion.

4. How to select the optimal or near-optimal seed set? The seed nodes selection algorithm is used to select the influential users from the social network in order to maximize the expected influence spread.

In this dissertation, we study a specific influence maximization problem, selecting a set of seed users to maximize the effectiveness of advertising campaigns on Twitter. All the questions listed above will be addressed.

1.3 Contributions

Based on the research problems described above, we make the following key contributions in this dissertation (for detailed contributions, see the respective chap-

ters).

1. In order to find principal features for measuring user influence on Twitter, we select manifest features according to the sociology knowledge and related work. These features are classified and analyzed. Principal manifest features are identified, including some ones which have been rarely used in measuring user influence. Furthermore, we analyze the hidden features, derived from the manifest features. We map the hidden features to the attributes of influencers in the study of social science. Our analysis reveals the most important social attributes that drive user influence in Twitter environment. To the best of our knowledge, our study provides the first comprehensive analysis of the principal features for measuring user influence on Twitter.

2. We propose a hybrid feature selection method for predicting user influence on Twitter. Based on the attributes of influencers defined in sociology, we explore the candidate features from Twitter. After evaluating the quality of features by utilizing a *filter* method, a reduced feature subset is obtained. Following the principles of *wrapper* methods, we assess the feature subset at each searching step. Finally, an optimal feature set is obtained for predicting user influence with a high degree of accuracy. This proposed method provides a solid foundation for studying complicated user influence evaluation and prediction. To the best of our knowledge, this work is the first

one to intensively study the feature selection for evaluating/predicting the online user influence.

3. We study a specific influence maximization problem, i.e., selecting a set of seed users on Twitter to maximize information propagation. Our approach takes into consideration of social ties, user interactions, and information propagation on Twitter. The influence probability is calculated based on users' action history including tweet, favorite, reply and retweet. An information diffusion model is proposed, which inherits the classic independent cascade model and captures the major characteristics of information spread on Twitter. A concise heuristic algorithm is developed for influence maximization accordingly.

4. Considering the characteristics of advertisement propagation on Twitter, we propose an improved diffusion model, which removes some constraints in the classic independent cascade model. A new metric *advertising effectiveness* is defined as the maximization objective. When calculating the influence probability, influence decay is also introduced to reflect the temporal features associated with influence. Experimental results and analysis are provided to show the soundness of the proposed model.

1.4 Thesis Organization

This dissertation is organized in seven chapters. In this chapter, the motivation, research problem and contributions are described. In Chapter 2, we provide the background knowledge and review the related work. Specifically, we introduce the basic concepts related to influence and the existing methods of measuring the influence in online social networks. The state-of-the-art research progress of the Influence Maximization problem is discussed. Chapter 3 is based on my publication [43] in *2015 IEEE International Conference on Big Data Computing Service and Applications*. We provide in-depth analysis on the effectiveness of the principal features for measuring user influence on Twitter. Both key manifest features and important hidden social attributes are analyzed when identifying influential users. Chapter 4 is based on my publication [44] in *2015 International Conference on Web Information Systems Engineering*, which is an extension for Chapter 3. We propose a hybrid feature selection method for predicting user influence on Twitter. The method inherits the advantages of commonly used *filter* and *wrapper* approaches to achieve a high degree of efficiency and accuracy in the optimization. Chapter 5 is based on my publication [45] in *2017 IEEE International Conference on Communications*. We study the influence maximization problem on Twitter. Our approach has taken into consideration of social ties, user interactions, and information propagation on Twitter. Chapter 6 is based on my publication [46] in *2017 IEEE International Congress on Big Data*, which

is an extension for Chapter 5. We utilize the advertising theory from the marketing area and identify a specific influence maximization problem for maximizing the effectiveness of advertising campaigns on Twitter. A new influence probability model and diffusion model are proposed, which can better reflect the real situations of advertising information spread on Twitter. Chapter 7 concludes the dissertation and discusses some future research directions.

2

Background

2.1 Influence in Online Social Networks

2.1.1 Influence Definition

It is difficult to give a precise definition of the term "influence", since this concept is abstract. Generally, influence means "change in a person's cognition, attitude, or behavior, which has its origin in another person or group" [5]. Merriam-Webster dictionary defines influence as "the power to change or affect someone or something; the power to cause changes without directly forcing them to happen" [1].

When influence is studied in OSNs, e.g. Twitter, researchers have given their own explanations. Cha Meeyoung, et al. [47] focus on "an individual's potential to lead others to engage in a certain act". Leavitt Alex, et al. [48] describe influence on Twitter as "the potential of an action of a user to initiate a further action by another user". Rosenman Evan TR [49] interprets the term as "the

[1] http://www.merriam-webster.com/

ability to, through one's own behaviour on Twitter, promote activity and pass information to others". According to the description on Klout [1] (a well-known Influence Scoring System (ISS)), influence is the ability to drive action. When you share something on social media or in real life and people respond, that is influence.

We believe that, influence is a concept which could show in various ways in different contexts. For example, it might refer to passing a message successfully to others in a task of information diffusion. It might mean that audiences agree with the speaker's arguments in a campaign speech. It might imply that customers are persuaded to buy products in a marketing activity. A clear definition / description for influence is crucial when studying a specific research problem. In this work, we give the definition of influence in 5.2, 6.2.

2.1.2 Types of Influencers

In sociology area, there is plenty of research work around social influence. In order to better understand the characteristics of the influencers and develop an effective influencer marketing strategy, influencers are grouped into different types. Malcolm [50] described influential people in the following ways:

- **Connectors** are the people in a community who know large numbers of people and who are in the habit of making introductions. Their ability to

[1] https://klout.com/

connect with people is a function of something intrinsic to their personality, some combination of curiosity, self-confidence, sociability, and energy.

- **Mavens** are information specialists, or people we rely upon to connect us with new information. They are really information brokers, sharing and trading what they know.

- **Salesmen** are persuaders, charismatic people with powerful negotiation skills. They tend to have an indefinable trait that goes beyond what they say, which makes others want to agree with them.

Klout's matrix of influence (Fig. 2.1) describes 16 types of social media influencers, divided into four quadrants:

- **Participating and Sharing:** Feeder, Broadcaster, Syndicator, Curator.

- **Listening and Casual:** Conversationalist, Dabbler, Explorer, Observer.

- **Focused-in-scope and Consistent:** Socializer, Activist, Networker, Specialist.

- **Broad-in-scope and Creating:** Thought leader, Tastemaker, Pundit, Celebrity.

People are classified and labeled based on their behavior and how other people respond to their content. For making contrast, Lisa Barone proposes a simpler list in Small Business Trends [1]: *The Five Types of Influencers On The Web.*

[1] https://smallbiztrends.com

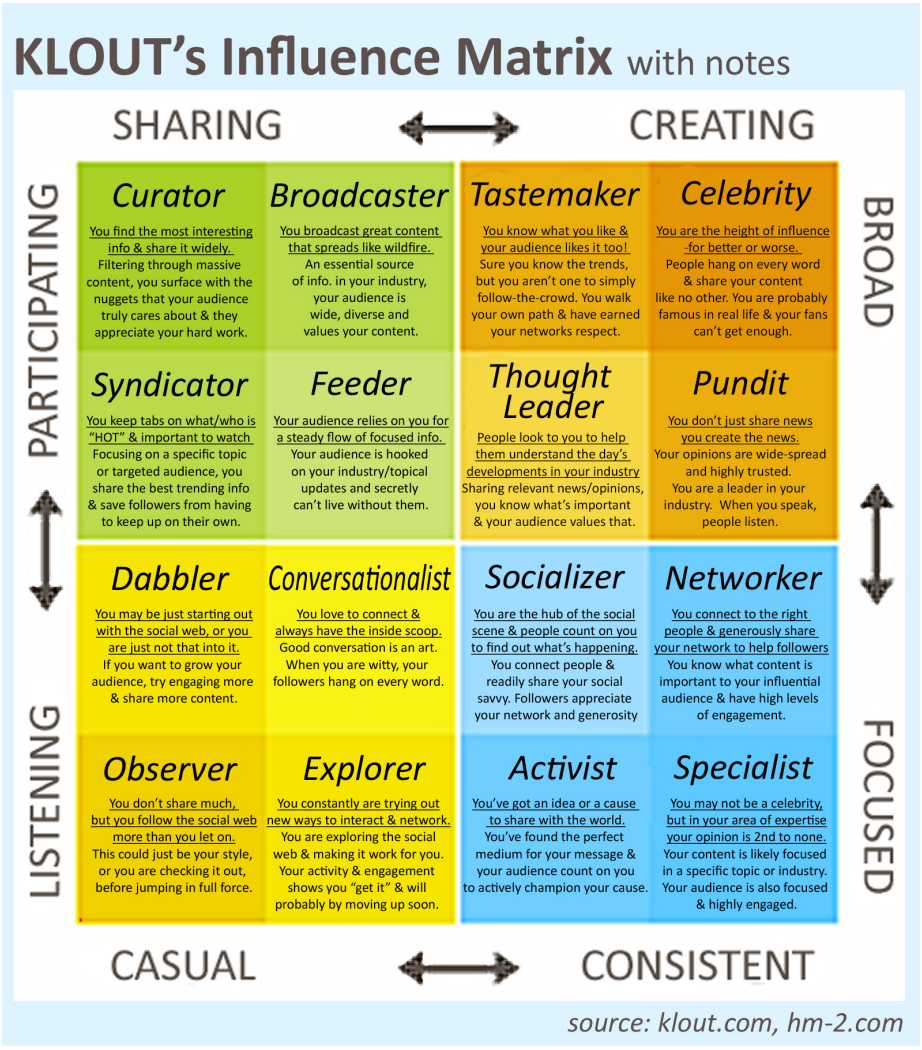

Figure 2.1: Klout Influence Matrix

- **Networkers (Social Butterflies):** The networker has a huge contact list residing across all social platforms. He or she knows everybody and everybody knows him or her.

- **Opinion Leaders (Thought Leaders):** The opinion leader is the best ambassador of a brand. He or she has built a strong authority in his or her field based on credibility. Their messages are most often commented on and retweeted.

- **Discoverers (Trendsetters):** The discoverer is the early adapter of the latest things. Constantly on the lookout for new trends, they are the "hub" in the sector.

- **Sharers (Reporters):** The sharer distributes information to the bloggers or journalists through the specialized webzines. He or she usually amplifies messages.

- **Users (Everyday Customers):** The user represents the regular customer. He or she does not have a network as large as the networker, but his or her network remains equally important.

During influencer marketing, the advertiser will choose certain types of influencers according to the strategies and objectives of the campaign.

Traackr [1] is an influencer management platform. It describes multiple faces of influence built from the 10 most common influencer archetypes (Fig. 2.2). Using the Traackr platform, influencers are vetted across three metrics: reach, resonance, and relevance:

- **Reach** measures the total size of an influencer's audience online across all social platforms.
- **Resonance** measures how engaged an influencer's audience is with their content. Engagement is measured by shares, likes, comments, and retweets. Measuring resonance is essential to ensure brands find influencers with engaged audiences who respond to their content.
- **Relevance** measures how "on topic" the influencer is. Relevance is the most important of the three metrics. Just because an influencer has a million followers does not mean they are relevant to the topics your audience cares about.

2.2 Evaluating and Measuring Influence

2.2.1 Feature Selection

Feature selection, also known as variable selection, is an important problem in the area of machine learning and statistics. The aim of feature selection is to find

[1] http://www.traackr.com

Figure 2.2: The Many Faces of Influence by Traackr

an optimal feature subset, which has significant impact on the target variables (i.e. user influence in this dissertation) from the original feature set, by reducing effects from noise or irrelevant variables. Feature selection provides us a way of reducing computation time, improving prediction performance, and a better understanding of the data in machine learning or pattern recognition applications [51].

Feature selection is different from feature extraction, which is another way of dimensionality reduction. Both approaches aim towards reducing the number of random variables under consideration. Feature extraction is to transform the existing features into a lower dimensional space, while feature selection is to select a subset of the existing features without a transformation (Fig. 2.3). For instance, Principal Component Analysis (PCA) as a feature extraction method, reduces the

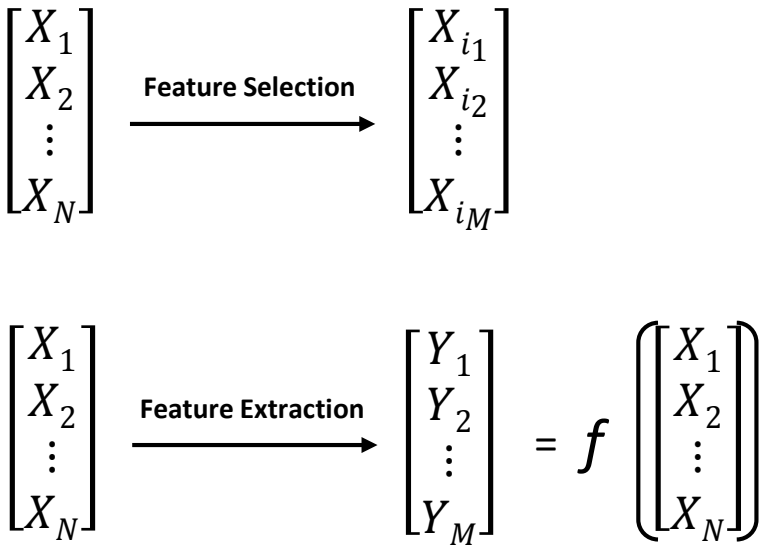

Figure 2.3: Feature Selection vs. Feature Extraction

dimensionality by making new synthetic features from linear combination of the original ones, and then discarding the less important ones. Feature selection does not create new features since it uses the input features itself to reduce their number.

To remove an irrelevant feature, a feature selection criterion is required which can measure the relevance of each input feature with the output value. Once a feature selection criterion is determined, a procedure must be developed to find the optimal subset of the original features. Directly evaluating all the subsets of features has a computational complexity of $O(2^N)$. Hence a suboptimal procedure might be utilized which can remove redundant data within an acceptable period of time.

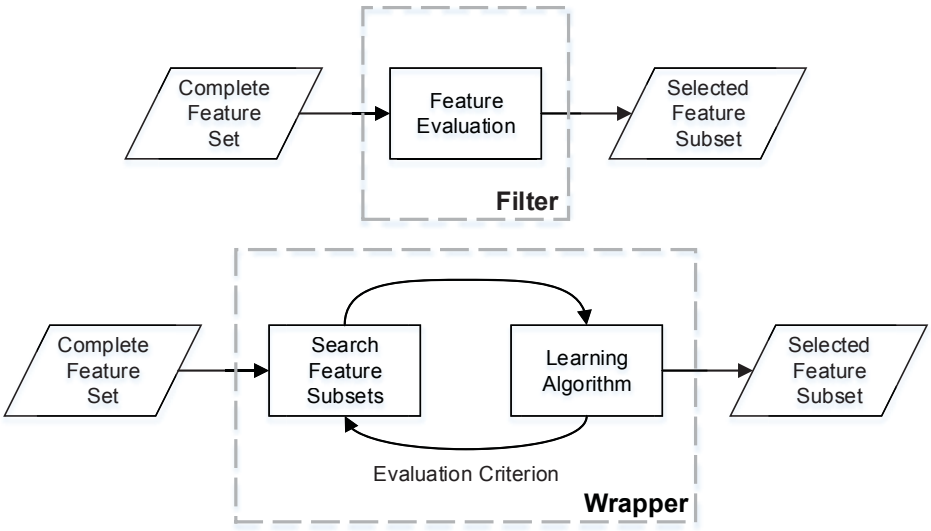

Figure 2.4: The processes of filter and wrapper methods

According to the evaluation criteria, existing feature selection methods can be divided into two main categories: *Filter* methods [52] and *Wrapper* methods [53], which are demonstrated in Fig.2.4.

2.2.1.1 Filter Methods

Filter methods select features based on performance evaluation functions calculated directly from the data. Ranking methods are used to score the features and less relevant features will be removed based on a defined threshold. A number of researchers [52, 54, 55, 56, 57, 58] have presented various definitions and measurements for the feature relevance.

Two well-known filter methods for feature selection are RELIEF [59] and FOCUS [60]. The original RELIEF algorithm estimates the quality of features

according to how well their values distinguish between instances that are near to each other. The FOCUS algorithm conducts an exhaustive search of all possible feature subsets, then selects the smallest feature subset.

Filter methods do not rely on learning algorithms which are biased. The advantage of filter methods is the high computational efficiency. However, the effectiveness is sometimes unsatisfactory since they separate the feature selection from model building.

2.2.1.2 Wrapper Methods

A wrapper method utilizes a predetermined learning algorithm and uses its estimated performance as the evaluation criterion for feature selection. This method employs a search strategy through the space of feature subsets and uses the estimated accuracy from the learning algorithm to measure the goodness of the selected feature subset.

Wrapper methods can be classified into two groups: *Sequential Selection Algorithms* [61, 62] and *Heuristic Search Algorithms* [63, 64, 65]. The sequential selection algorithms start with an empty set (or a full set) and add features (or remove features) until the maximum predictor performance is obtained. The heuristic search algorithms evaluate different subsets to find the optimal feature subset. The subsets to be evaluated are generated by heuristic algorithms, such as Genetic Algorithm (GA) [66, 67] and Particle Swarm Optimization (PSO) [68, 69].

Normally, a wrapper method can provide more accurate result but it has a higher degree of computational complexity in comparison with a filter method.

2.2.2 Modelling Influence

In the past ten years, a lot of research work has emerged on exploring the online social influence evaluation or measurement. Riquelme et al. [70] collected and classified many different influence measures on Twitter.

In this section, we review the related work on ranking user influence on Twitter, which can be summarized into two main categories. The studies in the first category evaluate user influence based on the topological graph of Twitter, whereas those in the second category focus on users' activities. We also discuss several studies on feature selection on Twitter.

In the topological graph of Twitter, nodes represent users, and edges represent *following* relationship. This kind of graph has been widely used to analyze influential users on Twitter, since the topological relationship within a graph has successfully demonstrated great value for ranking influential web pages in Google's *PageRank* algorithm [71]. Kwak et al. [72] ranked Twitter users by the number of followers and by *PageRank*, and found that the two rankings were similar. TwitterRank [73], which is an extension of *PageRank* algorithm, measures user influence taking both the topical similarity and link structure into account. TunkRank [74] is another adaption of *PageRank*, introducing the probability of retweeting and the distribution of attentions. Similarly, other *PageRank*-based

algorithms have been proposed as well, such as KHYRank [75] and InfluenceRank [76].

In addition, Brown et al. [77] proposed a modified k-shell decomposition algorithm for computing user influence on Twitter. The input to this algorithm is the connection graph between users. User influence is measured by the k-shell level, which is the output of the k-shell decomposition algorithm. However, we believe this kind of approach has its limitations. Recent studies have pointed out that top influencers on Twitter show a strong correlation with *retweets* and *mentions* rather than *followers* [47, 72]. Therefore, the count of *followers* alone does not actually reflect individual's influence. Without the consideration of users' interactions, it is difficult to measure user influence on Twitter.

In recent years, researchers started to focus more on other users' activity when measuring user influence on Twitter. Cha et al. [47] defined three types of influence: *Indegree influence* (i.e. the number of followers), *Retweet influence* and *Mention influence*. They compared the three types of influence ranking for 6 million users, and found that the top users showed a strong correlation with the *Retweet influence* and *Mention influence*, however, not so much related to the *Indegree influence*. IARank [78] is a model to continuously rank influential Twitter users in real-time, based on a concept of "information amplification". The information amplification is characterized by three activities: event activity, attention obtained and social connectivity. Leavitt Alex, et al. [48] categorized users'

actions from the perspectives of conversation and content, mapping to *replies* and *retweets* respectively. This study analyzed how these actions represented the influence of a user. Romero et al. [79] developed an IP (Influence-Passivity) algorithm to address the observation that the majority of users on Twitter act as passive information consumers and do not forward the content to the network. The IP algorithm interactively estimates the influence and passivity of users based on their information forwarding activity. We also consider users' actions as the essential metrics for measuring user influence. However, along with *reply* and *retweet*, there are a wider range of others' actions reflecting user influence, such as *follow*, *mention*, *create public lists*, etc.

Although there have been many studies on quantifying user influence on Twitter, no matter whether it is complex or simple, there is still a lack of a comprehensive study of analyzing the key features for measuring user influence. Luiten et al. [80] investigated the relations between certain features and so-called topical influence on Twitter. However, only four manifest features were taken into consideration (*followers*, *friends*, *mentions* and *retweets*). Similarly, the study in [81] analyzed the principle features for tie strength estimation in Sina Micro-blog network. Although this study investigated the pairwise tie strength in a different platform, the idea of the study inspires us to make comprehensive and in-depth analysis of the principal features for measuring user influence on Twitter.

On the other hand, the study by Wu et al. [82] utilized Twitter "lists" to

classify users to "ordinary" users and four types of "elite" users. As this study claimed, it paid more attention to how the information flowed among different categories of users, and how the information originating from traditional media sources reached the masses. Compared with these aforementioned studies, our paper makes use of a wide range of manifest features, focusing on the principal features for measuring user influence. As a result of analyzing these principal features, our work reveals certain novel findings which have not been discussed in the existing literature.

In recent years, some influence scoring services have emerged on the Internet, which developed their own models to measure online user influence. For example, Klout is a social media tool that measures your online influence by evaluating your activity on a variety of social media sites like Facebook, Twitter, Google Plus, LinkedIn, Foursquare, YouTube, and others. The Klout score is a number between 1-100 that represents your influence. The more influential you are, the higher your Klout score. The algorithm uses more than 400 signals from eight different networks to update the Klout score every day [1].

According to the philosophy of Klout, influence is the ability to drive action. It is great to have lots of connections, but what really matters is how people engage with the content you create. It is better to have a small and engaged audience than a large network that does not respond to your content. Being active

[1] https://klout.com/corp/score

is different than being influential. Posting a thousand times and getting zero responses is not as influential as posting once and getting a thousand responses. It is not about how much someone talks, but about how many people listen and respond.

Traackr is an influencer marketing platform that helps you get results with social media marketing by finding the right influencers and opportunities. Traackr uses three metrics for measuring online influence: *Reach*, *Resonance* and *Relevance*. Reach is a person's audience size. Resonance measures how engaging their content is and Relevance tells you how closely their content matches your topic.

2.3 Influence Maximization Approaches

Motivated by the marketing applications, the problem of *Influence Maximization* was firstly proposed by [28], and has attracted a lot of interest in the research field of online social networks [29, 32, 41, 83, 84, 85, 86, 87]. The influence maximization problem can be described as follows. A social network is represented by a directed graph $G = (V, E)$, where the nodes V represent users, and the directed edges E represent social ties between users. We are also given a budget k, which is a integer. The goal of influence maximization is to find k users (seed nodes) in the social network so that the spread of influence (defined as the excepted number of influenced users) could be maximized.

2.3.1 Information Diffusion Models

Information diffusion refers to the process that information spreads out among users in a social network as time goes on. Granovetter [88] developed a formal mathematical model for the diffusion process. Afterwards, various diffusion models have been proposed by researchers. *Linear Threshold Model* (LT) and *Independent Cascade Model* (IC) [28] are two well-known fundamental ones in studying the social influence problems. There are quite a few variations and extensions based on these two basic models, such as *Majority Threshold Model* [89], *Small Threshold Model* [89], *Decreasing Cascade Model* [90]. Some more diffusion models have been developed when studying particular research problems or in practical application scenarios. Chen et al. [34] proposed *Independent Cascading Model with Negative Opinion (IC-N)* that incorporates the emergence and propagation of negative opinions. Lu and Lakshmanan [35] extended the classical LT model by incorporating prices and valuations to capture monetary aspects in product adoption. He et al.[91] developed a heterogeneous network based epidemic model to describe the propagation dynamics of rumors in online social networks. Different diffusion models should be applied in different situations.

A social network is a social structure made up of a set of social actors (such as individuals / companies / organizations), and a set of complex social relations. Formally, a social network can be represented as a graph $G = (V, E)$, which can be either directed or undirected according to the specific characteristics of a real

social network. In the graph G, each vertex $v \in V$ represents an individual user. In a directed graph, an edge $(u, v) \in E$ represents the social connection is from u to v, not vice versa. In an undirected graph, an edge (u, v) represents the mutual connection between u and v. Particularly, an undirected graph can be viewed as a directed graph by considering each edge as a bidirectional edge with no distinction on both direction. For example, Twitter and Google+ are directed networks, while Facebook is an undirected network.

2.3.1.1 Linear Threshold Model

Linear Threshold (LT) model has been extensively discussed in studying diffusion models. In this model, each edge $(u, v) \in E$ is associated with a non-negative influence weight $w_{u,v}$, and for all $v \in V$, the sum of incoming weights is no more than 1.

$$\sum_{u \in N_{in}(v)} w_{u,v} \leq 1, \tag{2.1}$$

where $N_{in}(v)$ denotes the set of in-neighbours of v. Each node v has a threshold $\theta_v \in [0, 1]$, which represents the minimum total influence weight that are needed to activate v. Influence diffusion proceeds in discrete time steps. At the beginning (time step $t = 0$), a seed set S is activated. At any time step $t \geq 1$, an inactive node v becomes active if the total influence weight from its active

in-neighbours reaches or exceeds θ_v.

$$\sum_{u \in N_{in}^{a}(v)} w_{u,v} \geq \theta_v, \tag{2.2}$$

where $N_{in}^{a}(v)$ denotes the set of active in-neighbours of v. Every activated node remains active, and the diffusion process terminates if no more nodes can be activated.

Fig. 2.5 shows an example of influence diffusion in LT model. The number in the circle is the threshold for this node. The influence weight is labelled on the edge. At time step $t = 0$, a seed node s is activated. At time step $t = 1$, node s successfully influences two of its neighbours because the influence weights are greater than the thresholds. These newly activated nodes try to influence their own neighbours at the next step ($t = 2$), and two more nodes are activated. The influence propagates in such a way and this process stops at time step $t = 3$, after which no more activation is possible.

2.3.1.2 Independent Cascade Model

Independent Cascade (IC) Model is a dynamic cascade model, which originates from probability theory and was firstly studied by Goldenberg et al. [92] in the context of marketing. In this model, when a node u first becomes active, it has only a single chance to influence its inactive neighbour v, with a probability $p_{u,v}$, which is independent of the diffusion history. Node u can not try to influence v again whether the first attempt succeeds or not. If u succeeds in activating v at

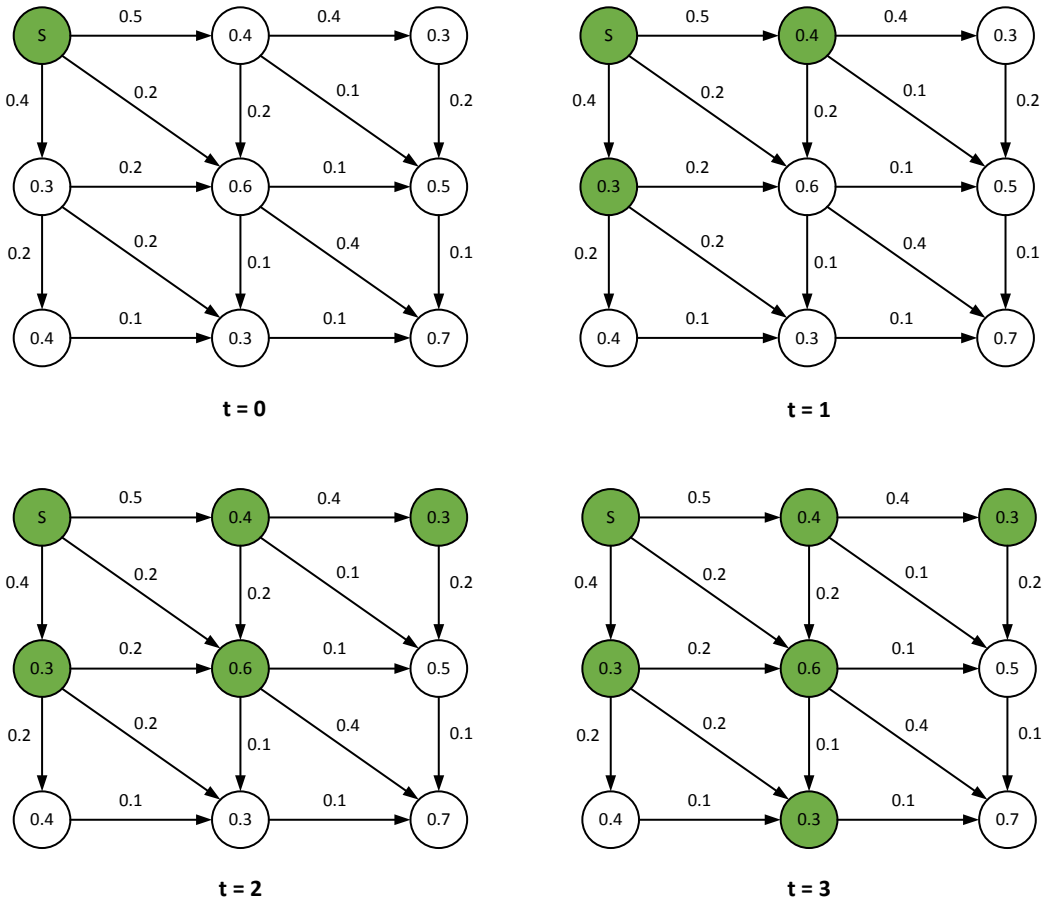

Figure 2.5: Linear Threshold Model

step t, then v can make one attempt to influence its inactive neighbours at step $t+1$. If an inactive node has more than one newly activated in-neighbour, these active nodes will make one attempt to influence the inactive node independently. This inactive node will switch to active status if one of its neighbour succeeds. The diffusion process stops until every active node has tried its single chance and there are no more activations. Since influence is propagated based on probability, each running of the simulation process will obtain a different result. Therefore, in order to obtain an accurate estimation, we need run a large number of simulations.

Fig. 2.6 shows an example of influence diffusion in IC model. This example is a running instance, and the result is only one of the possibilities. In the graph, a green edge means a successful attempt, and a red edge means an unsuccessful attempt. At time step $t = 1$, the seed node s succeeds in activating A and C but fails on D. At time step $t = 2$, the newly activated nodes A and C successfully activate D and F. At the last time step, node H is activated by D. During this round of simulation, the seed node influences 5 nodes in the graph.

2.3.1.3 Extension Models

Besides the two well-known models mentioned above, researchers have also proposed many variation and extension models to reflect the more complicated real-world situations. Some examples are given as follows.

Liu et al. [36] proposed the time constrained influence maximization problem, which is based on the *Latency Aware Independent Cascade (LAIC)* diffusion

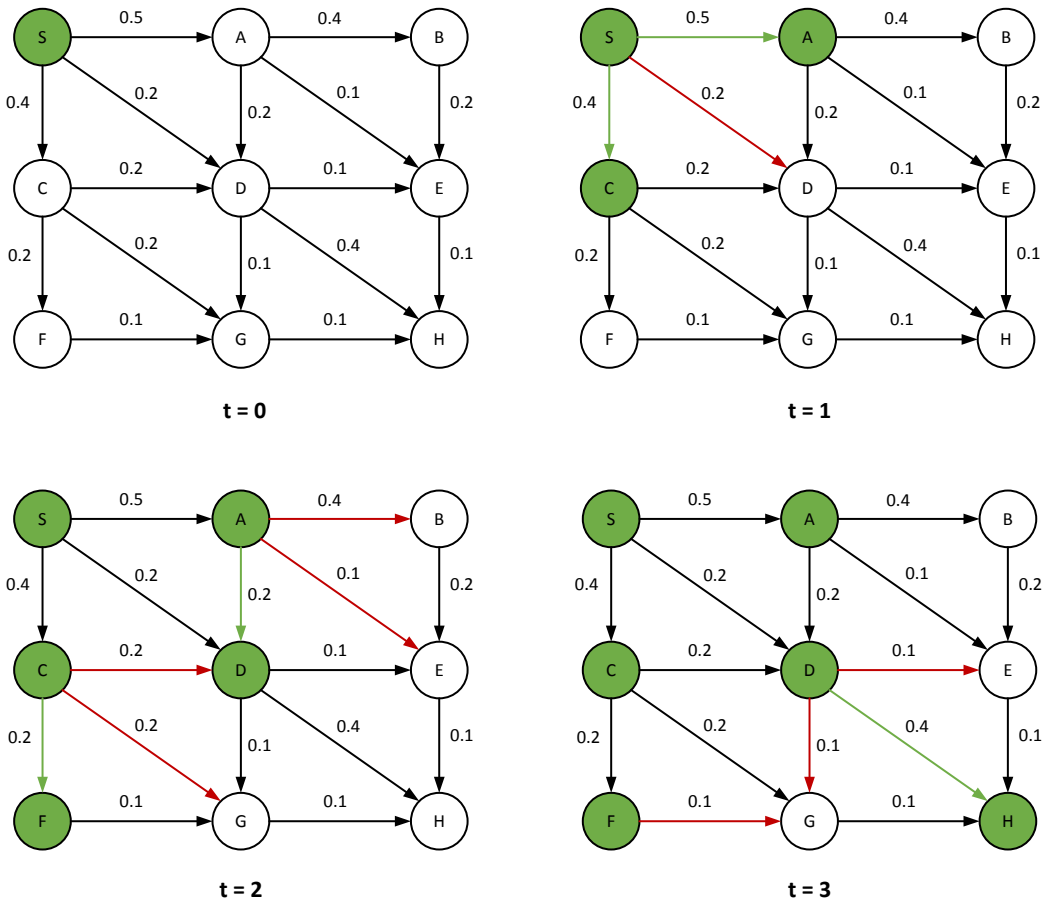

Figure 2.6: Independent Cascade Model

model. They considered time factor in the influence propagation process from two aspects: 1) a time bound is given, which means only the influence spread before a fixed time is calculated; 2) influence delay is incorporated which follows some distribution. These assumptions can reflect some real-world situations to a certain extent, but there is no further analysis on how influence delay occurs.

Chen et al. [37] added a deadline constraint in the influence maximization problem, and they proposed two new diffusion models, the *Independent Cascade model with Meeting events (IC-M)* and the *Linear Threshold model with Meeting events (LT-M)* to capture the time delay of influence diffusion between users. However, they didn't clearly explain in which application scenarios the proposed models should be applied and it lacked justification in the parameter setting for influence probabilities and meeting probabilities.

Lu et al. [40] proposed the *Comparative Independent Cascade (Com-IC)* model that allows any degree of competition or complementarity between two different propagating items. In Com-IC model, users' adoption decisions depend not only on edge-level information propagation, but also on a node-level automaton whose behavior is governed by a set of model parameters. However, how to simplify the model and make it tractable when extending the model to multiple items, and how to reason about the complicated two-way or multi-way competition and complementarity, still remain as interesting challenges.

Tong et al. [41] developed the *Dynamic Independent Cascade (DIC)* model

which is able to capture the dynamic aspects of real social networks. In the classic IC model a seed node is guaranteed to be activated after selected and the relationship between two users is simply represented by a fixed probability, while the seed nodes in DIC model could fail to be activated with a certain probability and the propagation probability between two users follows a certain distribution which reflects the change of topology of a social network.

Zarezade et al. [93] proposed a social *behavior adoption* model in which multiple correlated cascades spread over the network. Multidimensional Hawkes process is utilized for the behavior or product adoption with its marks capturing the decision making procedure of the users. The advantage of the proposed model is twofold; it models correlated cascades and also learns the latent diffusion network.

2.3.2 Seed Selection Algorithms

After the diffusion model is determined, the next step is to select a set of seed users who can spread the influence in the network as much as possible. Kempe et al. [28] prove the influence maximization problem is *NP-hard* for both LT model and IC model. In recent years, researchers in this field have conducted extensive studies on various algorithms for the influence maximization problem. Generally, there are two metrics to measure a seed selection algorithm, *effectiveness* and *efficiency*. Effectiveness means that the selected set of seed users can reach the maximum coverage in the social network. More users can be influenced by the seed users, more effective the algorithm is. Efficiency is also important for a seed

selection algorithm. A good algorithm should be computationally acceptable and be able to scale up to large real world social networks. If a seed selection algorithm takes quite a long time or can only handle a small-sized dataset, it is useless in practical applications. There is a trade-off between effectiveness and efficiency, which researchers need to consider in the particular problem.

2.3.2.1 Greedy Algorithm

Kempe et al. [28] prove the influence function $f(\cdot)$ is monotone and submodular. They propose a greedy hill-climbing algorithm, which guarantees to achieve an approximation solution within a factor $(1 - 1/e - \varepsilon)$ to the optimum in both the LT model and the IC model. Here e is the base of the natural logarithm and ε is any positive real number. Thus, this is a performance guarantee slightly better than 63%.

The idea of the greedy algorithm is quite straightforward. Suppose the objective is to choose a seed set of size k, then there will be k rounds of selection. In each round, a new user that gives the largest marginal gain in influence spread will be selected as a seed. The experiments in [28] show that the greedy algorithm significantly outperforms the classic high-degree and distance-centrality heuristic algorithms in terms of influence spread.

2.3.2.2 CELF and CELF++

Some recent research work has been done to tackle the efficiency issue of the greedy algorithm. One of the notable work is [24], in which the property of "submodularity" is exploited to develop an efficient algorithm called Cost-Effective Lazy Forward (CELF) selection algorithm, achieving near optimal solution, while being 700 times faster than a simple greedy algorithm. The idea is that the marginal gain of a node in the current step cannot be better than its marginal gain if selecting this node in the previous steps. Therefore, this optimization avoids the re-computation of marginal gains for all the nodes in any step, except in the first step.

To further reduce the running time, Goyal et al. [94] propose an extension of CELF, called CELF++, which is 35-55% faster than CELF. CELF++ further optimizes CELF by exploiting the property of submodularity of the spread function to avoid unnecessary re-computations of marginal gains incurred by CELF. Their empirical studies on real world social network datasets show that CELF++ works effectively and efficiently, resulting in significant improvements in terms of both running time and the average number of node look-ups.

2.3.2.3 SPM and SP1M

In [95, 96], Kimura et al. propose two shortest path based models *SPM* and *SP1M*, which are special cases of the IC (Independent Cascade) model. For two nodes u and v, SPM only considers the influence that flows via the shortest path

from u to v. SP1M, instead considers the top-2 shortest paths from u to v. The idea is that the majority of influence flows through shortest paths.

The experimental results show that SP1M performs better than SPM. Moreover, the authors experimentally demonstrate that when the propagation probabilities through links are small, the proposed methods can give good approximations to the IC model for finding sets of influential nodes in a social network. However, a critical issue with this approach is that it ignores the influence probabilities between users. It is not very convincing if only considering the shortest paths.

2.3.2.4 Maximum Influence Paths

Based on the above contribution in SPM and SP1M, Chen et al. [32] extend this idea by considering *Maximum Influence Paths* (MIP) instead of shortest paths. A maximum influence path between a pair of nodes (u, v) is the path with the maximum propagation probability from u to v. The main idea of this heuristic is to use local arborescence structures of each node to approximate the influence propagation.

The maximum influence paths (MIPs) are computed via the Dijkstra shortest-path algorithm, and then MIPs with probability smaller than an influence threshold θ are ignored, which effectively restricts influence to a local region. The MIPs starting or ending at each node form the arborescence structures, which represent the local influence regions of each node. Only influence propagated through these

local arborescences is taken into consideration, and this model is referred to as the *Maximum Influence Arborescence* (MIA) model.

The authors later propose an extension model *Prefix excluding* MIA (PMIA). Their experimental results show that the algorithm is scalable to million-sized graphs where the greedy algorithm becomes infeasible, and in all size ranges, the algorithm performs consistently well in influence spread.

2.3.2.5 SIMPATH

Goyal et al. [97] propose SIMPATH, an efficient and effective algorithm for influence maximization under the Linear Threshold model. SIMPATH algorithm builds on the CELF optimization that iteratively selects seeds in a lazy forward manner. However, instead of using expensive Monte Carlo simulations to estimate the spread, it is shown in [97] that under the LT model, the spread can be computed by enumerating the simple paths starting from the seed nodes.

SIMPATH algorithm leverages two optimizations. Firstly, the *VERTEX COVER OPTIMIZATION* reduces the spread estimation calls in the first iteration, thus addressing a key limitation of CELF. Secondly, the *LOOK AHEAD OPTIMIZATION* improves the efficiency in subsequent iterations. Their experiments on four real data sets show that SIMPATH outperforms the state of the art, in terms of running time, memory consumption and the quality of the seed sets.

2.3.2.6 Other Heuristic Algorithms

In order to improve the efficiency, heuristic algorithms are applied in the selection of seed nodes. Compared with greedy algorithms, heuristic algorithms may not provide the best result, but they are able to obtain an acceptable result in much less time. Two widely used heuristic algorithms are *degree centrality* and *closeness centrality*.

In degree centrality algorithm, the nodes with a large number of connections are considered to be influential in the social network. Take Twitter as an example, seed nodes are selected based on the number of followers (in descending order). This is based on a simple and intuitive assumption: the more followers a user has, the more influential the user tends to be. Experimental results provided in [28] show that this heuristic algorithm can achieve the influence spread close to the greedy algorithm, and outperforms several other algorithms.

In closeness centrality algorithm, the average distance from each node to all other nodes in the social network is computed. The nodes that have a smaller average distance to other nodes are considered at more central position in the network and these nodes are selected as seed nodes.

In general, heuristic algorithms have not been studied extensively in the research field due to the low expectation of quality. In addition, the ideas of heuristic algorithms rely on the definition of specific research problems, and the performance is closely related to the diffusion model.

2.4 Twitter at a Glance

Twitter is one of the most popular research platforms when researchers studied the influence maximization problem, because of its asymmetric following relationship and open APIs. In this work, we also consider Twitter as our research platform.

Twitter was created in March 2006, and nowadays it has become one of the largest online social networks. Twitter is the place to find out about what is happening in the world right now, whether you are interested in music, sports, politics, news, celebrities, or everyday moments. As of June 2016, Twitter had more than 313 million monthly active users.

Users can subscribe to other users' tweets, which is known as "follow" and subscribers are known as "followers". Users can join conversations by replying to others and by mentioning others in their own tweets. Individual tweets can be forwarded by other users to their own feed, which is known as "retweet". Users can also "like" (formerly "favorite") individual tweets.

2.4.1 Network Structure

The Twitter network can be described by a graph, which contains the collection of nodes (representing users) and edges (representing following relationships). Whereas some social websites like Facebook and LinkedIn require the mutual acceptance of a connection between users, Twitter's relationship model allows you

to keep up with the latest happenings of any other user, even though that other user may not choose to follow you back or even know that you exist. That means following on Twitter is not mutual. Someone who thinks you are interesting can follow you, but you do not have to follow them back. Researchers have found that only about 22% of Twitter relationships are mutual [72]. The asymmetric following model makes Twitter unique, and many online social networks that appear after Twitter employ this model, such as Sina Weibo [1], which is one of the most popular websites in China.

2.4.2 User and Tweet Objects

Users can be anyone or anything. They tweet, follow, create lists, have a home timeline, can be mentioned, and can be looked up. From the perspective of data storage, user objects are structured data including many fields. Fig. 2.7 is a user object sample in JSON format (only part of the fields are displayed).

A Tweet is any message posted to Twitter which may contain photos, videos, links and up to 140 characters of text. Tweets are the basic atomic building block of all things on Twitter. Tweets are also known as "status updates". Fig. 2.8 is a tweet object sample in JSON format (only part of the fields are displayed).

In this work, we analyze the information from user and tweet objects. Considering both user profile and user behaviors (including actions and interactions),

[1] http://weibo.com/

```
{
        "name": "Twitter API",
        "location": "San Francisco, CA",
        "created_at": "Wed May 23 06:01:13 +0000 2007",
        "id_str": "6253282",
        "favourites_count": 24,
        "id": 6253282,
        "listed_count": 10713,
        "lang": "en",
        "followers_count": 1198334,
        "protected": false,
        "geo_enabled": true,
        "description": "The Real Twitter API.",
        "verified": true,
        "time_zone": "Pacific Time (US & Canada)",
        "statuses_count": 3331,
        "friends_count": 31,
        "screen_name": "twitterapi",
}
```


Figure 2.7: A user object sample in JSON format

we study how to select features for measuring user influence and how to solve the influence maximization problem.

2.4.3 Twitter APIs

Twitter is recognized for having one of the most open and powerful developer APIs of any major technology company. Developer interest in Twitter began immediately following its launch, prompting the company to release the first version of its public API in September 2006. Twitter allows developers to interact with its data i.e. tweets and several attributes about tweets using Twitter APIs. Developers need to know a server side scripting language like php, python or ruby to make requests to Twitter APIs and results would be in JSON format that can be easily read by the program.

```
{
	"created_at": "Mon Jan 12 03:31:45 +0000 2015",
	"id": 554480825480142850,
	"id_str": "554480825480142848",
	"text": "It is very warm today.",
	"source": "<a href=\"http://twitter.com\" rel=\"nofollow\">Twitter Web Client</a>",
	"in_reply_to_status_id": null,
	"in_reply_to_status_id_str": null,
	"in_reply_to_user_id": null,
	"in_reply_to_user_id_str": null,
	"in_reply_to_screen_name": null,
	"user": {
		"id": 2471109944,
		"id_str": "2471109944",
		"name": "suet",
		"screen_name": "sirpatammilehto",
		"location": "melb. australia",
		"url": null,
		"description": null,
		"protected": false,
		"verified": false,
		"followers_count": 4,
		"friends_count": 26,
		"listed_count": 0,
		"statuses_count": 117,
		"created_at": "Wed Apr 30 14:55:08 +0000 2014",
		"time_zone": "Melbourne",
		"geo_enabled": false,
		"lang": "en",
	},
	"retweet_count": 0,
	"favorite_count": 0,
	"favorited": false,
	"retweeted": false,
	"timestamp_ms": "1421033505207"
}
```


Figure 2.8: A tweet object sample in JSON format

There are two main types of Twitter APIs. The Streaming APIs [1] give access to (usually a sample of) all the latest tweets matching a search query on Twitter. On average, about 6,000 tweets per second are posted on Twitter and developers will get a small proportion (<= 1%). The other type is called REST APIs [2], which is more suitable for singular searches, such as searching historic tweets, reading user profile information, or posting Tweets. The Streaming APIs only sends out real-time tweets, while the Search API (one of the popular REST APIs) gives historical tweets up to about a week with a max of a couple of hundreds. In addition to these two types of APIs, the Webhooks APIs provide realtime access to account data, and the Twitter Ads API allows partners to integrate with the Twitter advertising platform in their own advertising solutions.

In this work, we utilize the REST APIs and the Streaming APIs to collect user and tweet data from Twitter. All the experiments are conducted based on real Twitter datasets.

[1] https://dev.twitter.com/streaming/overview
[2] https://dev.twitter.com/rest/public

3

Principal Features Analysis for Measuring User Influence

3.1 Introduction

In recent years, both academy and industry have shown great enthusiasm on the study of user influence on Twitter. In particular, many researchers speculate that the measurement of user influence on Twitter is similar to the situation of ranking influential web pages. Owing to this, a number of the studies used *PageRank* algorithm [71] or its variants to measure user influence on Twitter [73, 74, 75, 76]. In recent years, researchers started to utilize users' activities on Twitter to measure user influence [47, 48, 78]. Meanwhile, several influence ranking services on the Internet became available based on or incorporating the information from Twitter, such as Klout [1], Kred [2], PeerIndex [3], Followerwonk [4],

[1] http://klout.com
[2] http://kred.com
[3] http://peerindex.com
[4] http://followerwonk.com

etc.

Nevertheless, there are some controversial discussions on which features should be selected for measuring user influence. Intuitively, a user would be influential if he/she has a large number of followers, so a user's *followers* count is generally considered as an important feature for influence measurement [48]. However, estimating influence by only *followers* may lead to misunderstanding. For example, it is quite possible that many followers of a user may be faked accounts or even spammers. Also, against the *followers*, the study by Cha et al. [47] claimed that the top influencers showed a stronger correlation with *retweets* and *mentions* than *followers*. Furthermore, the existing popular influence ranking services are basically black-boxes, and no one knows which features are used for measurement. Meanwhile most studies in the area are working with the manifest features, whereas they hardly analyze the relationship between these features and the social attributes widely discussed in social science.

As such, there is a lack of comprehensive analysis of the principal features for measuring user influence on Twitter. In this chapter, we aim to find the commonly recognized major features for measuring user influence in order to address the issue of feature selection. In the meantime, we analyze the correlation between the principal features and the rankings made by popular influence scoring services, which helps users understand the different preferences and priorities adopted in these particular services. In addition, we identify certain social attributes

associated with these principal features so that we can provide researchers with a sound theoretical support for their study on user influence in the future.

The main contributions of this chapter can be summarized as follows:

- We employ *Entropy* method and Spearman's *Rank Correlation Analysis* to identify the major manifest features for measuring user influence on Twitter. Furthermore, we use *Principal Component Analysis* and *Stepwise Multiple Linear Regression* to investigate the hidden social attributes for influential users on Twitter.

- Our study reveals that, besides the direct features such as *retweets* or *mentions*, a combination of some other features is also fairly effective to predict user influence, e.g., *the number of public lists*, *new tweets*, *ratio of followers to friends*, etc.

- Our mapping from principal hidden features to social attributes demonstrates that *popularity*, *engagement* and *authority* are three most important social attributes to drive user influence on Twitter.

- Despite unawareness of the hidden algorithms of some popular influence ranking services, our study reveals that *new mentions* and *number of public lists* are the two most effective features reflecting the ranking results, and *popularity* is commonly considered as the most important social attribute by all these services.

The remaining of this chapter is organized as follows. Section 3.2 describes our work on finding the principal manifest features for measuring user influence on Twitter. In Section 3.3, we analyze the relationships between the principal features and the rankings made by online influence services. In Section 3.4, we identify hidden social attributes for the influencers on Twitter. Finally, in Section 3.5, we conclude this chapter with a brief discussion of future work.

3.2 Finding Principal Manifest Features

Our goal is to find and analyze the principal features for measuring user influence on Twitter. In this section, we present our work on finding major manifest features. Manifest features are users' explicit attributes or statistical data, which can be explicitly defined and directly retrieved from Twitter environment. For example, the number of *followers*, the number of *retweets*, etc. In the first two subsections, we give the details of features selection and data retrieving. Then, we discuss the Correlation analysis and Entropy method on the manifest features.

3.2.1 Determining Candidate Features

As we mentioned before, user influence is the ability to drive other users' actions. When a user appears or posts a tweet on Twitter, and other people respond, these responses demonstrate the user's influence. There are several kinds of actions or responses on Twitter. For example, *follow*, *read*, *reply*, *mention*, and *retweet*. These actions, if available, are certainly good indicators for user influence. The

more actions are driven by a user, the more influential power the user has. Meanwhile, we believe that users' attributes should also be taken into consideration when estimating the influence on Twitter, e.g., the number of tweets, since it demonstrates how involved a user is in the community to some extent. However, in terms of the effectiveness of these features, we should conduct in-depth analysis.

Furthermore, influence always changes over time. It is often seen that a user's influence suddenly increases due to some social event, or gradually drops as a result of low level of engagement. We believe that user influence is not only an accumulative effect of a user's activities, but also a real-time status that reflects the dynamics of the user's behavior in recent period of time. If a user had a large number of *followers*, *mentions* and *retweets* some time ago, but became inactive and obtained much less attention from others recently, then the user's influence on Twitter at this moment will be considered significantly decreased.

Based on the above discussion, we select eleven (11) candidate features. These features are also called as manifest features as all of them are explicitly available through Twitter APIs. These features are listed as follows (alphabetic order):

1. **Actions to tweets ratio (AT ratio):** the ratio of the number of actions (*retweets* and *mentions*) to that of tweets. This ratio is proposed by Leavitt et al.[48]. We select this metric because the relationship between the subsequent actions and the original tweets certainly reflects a user's influential

power. For instance, user A and user B receive the same amounts of actions(*retweets* and *mentions*), however, user A posts 10 times more tweets than user B. This means that user B obtains the same reactions as user A by far less efforts. Thus, user B seems more influential than user A.

2. **Age of Twitter account:** the number of months since the account was created. Generally speaking, a user having a long time account should have a wider social network than a user with a new account, therefore, the user having a long time account is likely more influential.

3. **Followers to friends ratio (FF ratio):** the ratio of a user's followers to the user's friends. The higher FF ratio is, the more people are interested in the user's status updates without the user showing interest in return. If FF ratio approaches 1, it is likely that the user follows back a majority of his/her followers. If FF ratio approaches 0, we can consider this user as a spammer or bot.

4. **New followers:** the number of new followers during a period of time. Whether a user's followers is increasing or decreasing is a good indicator for the user's influence.

5. **New mentions:** the number of new *mentions* or replies to the user during a period of time, i.e., the number of tweets including "@username" (excluding

"RT @username"). *Mention* represents the name value of a user. This feature is widely used to measure user influence in many studies [47, 48, 75].

6. **New retweets:** the number of new *retweets* of the tweets created by a user during a period of time, i.e., the number of tweets including "RT@username". Many researchers believe *retweet* is one of the most effective metrics for influence calculation[47, 48, 49, 75]. The action of *retweet* signals the retweeter has been influenced by the original author, no matter positive or negative impact.

7. **New tweets:** the number of tweets newly issued by a user during a period of time. We understand that influence is time sensitive, since one's influence in social network is changing over time. If an influential user does not post any tweets for a period of time, his/her influence probably starts declining.

8. **Number of followers:** the total number of followers this account currently has. In general, the more followers a user has, the more influential he appears to be. Although the study by Cha et al. [47] questioned that users with a large number of followers might be not necessarily influential users in terms of spawning *retweets* or *mentions*, we believe number of followers is an important feature for user influence.

9. **Number of public lists:** the number of public lists which a user is a member of. Being included in public lists indicates a user is visible and

people show interest in the user. Wu et al. [82] utilized lists to classify users, and they believed the number of lists on which a user appeared demonstrated his/her importance to the community.

10. **Number of tweets:** the total number of tweets posted by the user. This simply indicates how productive a user is. The study by Keller and Berry [98] proposed five attributes of influencers, one of which is "Activist". A user posting a large volume of tweets indicates the user's high level of engagement in communities.

11. **Verified:** whether a user's account is verified or not. Verification is currently used to establish authenticity of identities of key individuals and brands on Twitter [1]. These users are in various domains such as music, fashion, government, religion, media, sports, business, etc. Generally speaking, a user with a verified account is an influential user.

As a preliminary analysis of the natures of the above features, we categorize them into two dimensions: source (*Personal Profile* vs *Others' Action*), and time (*Accumulative* vs *Instantaneous*). As shown in Fig. 3.1, all features can be put in one of the four quadrants. For instance, in terms of the first dimension, *Age of twitter account* and *Number of tweets* belong to personal profile, while *Number of followers* and *Number of retweets* belong to others' action. In the dimension of Accumulative vs Instantaneous, for example, *Number of tweets* and *Number*

[1] https://support.twitter.com/articles/119135-faqs-about-verified-accounts#

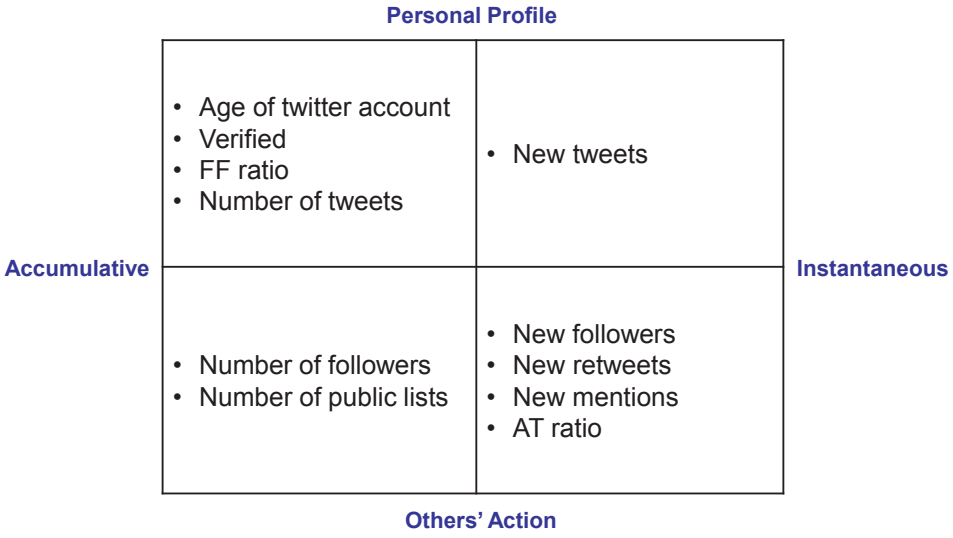

Figure 3.1: The natures of manifest features

of public lists are accumulative features, but *New tweets* and *New followers* are instantaneous ones, as the latter two features reflect the situation during a short period of time.

3.2.2 Retrieving Twitter Data

As discussed in the previous subsection, the candidate features can be categorized into two types: accumulative or instantaneous in terms of their temporal nature. Specifically, accumulative features are those which have status values at this moment, but can be accumulated for a long time since the creation of a user's account. On the other hand, instantaneous features reflect users' actions or responses occurring within a short period of time in the past. Different types of features are retrieved or calculated from different Twitter APIs.

We select top 100 most-followed Twitter users in Australia, who explicitly

indicate "Australia" in their location profiles. All the tweets posted by these users and all the responses to their tweets of these users are collected. For the purpose of capturing accumulative features, we also retrieve profile data of these users, such as account created time, number of followers, etc.

In order to analyze the features during different lengths of time period, we fetch the instantaneous features in one month and one week periods respectively. For details, the one month period is selected from April 4 to May 3 of 2014; and the one week period is between April 26 and May 3 of 2014. There are totally 12,537,115 tweets captured in the month, and 3,498,620 tweets in the week.

All the experimental data are retrieved via Twitter APIs. We capture accumulative features (such as number of public lists) by Twitter REST API [1], and retrieve all the tweets including *mentions* and *retweets* by Twitter Streaming API [2].

3.2.3 Correlation Analysis

Firstly, we apply Pearson Correlation Analysis to the candidate features to find highly correlated features. Pearson Correlation Analysis is widely employed to check how several variables are correlated. The Pearson Correlation Coefficient is a measure of the linear correlation or dependence between two variables, giving a value between $+1$ and -1, where $+1$ stands for total positive correlation, 0 for

[1] https://dev.twitter.com/rest/public
[2] https://dev.twitter.com/streaming/overview

Table 3.1: Correlation Coefficient table for candidate features

	Twitter Age	Verified	Tweets	Followers	FF Ratio	New Followers	Public Lists	New Tweets	Retweets	Mentions	AT Ratio
Twitter Age	1	.297	.248	.006	.015	-.119	.214	-.147	-.173	-.181	-.158
Verified	.297	1	-.254	.229	.149	.209	.282	-.114	.194	.171	.163
Tweets	.248	-.254	1	-.075	-.211	-.059	.097	.363	-.086	-.066	-.087
Followers	.006	.229	-.075	1	.006	.635	.846	-.056	.476	.460	.391
FF Ratio	.015	.149	-.211	.006	1	.000	-.062	-.098	-.100	-.094	-.057
New Followers	-.119	.209	-.059	.635	.000	1	.584	.012	.755	.823	.874
Public Lists	.214	.282	.097	.846	-.062	.584	1	-.002	.479	.475	.415
New Tweets	-.147	-.114	.363	-.056	-.098	.012	-.002	1	-.034	.049	-.045
Retweets	-.173	.194	-.086	.476	-.100	.755	.479	-.034	1	.972	.909
Mentions	-.181	.171	-.066	.460	-.094	.823	.475	.049	.972	1	.960
AT Ratio	-.158	.163	-.087	.391	-.057	.874	.415	-.045	.909	.960	1

no correlation, and -1 for total negative correlation [99].

For the original eleven candidate features, we calculate Pearson Correlation Coefficient on each pair of variables, resulting in a Correlation Coefficient table (Table 3.1). This coefficient table shows that there are three variables highly correlated with each other (correlation coefficient > 0.9): *New Mentions*, *New Retweets* and *AT Ratio*. This finding coincides with the study by Cha et al. [47] that also claimed a strong correlation between *mentions* and *retweets*.

3.2.4 Computing Weights by Entropy Method

In system science, Entropy is a measure of the disorder degree in the system. The bigger entropy value indicates higher degree of disorder. It can also be used to measure how effective information is provided for the evaluation objects

by the indicators in a dataset [100, 101, 102]. In our case, the objects are the users, the indicators are the features, and the dataset includes all the feature values for each user. Entropy method is an objective empowering method. It determines the weights from the characteristics of the dataset itself without any human factors. Owing to this, it is popularly employed for the determination of weights in comprehensive evaluation. In the following discussion, we denote the weight calculated by Entropy method as Entropy Weight.

For a particular feature, if the entropy weight of the feature is bigger than other features, it means this feature contributes more effective information than other features. In the case every user has the same value on a particular feature, the entropy reaches the maximum. This indicates that this feature makes no sense in the evaluation, and the entropy weight should be zero.

Before we compute entropy weights, we adjust the candidate features. Firstly, we remove the feature *verified*, since it only has two values (true or false), providing little effective information. Next, Entropy method requires no or small correlation among original feature set. Therefore, we put *AT ratio* in the evaluation list of the features, while remove *new retweets* and *new mentions*. These three features are highly correlated, and actually *AT ratio* contains the information of *new retweets* and *new mentions* in its calculation. As the result of adjustment, we have eight features left for the computation of entropy weights.

There are three steps to compute the entropy weights for the features [102].

In the following formulas, supposing there are m features and n users being evaluated. The original data matrix is $X = (x_{ij})_{m*n}$, in which x_{ij} is the value of the i-th feature for the j-th user.

1) Standardizing data matrix.

In order to normalize the feature values for all the users, we need to standardize the data matrix as follows:

(1.1) For those features playing a positive role,

$$r_{ij} = \frac{x_{ij} - \min\limits_{j}\{x_{ij}\}}{\max\limits_{j}\{x_{ij}\} - \min\limits_{j}\{x_{ij}\}}$$

(1.2) For those features playing a negative role,

$$r_{ij} = \frac{\max\limits_{j}\{x_{ij}\} - x_{ij}}{\max\limits_{j}\{x_{ij}\} - \min\limits_{j}\{x_{ij}\}}$$

In our case, all the features play a positive role.

After the standardization, we will get the new data matrix $R = (r_{ij})_{m*n}$, in which $r_{ij} \in [0, 1]$ is the standardized feature value of the i-th feature for the j-th user.

2) Defining Entropy.

The entropy of i-th feature is defined as follows:

$$H_i = -k \sum_{j=1}^{n} f_{ij} \ln f_{ij} \ , \ i = 1, 2, \cdots, m$$

$$f_{ij} = r_{ij} / \sum_{j=1}^{n} r_{ij} \ , \ k = 1/\ln n \ ,$$

$$when \ f_{ij} = 0 \ , \ f_{ij} \ln f_{ij} = 0$$

Table 3.2: Entropy weights of manifest features (one month period)

Features	Entropy Weights	Weight Order
AT_Ratio	0.3203	1
New_Tweets	0.1407	2
FF_Ratio	0.1329	3
Followers	0.1248	4
Public_Lists	0.0975	5
Total_Tweets	0.0874	6
New_Followers	0.0843	7
Twitter_Age	0.0121	8

Table 3.3: Entropy weights of manifest features (one week period)

Features	Entropy Weights	Weight Order
AT_Ratio	0.3095	1
New_Tweets	0.1505	2
FF_Ratio	0.1332	3
Followers	0.1250	4
Public_Lists	0.0977	5
Total_Tweets	0.0876	6
New_Followers	0.0845	7
Twitter_Age	0.0121	8

3) Computing Entropy Weight.

After the entropy is defined, we calculate the entropy weight for each feature as below.

$$w_i = (1 - H_i)/(m - \sum_{i=1}^{m} H_i) \ , \ 0 \leqslant w_{ij} \leqslant 1 \ , \ \sum_{i=1}^{m} w_i = 1$$

Table 3.2 and Table 3.3 list the entropy weights for all the eight features in one month and one week respectively. In both tables, all the features are listed

in descending order of their entropy weights. As shown in Table 3.2, the feature *AT ratio* obtains an entropy weight of 0.3203, accounting for nearly one third of the total weights. The next three features are *new tweets*, *FF ratio* and *number of followers*. These features receive similar entropy weights and occupy around 40% of the total weights.

Although the entropy weights do not directly represent the importance of the features from the perspective of practical significance, however, they do show the effectiveness from the view of information provided by the features. From these two Tables 3.2 and 3.3, it can be seen that the entropy weights for all the features are very similar, and the weights orders are exactly the same as each other. For example, the *AT_Ratio* feature provides the biggest part of effective information (more than 30%), while *Twitter_Age* has the least contribution for effective information (1.21% in both tables). This result indicates that each feature contributes similar effective information regardless of the length of observation time period.

3.3 Analysis of Commercial References

In recent years, a number of commercial influence scoring services gained a great deal of attention. All these services provide numerical scores that quantify users' influence in online social networks. Of these services, *Klout* is one of the most well-known influence scoring services [103]. Launched in 2008, *Klout* uses Twitter, Facebook, Google+, LinkedIn, and other social media data to create user profiles,

which are used to generate a unique *Klout Score*. The *Klout Score* is a number between 1 and 100, representing a user' influence. As *Klout* claims, the higher a user's *Klout Score*, the more power the user's influence.

Similarly, *Kred* measures user influence in online communities. It has dual metrics for Influence and Outreach. *Kred*'s influence score ranges from 1 to 1,000. Higher scores represent greater influence. Likewise, *PeerIndex* is a popular social ranking site, founded in 2009. *PeerIndex* provides *PeerIndex Scores* to indicate how influential an individual is, by analyzing the user footprints from major social media services. Another popular influence service is *Followerwonk*, which also offers its own measures of Twitter users' influence.

Influence scoring services have been helping many marketers to successfully run marketing campaigns on social media sites. For instance, some of the largest brands in the world, such as Sony, Nike, P&G, Disney, are using *Klout* to find the most relevant people for their campaigns and increase brand awareness.

All these influence scoring services claim that they take hundreds of signals into consideration, and each service is said having distinct algorithms of computing influence scores. However, there is no public information about the actual features being used by the services. We believe that understanding the representative features used by these services will greatly help people understand the different preferences and priorities adopted in particular services.

Therefore, we employ Spearman's *Rank Correlation Analysis* or *RCA* to ana-

lyze the most principal features used by the four services (*Klout*, *Kred*, *PeerIndex* and *Followerwonk*). Compared with other correlation analysis methods, Spearman's Rank Correlation Analysis is more appropriate for small-sized samples, and for the situations where the variables are not normally distributed or the relationships are not linear [104].

In *RCA*, correlation coefficient r_s is calculated between each manifest feature and each influence score made by a particular service. The coefficient is a value between $+1$ and -1 inclusive, where $+1$ indicates perfect positive monotonic correlation, 0 means no correlation, and -1 stands for perfect negative correlation. The closer r_s is to $+1$, the stronger the positive correlation is. For this reason, the calculated correlation coefficient can be used as an important measure of the correlation between a feature and a specific kind of influence score.

We conduct two experiments of applying *RCA* between manifest features and commercial influence scores, for the time periods of one month and one week respectively. In both experiments, we remove the feature *verified* from the investigation, as it does not make sense to take a Boolean variable in the *RCA*.

Table 3.4 shows the *RCA* results for the one-month experiment. As shown in the table, among the top three features which have biggest values of correlation coefficient, there are two common features: *new mentions* and *number of public lists*. That means these two features have the strongest correlation with the influence scores given by all the four influence scoring services.

Table 3.4: Spearman's RCA between manifest features and influence scoring services (one month period)

	Klout	Kred	PeerIndex	Followerwonk
Twitter_Age	0.365	0.199	0.333	0.173
Total_Tweets	0.141	0.375	0.308	0.368
Followers	0.362	0.473	0.607	0.462
FF_Ratio	0.289	-0.018	0.133	0.103
New_Followers	0.401	0.379	0.453	0.450
Public_Lists	0.731	0.652	0.819	0.581
New_Tweets	0.292	0.347	0.324	0.416
Retweets	0.625	0.732	0.604	0.766
Mentions	0.773	0.817	0.765	0.740
AT_Ratio	0.685	0.627	0.568	0.562

Besides these two features, we find that the coefficient values of *new retweets* and *AT ratio* are both greater than 0.5, indicating that both of these two features are also effective indicators for representing the influence scores generated by all these services.

A special case in Table 3.4 is the coefficient value (0.607) of the feature *number of followers* for the service *PeerIndex*. This shows that *PeerIndex* may assign more weights to *number of followers* in its ranking algorithm.

In the experiment of *RCA* for one week period, there is no difference between the accumulative features. Therefore, Table 3.5 only lists the correlation coefficient values between the instantaneous features (i.e. *new followers*, *new tweets*, *retweets*, *mentions* and *AT ratio*) and all the services. As seen from the table, these coefficient values are slightly smaller than those in one-month experiment,

Table 3.5: Spearman's RCA between instantaneous features and influence scoring services (one week period)

	Klout	Kred	PeerIndex	Followerwonk
New_Followers	0.423	0.381	0.452	0.442
New_Tweets	0.302	0.350	0.331	0.421
Retweets	0.589	0.719	0.568	0.745
Mentions	0.771	0.826	0.767	0.741
AT_Ratio	0.631	0.622	0.551	0.571

but remain the same order. The result demonstrates that one week period might be relatively too short to reflect a user's influence performance. This also explains the reason why Twitter also defines the active users in a monthly basis in its own usage statistics.

The above experiments and analysis reveal several novel findings which are hardly discussed in existing studies of measuring user influence on Twitter. Firstly, besides the *retweets* and *mentions* related features which have been widely used to measure user influence in literature [47, 48, 49, 75, 105], we find that the features of *new tweets* in Table 3.2 and *number of public lists* in Table 3.4 are also good indicators to reflect the activeness and popularity of the user, therefore, fairly effective signals for user influence.

Secondly, in terms of the natures of the manifest features (defined in Fig. 3.1), we find that most of the features in the quadrant of "instantaneous & others' action" show the strongest correlation with the influence scores of popular influence scoring services. Furthermore, the features of others' action are more

effective than those of personal profile, and the instantaneous features are better predictors compared with accumulative ones. This unfolds two important factors in measuring user influence on Twitter: the recent situation, and other users' actions.

3.4 Identifying Hidden Social Attributes

In the previous sections, we have analyzed which manifest features are more important for influence evaluation. These features are explicit variables which can be retrieved by Twitter API. In this section, we further study how the latent features drive influence on Twitter. Latent features are the hidden variables which cannot be calculated explicitly. However, they are the intrinsic attributes, driving the manifest features. For example, engagement is an abstract concept, which cannot be clearly represented by a specific value. In Twitter circumstance, we can explain strong engagement as joining in Twitter early, a large number of tweets, keeping high tweet frequency, etc. Our objective is to discover the principal latent variables behind the manifest features.

In sociology area, there is plenty of research work around social influence. Malcolm [50] described influential people in three ways: *Connectors*, *Mavens*, and *Salesmen*. Each type has their own distinct traits. Connectors are connected with large numbers of people. Mavens are information specialists. And salesmen are charismatic people with powerful negotiation skills. Similarly, Keller and

Berry [98] proposed five attributes of influencers: *Activists*, *Connected*, *Impact*, *Active minds*, and *Trendsetters*. It means that influencers get involved with their communities, have large social networks, are trusted by others, have diverse interests, and tend to be early adopters in markets. However, these social attributes contribute to influence in different ways among different fields. We try to map hidden features to social attributes, and to find which ones are the principal factors in Twitter environment.

First of all, we employ Principal Component Analysis (PCA) [106] to discover latent features. PCA is a statistical method that uses an orthogonal transformation to convert a set of correlated variables into a set of linearly uncorrelated variables called principal components. Then we use Stepwise Multiple Linear Regression (SMLR) [107] to work out key latent features. SMLR is an approach to find the most parsimonious set of predictors that are most effective in predicting the dependent variable. In SMLR, variables are added to the regression equation one at a time, using the statistical criterion of maximizing the R^2 (coefficient of determination) of the included variables. The process of adding more variables stops when all of the available variables have been included or when it is not possible to make a statistically significant improvement in R^2 using any of the variables not yet included.

We perform PCA with IBM Statistical Package for the Social Sciences (SPSS), in order to discover the underlying constructs. Table 3.6 is the computed com-

Table 3.6: Result of Principle Component Analysis

	Comp-1	Comp-2	Comp-3	Comp-4
Twitter_Age	-.132	.526	.555	-.298
Total_Tweets	-.083	.796	-.282	.000
Followers	.703	.224	.431	.255
FF_Ratio	-.074	-.373	.384	.661
New_Followers	.918	-.010	-.003	.076
Public_Lists	.692	.435	.430	.145
New_Tweets	-.018	.449	-.615	.534
New_Retweets	.920	-.114	-.165	-.134
New_Mentions	.941	-.092	-.221	-.086
AT_Ratio	.918	-.156	-.187	-.138

ponent matrix in SPSS. As shown in the table, four principal components have been extracted.

To further understand which components have a significant relationship to the dependent variable (i.e. influence scores), we use SMLR to identify the most important predictors for influence evaluation. In our study, we regard the average of standardized scores of the four influence services as the dependent variable. We believe that, each popular influence scoring service has its own rationality in the scoring scheme, but the average influence score should better reflect the level of user influence from more comprehensive aspects.

Table 3.7 shows that, the first three components extracted from PCA have been included in the final result. As can be seen by examining the *Beta* weights in Table 3.7, the first component (Comp-1) followed by the second component (Comp-2) followed by the third component (Comp-3) are making relatively larger

Table 3.7: Result of Stepwise Multiple Linear Regression

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.012E-013	.086		.000	1.000
	Comp-1	.072	.020	.346	3.649	.000
2	(Constant)	-1.012E-013	.082		.000	1.000
	Comp-1	.072	.019	.346	3.872	.000
	Comp-2	.194	.053	.325	3.635	.000
3	(Constant)	-1.012E-013	.080		.000	1.000
	Comp-1	.072	.018	.346	3.958	.000
	Comp-2	.194	.052	.325	3.715	.000
	Comp-3	.133	.057	.202	2.306	.023

contributions to the prediction model.

Although generally, the extracted principal components by PCA are uninterpretable, we are able to understand the hidden factors through the component loadings in Table 3.6. If we focus on the values above 0.4, we can find the first component has high correlations with the manifest features *new mentions*, *new retweets*, *AT ratio*, *new followers*, *number of followers* and *number of public lists*. These features reflect the attentions users have gained. Therefore, we explain Comp-1 as "Popularity". Similarly, the second component (Comp-2) mainly contains the information from the features *number of tweets*, *age of Twitter account* and *new tweets*, which are signals of user "Engagement". The third component (Comp-3) has positive correlation with *age of Twitter account*, *number of followers* and *number of public lists*, while it has negative correlation with *new tweets*

and *number of tweets*. This means this component reaches bigger value when users make less efforts but obtain more attentions. We consider it is similar to the concept of "Authority".

Compared with the social theory we introduced before, we can see that these three intrinsic features are fairly relevant or similar to the social attributes of influencers. For example, "popularity" actually reflects the *connected* attribute. "Engagement" is a synonym for the *activist* attribute. And "authority" expresses the character of *mavens* or the *impact* attribute.

Based on discussion above, we conclude that, in Twitter environment, the most important social attributes which drive the influence are: *popularity*, *engagement* and *authority*.

Moreover, to better understand the principal features in the popular influence scoring services, we demonstrate the top three important manifest features and hidden attributes for each service in Table 3.8, based on Spearman's correlation analysis. As we can see, in the ranking algorithms, they consider various features in different ways. The following findings provide new insights for influence evaluation nowadays.

- *New mentions* and *Number of public lists* are the two common features among the top three manifest features for the four influence scoring services. The latter has hardly been mentioned in recent research on influence evaluation, but it is really an efficient indicator according to our study.

Table 3.8: Top three manifest features and hidden attributes for influence scoring services

		Klout	Kred	PeerIndex	Followerwonk
Manifest Features	#1	New mentions	New mentions	Number of public lists	New retweets
	#2	Number of public lists	New retweets	New mentions	New mentions
	#3	AT ratio	Number of public lists	Number of followers	Number of public lists
Social Attributes	#1	Popularity	Popularity	Popularity	Popularity
	#2	Authority	Engagement	Authority	Engagement
	#3	Engagement	Authority	Engagement	Authority

- *New mentions* and *New retweets* are explicit signals to evaluate influence, just as many researchers claimed [47, 48, 49, 75]. However, in these four influence scoring services, they pay more attention to *New mentions* than *New retweets*, except Followerwonk.

- "Popularity" is the most important social attribute for influencers. After that, Klout and PeerIndex emphasize "Authority", while Kred and Followerwonk consider "Engagement" a higher priority.

3.5 Summary and Discussion

This chapter analyzes the principal features for measuring user influence on Twitter. We select eleven manifest features based on sociology knowledge and related work. To classify these features, we characterize these features by two dimensions: time (accumulative vs instantaneous) and source (personal profile vs others' action).

We collect the dataset by Twitter APIs, targeting at top 100 most-followed users in Australia. We analyze the principal manifest features in two ways. On

one hand, we employ Entropy method to compute the weights for each feature. On the other hand, we apply Spearman's RCA between features and four popular influence scoring services. Both results show that the features in the quadrant of "instantaneous & others' action" (i.e. *new mentions*, *new retweets*, *AT ratio*) are the most effective predictors for influence. Besides, *number of public lists* is a quite effective feature to demonstrate the influence scores worked out by popular influence services. We also find that the features *new tweets*, *FF ratio* and *number of followers* are also worth paying more attention from the perspective of information effectiveness.

Furthermore, we analyze the hidden features, derived from the manifest features. PCA is employed to discover the hidden components and then SMLR is used to identify the most important hidden features. Finally, we try to map the hidden features to the attributes of influencers in the study of social science. Our analysis discovers that three social attributes are most important for measuring user influence on Twitter. They are: *popularity*, *engagement* and *authority*.

To the best of our knowledge, the study of this chapter provides the first comprehensive analysis of the principal manifest features and hidden social attributes for measuring user influence on Twitter. Nevertheless, this study is based on the actual dataset with limited sample size and time span. In future work, we will recruit large-sized tweet datasets and introduce content-related features for further study. Moreover, we are planning to employ state-of-the-art machine

learning techniques to perform feature selection and influence prediction on Twitter platform. Finally, we will move forward to domain-based or topic-based user influence in next steps, since these kinds of influence are very helpful and beneficial for information dissemination and marketing campaigns.

4

A Hybrid Feature Selection Method for Predicting User Influence

4.1 Introduction

Nowadays, Online Social Networks (OSNs) have gained increasing attention from all over the world, and they are among the most popular websites on the Internet. OSNs not only provide individual users a platform to share information and keep in touch with their friends, but also become an important marketing channel for companies and organizations. It is crucial for companies to establish themselves in OSNs. At the same time it is important for them as well to identify influential people as the marketing targets in OSNs. Both academic community and industry have shown great enthusiasm on the study of user influence in OSNs.

The study of influence originated from psychology and sociology. Briefly speaking, influence is the ability to cause a change in others' thoughts or ac-

tions. Due to this nature, it is difficult to define a quantitative measure for influence in OSNs. In the context of Twitter, we are unable to detect whether there are some changes in others' mind, and despite no online actions (such as reply or retweet) observed, there might be some offline actions (such as buying a product). Therefore, to predict user influence on Twitter, we can only speculate from available features, including characteristics of a user and explicit actions from others.

Intuitively we can say that a user with a large number of followers is influential. However, estimating influence only by followers may introduce noises. For example, it is quite possible that some followers of a particular user are from faked accounts or even spammers. The study by Cha et al. [47] claimed that the top influencers showed a stronger correlation with retweets and mentions than followers. A similar metric for measuring user influence is the ratio between follower count to friend count. This ratio probably describes the types of users in the community, but it is imprecise to measure user influence [48].

In recent years, some Influence Scoring Services (ISS) have gained attention by offering numerical scores that quantify users' social media influence. Klout [1] utilizes social media analytics to rank users according to their online social influence via Klout Score, which is a numerical value between 1 and 100. Klout Score turns out to be a good reference for user influence and gets widely used in

[1] https://klout.com/

industry. Some famous brands in the world, such as Sony, Nike, Disney, are using Klout for business to run successful marketing campaigns by targeting valuable influencers. Kred [1] also measures user influence in online communities. It has dual metrics for Influence and Outreach. Kred influence score ranges from 1 to 1,000. PeerIndex [2] is another social ranking site, which provides social media analytics based on footprints from use of major social media services, and works out PeerIndex scores to indicate how influential an individual is. Another popular online influence tool is Followerwonk [3], which also offers its own measures of user influence (called as Social Authority) on Twitter.

Many researchers have also proposed their own algorithms for predicting user influence on Twitter [47, 48, 73, 74, 75, 76, 78]. However, most studies directly utilize their own pre-defined features to build the model without a pre-evaluation process for these selected features. No literature has extensively investigated and evaluated the potential features that can be used to measure user influence in Twitter environment. We believe that identifying important features that are crucial for influence measurement is the first step towards influence model construction, and this is what we are going to investigate in this chapter.

As an extension of the research work in Chapter 3, this chapter focuses on the feature selection process, which is the most important step to build influence model. A feature selection method is proposed to obtain an optimal feature set

[1] http://kred.com/
[2] http://www.peerindex.net/
[3] https://followerwonk.com/

for predicting user influence on Twitter. The main features of this work are summarized as follows:

- The method inherits the advantages of commonly used *filter* and *wrapper* approaches to achieve a high degree of efficiency and accuracy in the optimization.

- To the best of our knowledge, this work is the first one to intensively study the feature selection for evaluating/predicting the online user influence. This work can provide a solid foundation for further analysis of user influence to cover complicated situations.

- This work employs the five attributes of influencers defined in sociology as the criteria to explore the candidate features in online social networks. Experiments based on a real world Twitter dataset show the effectiveness of the proposed method.

The remainder of the chapter is as follows. Section 4.2 investigates the candidate features. Section 4.3 describes our hybrid feature selection method in the context of Twitter. Section 4.4 provides the detailed description of our experiments and discusses our experimental results. Finally, in Section 4.5, we conclude the chapter with a brief discussion on future work.

4.2 Determining Candidate Features

As we mentioned earlier, algorithms have been developed for predicting user influence based on different features, such as followers, retweets, mentions, tweet content, etc. In this section, we firstly introduce a background theory from sociology. Then we select corresponding features from Twitter based on this theory.

4.2.1 Five Attributes of Influencers

In sociology area, there are numerous studies on identifying influencers. For example, Keller and Berry [98] define the influentials from five attributes:

- **Activists:** Influencers are active in their communities. They attend community events, serve on committees, and persuade others of their opinions.
- **Connected:** Influencers have richer social connections than the average. Their contacts are likely to bring them into connection with more people, in an ever-widening network.
- **Impact:** Influencers are people others look up to for advice. They are trustworthy and reliable because of their reputation or expertise.
- **Active minds:** Influencers have interests in many areas, and they always share their new experience or ideas with others.
- **Trendsetters:** Influencers tend to be early adopters in markets. A new fashion or trend usually originates from some influential people.

The study of the traditional influence in sociology starts much earlier and goes further than that of online influence. Thus, we consider these five attributes of influencers as our basis for feature selection. However as we observed, influence and characteristics of influencers might appear differently in different environments.

4.2.2 Candidate Features from Twitter

Firstly, we select candidate features which are possible predictors for target value (i.e. user influence) as the original feature set. We capture the relevant features from Twitter, and map them to the five social attributes mentioned before. And these features are available through public Twitter APIs.

User influence is changing over time. For instance, it is often seen that a user's influence suddenly increases due to some emerging news, or gradually drops as a result of low level of engagement. Therefore, we do not consider only the *long-term* features which reflect a user's accumulative efforts or achievements, but also the *short-term* features which reflect a user's dynamic situation in a recent period of time.

Based on the above thoughts, we select 17 candidate features (listed in Table 4.1) for predicting user influence on Twitter. All the features starting with "New" are short-term features. Besides, the other five features (*Topic Diversity*, *Average Length of Tweets*, *Original Tweets*, *Original Tweet Ratio* and *Average Retweets of Original Tweets*) are also calculated based on the tweet data in a specific period

Table 4.1: Candidate features for predicting user influence on Twitter

Social Attributes	**Features on Twitter**
Activists	Tweet Frequency
	New Tweets
Connected	Followers
	Friends
Impact	Verified
	Public Lists
	New Public Lists
	New Followers
	New Mentions
	New Retweets
	Followers to Friends Ratio
Active Minds	Topic Diversity
	Average Length of Tweets
Trendsetters	Account Age
	Original Tweets
	Original Tweet Ratio
	Average Retweets of Original Tweets

of time. Note that the complete historical tweet data is not available through Twitter APIs.

These candidate features are briefly explained as follows. The features mapped to *Activists* attribute include:

(1) Tweet Frequency: the average number of tweets a user posted per month, since his/her account was created. It represents the active level of the user.

(2) New Tweets: the number of tweets issued by a user during a recent period of time. We understand that influence is time sensitive, since user influence

in a social network is changing over time. If an influential user does not post any tweet for a period of time, his/her influence probably starts declining.

The features mapped to *Connected* attribute include:

(3) Followers: the total number of followers a user has. It is an explicit metric of connectivity. Generally, more followers mean more potentials of a high degree of influence.

(4) Friends: the total number of friends a user has. It is a metric of outgoing connection.

The features mapped to *Impact* attribute include:

(5) Verified: whether it is a verified account. Twitter verifies accounts on an ongoing basis, focusing on popular users in interest areas, such as music, acting, politics, media, sports, business and others.

(6) Public Lists: the number of public lists which a user is a member of. Being included in public lists indicates a user is visible and people show interest in the user.

(7) New Public Lists: the number of new public lists which include the user in a recent period of time. An increase of public lists implies a user's continued impact.

(8) New Followers: the number of new followers during a recent period of time. Whether a user's follower count is increasing or decreasing is a good metric for user influence.

(9) New Mentions: the number of new mentions or replies to a user during a recent period of time, i.e., the number of tweets including "@username" (excluding "RT @username"). Mention is an explicit signal reflecting a user's impact to others.

(10) New Retweets: the number of new retweets of the tweets created by a user during a recent period of time, i.e., the number of tweets including "RT @username". Retweeting indicates the retweeter has been influenced by the original author, no matter positive or negative impact.

(11) Followers to Friends Ratio: the ratio of a user's follower count to friend count. The higher the ratio is, the more people are interested in the user's status updates without the user showing interest in return.

The features mapped to *Active Minds* attribute include:

(12) Topic Diversity: a metric measuring how many different topics a user's tweets might cover. To understand topic diversity, we train a Latent Dirichlet Allocation (LDA) model based on the corpus of tweets. We combine all tweets posted by a user during one-month period as one document and obtain the document-topic distribution by training LDA model with Stanford

Topic Modeling Toolbox [1]. Then the entropy of document-topic distribution is computed to represent the topic diversity. If a user only concentrates on one or two topics, the entropy is relatively small, while larger entropy indicates more diverse topics are covered.

(13) Average Length of Tweets: The activists tend to share their ideas with more words, and the average length of tweets is an obvious indicator. Studies have shown that on Twitter with 140-character limit, too short text conveys little information and informative content with enough words is critical to gaining attention [108].

The features mapped to *Trendsetters* attribute include:

(14) Account Age: the number of months since a user's Twitter account was created. It reflects whether a user is an early adopter on Twitter.

(15) Original Tweets: the number of original tweets (excluding replies and retweets) a user created during a recent period of time. A trendsetter should have some original thoughts, rather than always joining conversations or forwarding information.

(16) Original Tweet Ratio: the ratio of original tweets to total tweets. It reflects the relative originality based on the user's total tweets.

[1] `http://nlp.stanford.edu/software/tmt/tmt-0.4/`

(17) Average Retweets of Original Tweets: the average number of retweets obtained per original tweet. This metric reflects a user's performance in trendsetting.

4.3 A Hybrid Feature Selection Method

There are two main categories of feature selection methods: filter methods and wrapper methods, which are demonstrated in Fig.2.4. A filter method directly evaluates the the quality of features according to their data values. A wrapper method employs learning algorithms as the evaluation criteria to select optimal feature subsets. Comparing with a filter method, a wrapper method is more effective, but it often brings in a higher degree of computational complexity. In this work, we combine the advantages of these two types of methods and propose a hybrid filter-wrapper method for predicting user influence on Twitter. The filter method provides a quick way to eliminate the less relevant features and then the wrapper method is employed to achieve a high accuracy.

4.3.1 The Proposed Method

The proposed method is illustrated in the flowchart (Fig.4.1).

We explain the detailed procedure in the following seven steps.

Step 1: Determine the candidate feature set $F = \{F_1, F_2, ..., F_f\}$.

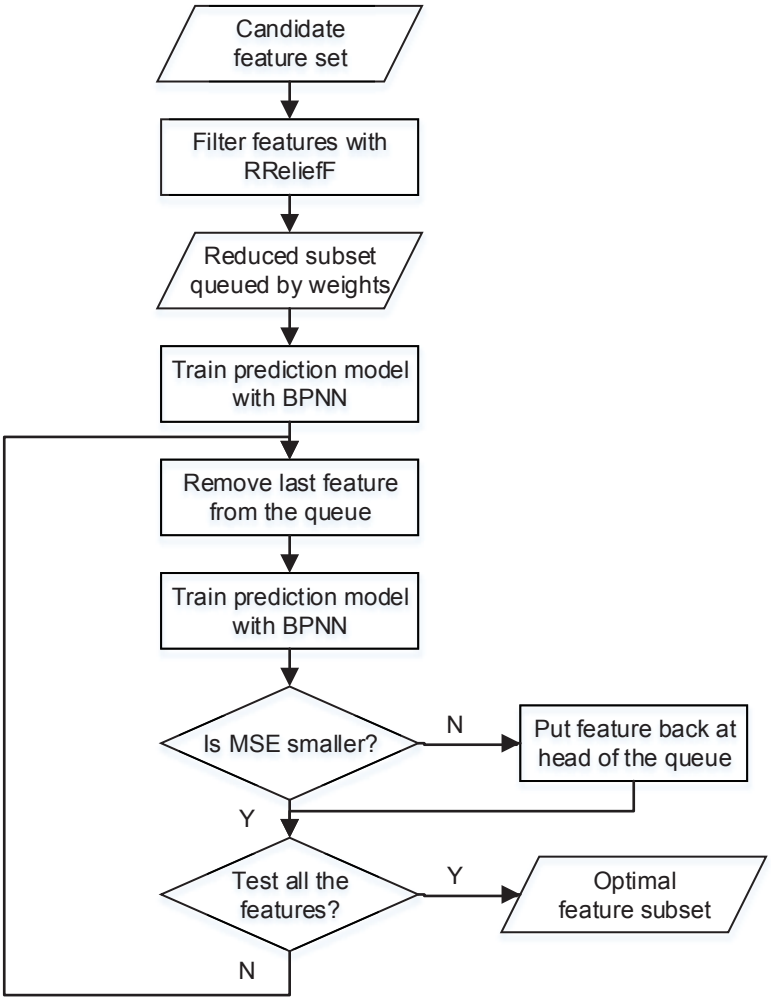

Figure 4.1: Our proposed hybrid feature selection method

Step 2: Utilize the feature weighting algorithm (*RReliefF*) to compute the weights for each feature, and filter out the features which have little relevance to the target user influence. The remaining features form a reduced feature set $F' = \{F_1, F_2, ..., F_n\}(n <= f)$, in which features are sorted in descending order based on their weights.

Step 3: Train prediction model with a learning algorithm (BPNN) on training set and calculate Mean Square Error (MSE) on testing set, denoted as e. Meanwhile, set $i = n$.

Step 4: Remove the feature F_i which is the last feature in F', set $F' = F' - \{F_i\}$, then train prediction model based on F' and calculate new MSE, denoted as e'. If $e' < e$, then set $e = e'$ and go to step 6.

Step 5: Move the feature F_i back into F' as the new first feature, and set $F' = \{F_i\} + F'$.

Step 6: Set $i = i - 1$, if $i > 0$, go back to Step 4.

Step 7: The feature subset F' is the optimal feature subset.

4.3.2 Filter - Feature Ranking

Estimating the quality of features is critical in feature selection. A robust feature weighting technique is the *Relief* algorithmic family [109]. *Relief* algorithms estimate how well the features' values distinguish between instances. The output of *Relief* algorithms is a set of numerical weights representing the percentages of the features' contribution to the variance in dependent variable. Features that are assigned weights larger than zero can cause the dependent variable to vary, while those features with zero or negative weights are believed to have no contribution to the variance of the dependent variable. *Relief* algorithms perform well even when strong dependencies exist between features and have been used successfully in a variety of contexts. In our method, we employ *RReliefF* [110], which is designed for regression problems, to rank all the candidate features.

Assume $I_1, I_2, ..., I_j$ are the instances with n features $F_1, F_2, ..., F_n$ and target values. To estimate the weights of all features (denoted by $W[F]$), *RReliefF* starts with selecting k nearest instances around a randomly selected instance I_i and then updates the weight estimation $W[F]$ for all features F based on probabilities of difference. This whole process is repeated for m times. Here k and m are user-defined parameters, and $W[F]$ is calculated as below:

$$W[F] = \frac{P_{diffC} \mid_{diffF} P_{diffF}}{P_{diffC}} - \frac{(1 - P_{diffC} \mid_{diffF}) P_{diffF}}{1 - P_{diffC}}$$

where

$$P_{diffF} = P(\text{diff. value of } F \mid \text{nearest instances})$$

$$P_{diffC} = P(\text{diff. prediction} \mid \text{nearest instances})$$

$$P_{diffC} \mid_{diffF} = P(\text{diff. prediction} \mid \text{diff. value of } F \text{ and nearest instances})$$

We implement *RReliefF* in the Weka tool [1] (a popular software tool of machine learning) to calculate the feature weights. Due to the problem of underestimating numerical attributes shown in the work of [109] when both numerical and nominal features are in the feature set, we initially remove the nominal feature *Verified* and put the remaining sixteen features into the *RReliefF* algorithm. We set the parameter k, which is the number of nearest neighbours, to 10 as proposed in [109] and keep other parameters as the default settings in Weka.

4.3.3 Wrapper - Feature Search Strategy and Learning Algorithm

As we show in Fig.2.4, feature search strategy and learning algorithm are the two main parts in wrapper methods. Various search strategies have been proposed under two main ideas: exhaustive and heuristic searches. Exhaustive search can guarantee the optimal feature subset but with a high complexity (2^N possible feature subsets, where N is the number of features), while heuristic search can achieve near optimality more efficiently. Among the heuristic search strategies, floating search methods, including Sequential Forward Floating Selection (SFFS) and Sequential Backward Floating Selection (SBFS), are proven to be qualified [111], since they can provide near-optimum or optimum results in most situations.

[1] http://www.cs.waikato.ac.nz/ml/weka/

We go through the feature subsets with SBFS strategy, starting from the full feature set. All the candidate features are removed one by one according to the feature weights (from smallest to largest). If the training model performs better without a certain feature, then this feature will be deleted and the model's performance will be set as the current optimum. However, if the model performs worse, then the tentatively-deleted feature will be kept and put back into the feature set. The search stops when all the features have been examined once.

Neural Networks (NN) are important learning algorithms for modelling complex non-linear systems [112, 113, 114]. A basic NN model has three layers: input layer, hidden layer and output layer. On each layer, there are a number of nodes (or neurons). Nodes on input layer are connected to nodes on hidden layer, and nodes on hidden layer are connected to nodes on output layer. These connections between nodes represent the weights. Back Propagation Neural Network (BPNN) is one of the most popular NN algorithms. The main idea in BPNN is that, the network output is compared to the target output, and the errors propagate backwards from the output nodes to the input nodes. If results are not satisfactory, the connections (weights) between layers are modified and this process is repeated again and again until some stopping criterion is satisfied.

We implement the wrapper method with Neural Network Toolbox [1] in MATLAB to select the optimal feature subset from the remaining features after filter-

[1] http://www.mathworks.com/products/neural-network/

ing. Tan-Sigmoid is selected as the transfer function for the hidden layer. It is defined as:

$$f(x) = \frac{2}{1 + e^{-2x}} - 1$$

And we select the linear transfer function for the output layer, which is defined as:

$$f(x) = purelin(x) = x$$

All the samples are randomly divided to three sample sets: training set (70%), validation set (15%) and testing set (15%). Training stops when generalization stops improving, as indicated by an increase in the mean square error of the validation samples, or the maximum iteration limit is reached. Since training multiple times generates different results, we train the BPNN model for 20 times and utilize the average Mean Square Errors (MSE) to evaluate the performance. The MSE is calculated as below:

$$E = \frac{1}{n}\sum_{k=1}^{n} e(k)^2 = \frac{1}{n}\sum_{k=1}^{n}(t(k) - a(k))^2$$

Here n is the number of samples in testing set, $t(k)$ is the target output and $a(k)$ is the network output. If the MSE becomes smaller than the current optimum, the out feature will be deleted; otherwise, the out feature will be put back into the feature set.

4.4 Experiment and Analysis

We select 200 most-followed Twitter users in Australia, who explicitly indicate "Australia" in their location profiles. All the tweets posted by these users and all the responses (including replies, mentions, retweets) are collected during one month period (from 12 January to 12 February in 2015). There are totally 6,770,715 tweets (around 36.8 gigabytes data) captured in the month.

Different types of features are retrieved or calculated from different Twitter APIs. For example, user profile data are captured through Twitter REST APIs. All the long-term features are calculated from user profile data. Real-time tweet data are captured through Twitter Streaming APIs. Most of the short-term features (e.g. *New Retweets*) and tweet content related features (e.g. *Topic Diversity*) are calculated by analyzing these tweet data.

In our study, we consider the average of standardized scores of the four popular ISS (i.e. *Klout*, *Kred*, *PeerIndex*, *Followerwonk*) as the relative truth of user influence. We believe that, each system has its own rationality in the scoring scheme, but the average influence score should better reflect the level of user influence from more comprehensive aspects. This is also used as the desired output in the supervised learning algorithms involved (such as BPNN).

Firstly, we pre-process the candidate features with the filter algorithm *RReliefF*. The results of calculated feature weights are shown in Table 4.2.

Based on the ideas of *RReliefF*, the last five features with negative weights

Table 4.2: Feature Weights computed by RReliefF algorithm

Rank	Features	Weights
1	Average Length of Tweets	0.016720
2	Account Age	0.015140
3	Followers	0.012932
4	Public Lists	0.012165
5	New Retweets	0.011357
6	New Followers	0.010440
7	New Mentions	0.009830
8	Average Retweets of Original Tweets	0.007138
9	New Public Lists	0.006059
10	Original Tweet Ratio	0.003991
11	Followers to Friends Ratio	0.000187
12	Friends	-0.00101
13	Tweet Frequency	-0.00622
14	New Tweets	-0.01013
15	Original Tweets	-0.01056
16	Topic Diversity	-0.01494

Table 4.3: Average MSE for each iteration of the loop

# of Loop	1	2	3	4	5	6
Avg. MSE	0.0334	0.0282	0.0277	0.0299	0.0280	0.0305
# of Loop	7	8	9	10	11	12
Avg. MSE	0.0261	0.0273	0.0338	0.0271	**0.0260**	0.0315

are filtered out, since they have no contribution to the variance of user influence. Then the eleven remaining features are tested by feature search strategy SBFS. For each feature subset, we train with BPNN algorithm and compute MSE on testing set for twenty times. The average MSE for each subset is shown in Table 4.3.

As we can see from Table 4.3, the average MSE reaches the minimum in the eleventh iteration of the loop. Therefore, this feature subset is the optimal feature subset, which includes seven features: *Average Length of Tweets, Followers, Public Lists, New Retweets, New Mentions, Average Retweets of Original Tweets* and *New Public Lists*.

In view of the mapping between social attributes and candidate features (shown in Table 4.1), we find that most of the features belonging to the *Impact* attribute are included in the optimal feature subset, and all features belonging to the *Activists* attribute are not included. According to the feature weights calculated by *RReliefF* filter algorithm (shown in Table 4.2), we believe that, in the Twitter environment, the social attribute *Impact* plays the most important role, followed by *Active minds*, *Connected* and *Trendsetters*. The attribute *Activists*

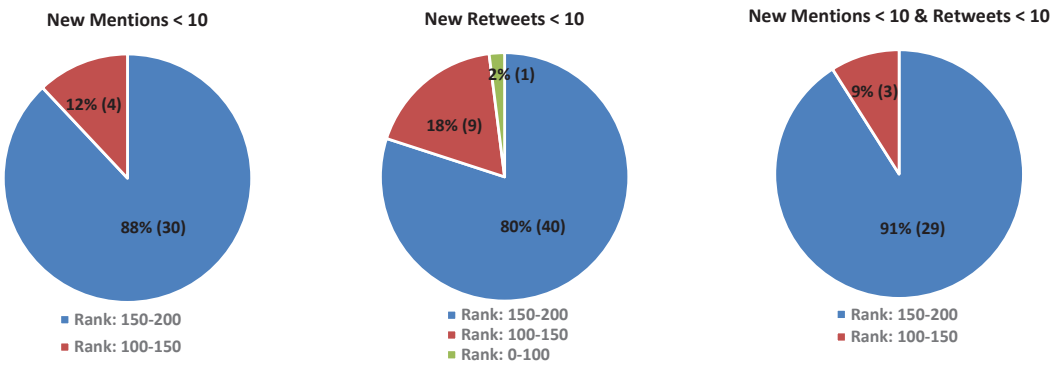

Figure 4.2: Influence analysis with few mentions and retweets

seems to contribute little to user influence.

The identified seven features can be used as the starting point to model the user influence. The influence models will be built up in different ways such as the linear regression or the rule-based evaluation with specific constraints on selected features. For example, we group the 200 users into three influence levels: top 100 influencers, users ranked between 100 and 150, users ranked between 150 and 200. The distributions of users with less than 10 mentions or/and retweets at different influence levels are shown in Fig.4.2. Over 90% of the users with less than 10 mentions and retweets are ranked between 150 and 200. These users have quite low influence scores. While identifying influencers, these users can be filtered out at an early stage.

4.5 Summary and Discussion

This chapter has proposed a hybrid feature selection method for predicting user influence on Twitter. Based on the five attributes of influencers defined in sociol-

ogy, we have explored the candidate features from Twitter and selected seventeen ones as the starting point. We collected the experimental dataset through public Twitter APIs including all tweets associated with the 200 most-followed users in Australia. We employed the *RReliefF* algorithm (a filter method) to evaluate the quality of features. As the result, a reduced feature subset was obtained. Following the principles of wrapper methods, we developed our method by utilizing the SBFS search strategy and assessing the feature subset at each searching step by employing the BPNN learning algorithm.

The proposed method produce an optimal feature set for predicting user influence on Twitter. These features include: *Average Length of Tweets, Followers, Public Lists, New Retweets, New Mentions, Average Retweets of Original Tweets* and *New Public Lists*. Some of these optimal features such as *Average Length of Tweets* and *Public Lists* have rarely been discussed in the existing literature. These features will be used as the basis for further study about incentive mechanisms of user influence and methods to build up user influence on social networks. We will develop user influence models and study topic popularity by considering the characteristics of social network structures.

5

Influence Maximization on Twitter: A Mechanism for Effective Marketing Campaign

5.1 Introduction

With the rapid growth of Online Social Networks (OSNs), there has been a revolutionary change in the way people communicate with each other. Motivated by applications such as viral marketing [115, 116, 117, 118, 119], information diffusion in online social networks has received tremendous attention. Online social networks have become new channels for companies to carry out their marketing campaigns and brought in business opportunities for enterprises.

For instance, a company has released a new product or an online service. This company would like to select a set of users to help the propagation of the marketing information on Twitter. The company expects that these seed users will influence their followers, and then these followers will influence their own

followers as well. As a result, a large number of users could receive this marketing information through the online word-of-mouth effect.

The problem of selecting seed users, referred to as *Influence Maximization* has attracted a lot of interest in the research field of online social networks. Most of existing works on this topic focus on algorithms for the selection of seed nodes. In particular, when the influence maximization problem is studied in a specific social network, the following key questions have often been ignored.

Firstly, what is the definition of influence? Influence is a concept which could show in various ways in different contexts. For example, it might refer to passing a message successfully to others in a task of information diffusion. It might mean that audiences agree with the speaker's arguments in a campaign speech. It might imply that customers are persuaded to buy products in a marketing activity. A clear definition/description for influence in a specific context is crucial when studying the influence maximization problem.

Secondly, how are influence probabilities obtained or computed? The data of influence probabilities are essential in this problem. Most of the studies in this area assume these probabilities are given as input. Only some recent studies [30, 31] have shown how to learn influence probabilities from the historical data of user actions. It is necessary to identify which types of actions in a specific social network should be used in the calculation of influence probabilities.

Thirdly, how is a diffusion model defined? The diffusion model determines

how influence propagates in the networks. A well-defined diffusion model should capture the major characteristics of information spread in a specific social network. It plays an important role in dealing with the influence maximization problem.

In this work, we specify an influence maximization problem which can cover a wide range of marketing campaign scenarios on Twitter. The main contributions are summarized as follows:

- An influence maximization approach has taken into consideration of social ties, user interactions, and information propagation on Twitter. The proposed approach provides a solid generic solution for promoting products and services in online social networks like Twitter.

- An influence probability model is proposed. The influence probability during a specific time period is calculated according to users' action history including tweet, favourite, mention/reply, and retweet.

- An information diffusion model is proposed to capture the major characteristics of information spread on Twitter. This model inherits the classic *independent cascade model* and has the capability to adopt the assumption that a user can have multiple chances to be influenced by others in a considered time period.

- A heuristic algorithm is designed for influence maximization on Twitter.

Experimental results show that this algorithm achieves better influence spread than classic heuristics, and has the influence spread quite close to that of the well-known improved greedy algorithm but uses less than one-thirtieth of its running time.

The remainder of the chapter proceeds as follows. In Section 5.2, we specify the influence maximization problem on Twitter. In Section 5.3 and 5.4, we propose the influence probability model and information diffusion model. In Section 5.5, we develop a heuristic algorithm for the selection of seed users. Section 5.6 provides the details of a set of experiments and discusses the results. In Section 5.7, we conclude the chapter with a discussion on the future work.

5.2 Influence Maximization Problem

The influence maximization problem can be described as follows. A social network is represented by a directed graph $G = (V, E)$, where the nodes V represent users, and the directed edges E represent social ties between users. We are also given a budget k, which is a integer. The goal of influence maximization is to find k users (seed nodes) in the social network so that the spread of influence (defined as the excepted number of influenced users) could be maximized.

Based on the general definition of influence maximization problem above and the specific characteristics of Twitter network, we have the assumptions as follows.

- *Influence*: As a general term, *influence* means "change in a person's cognition, attitude, or behaviour, which has its origin in another person or group" [5]. When the term *influence* is used in the research community of OSNs, many researchers have provided their own explanations about *influence* [47, 49] in their interested contexts. In this work, this term is referred to as "the ability to let someone know something, or pass information to others". We consider u influencing v if v gets the information from u.

- *Influence Probability*: A directed edge $(u, v) \in E$ between users u and v represents the probability of u influencing v, which is denoted as $p_{u,v} \in (0, 1)$. This probability will be calculated according to the *action history* on Twitter, including individual user's actions and interactions between users. More details will be provided in Section 5.3.

- *Information Diffusion*: We assume that the information diffusion can be simulated as a process with multiple discrete steps. A user can have multiple chances to be influenced by activated neighbours during the considered time period. At step t, the nodes which were active at step $t-1$ remain active, and other inactive nodes might be activated based on our probability model. More details will be provided in Section 5.4.

- *Information Maximization*: We specify influence maximization as the problem of selecting a set of users in order to maximize the influence spread

within a specific time period. In this work, we simulate the information diffusion with N discrete steps during this period.

In this work, we utilize influence maximization techniques to support the development of marketing campaigns on Twitter. The maximal information propagation is the goal of the proposed approach. Here the influence maximization problem is studied in the context of a Twitter network with a directed graph $G = (V, E)$ and historical data of actions as inputs. Models of influence probability and information diffusion are proposed. An efficient heuristic algorithm is developed to solve the specified maximization problem.

5.3 Influence Probabilities

Given a graph of a social network $G = (V, E)$, each directed edge $(u, v) \in E$ is labelled with a weight $p_{u,v}$, representing the influence probability with which u will succeed in activating his neighbour v. We assume that the *action history* is given. The *action history* includes information of individual user's actions and interactions between users. Let A_u denote the total number of actions user u performs and R denote the set of interaction types. $I(u, v, a)$ is a function to calculate the number of interaction $a \in R$ with which user v reacts to user u.

On Twitter, users deliver messages by posting tweets. After other users read a tweet, they can respond to the tweet by means of favouriting, replying or retweeting. We assume user u is likely to influence user v only in a fixed-size

time-frame T since u posts a message, and the influence probability does not change over time. In the influence maximization problem, we consider A_u as the total number of tweets the user u posts in a certain time period T. The *action history* contains three kinds of interactions (denoted by R), which are *favourite*, *mention/reply*, and *retweet*.

If user v reacts to user u, it means u has successfully passed the information to v, say u has influenced v. The *influence factor* ($infl(u, v) \geq 0$) from user u to v is defined as the ratio of $v \rightarrow u$ reactions to the total actions performed by u.

$$infl(u, v) = \sum_{a \in R} I(u, v, a) \ / \ A_u \tag{5.1}$$

The influence probability $P_{u,v}$ ($0 \leq P_{u,v} < 1$) is calculated based on the *influence factor* as:

$$P_{u,v} = 1 - \exp(-infl(u, v)) \tag{5.2}$$

We assume that there is always a small influence probability between connected users even if there are no historical interactions between them. If $P_{u,v} < 0.01$, $P_{u,v}$ will be set to 0.01. This constant value 0.01 has also been used in [28, 29]. In our work, the influence time-frame T is divided equally into N slots. The diffusion process moves one step forward in each time slot. $P_{u,v}$ is calculated

as follows.

$$P_{u,v} = 1 - (1 - p_{u,v})^N \tag{5.3}$$

where $p_{u,v}$ is the probability of u influencing v at each step of propagation. Based on Eq. 5.3, $p_{u,v}$ is:

$$p_{u,v} = 1 - (1 - P_{u,v})^{1/N} \tag{5.4}$$

5.4 Information Diffusion Model

In cascade models, when a node u first becomes active, it has a single chance to influence its inactive neighbour v, with a probability $p_{u,v}$. If u succeeds in activating v at step t, then v can make an attempt to influence its inactive neighbours at step $t + 1$. The diffusion process stops until every active node has tried its single chance and there are no more activations.

Based on the problem definition in Section 5.2, we propose a *R-J cascade model* for the information diffusion process. There are two modifications in *R-J cascade model* compared with the basic independent cascade model. Firstly, a user u always has the chance to activate his inactive neighbour v at each step, which means that the attempt from u to v can be *repeatable*. In the context of Twitter, a user can obtain the information (or to say, read the tweet) from others whom he/she follows any day after the information / tweet is posted.

Secondly, if an inactive user v has a set of activated neighbours denoted by S, we predict whether v will be activated based on a *joint influence probability* denoted by $p_v(S)$. At each step, the user v will become active if any of his/her active neighbours succeeds in activating v. Thus, the joint influence probability is calculated as below:

$$p_v(S) = p_{w,v} + (1 - p_{w,v}) * p_v(S \setminus \{w\})$$

where $w \in S$. The user v will be activated unless all his active neighbours fail to activate v. This formula can be expressed as follows.

$$p_v(S) = 1 - \prod_{u \in S} (1 - p_{u,v})$$

The algorithm for estimating the expected influence spread with *R-J cascade model* is provided as Algorithm 1. A table jp is created to store the joint influence probabilities for all users. The joint influence probability of user v is denoted as $p_v(S)$. This table is initialized based on the set of seed users SS (Lines 3-11). During each round of simulation, if there are new activated users, the joint influence probabilities of these users' followers will be updated in the table jp (Lines 20-26).

Algorithm 1 RJCascade

Input: SS, r, N, $p_{u,v}$

Output: s

initialize activated set $AS = SS$
create a table jp to store the joint probability for each user
for each user $u \in V$ **do**
    $p_u(S) \leftarrow 0$
end for
for each user $u \in AS$ **do**
    find the follower set FS of user u
    for each user $v \in FS$ **do**
        $p_v(S) \leftarrow p_v(S) + (1 - p_v(S)) * p_{u,v}$
    end for
end for
$s \leftarrow 0$
for $i \leftarrow 1, r$ **do**
    reset the table jp to initial values
    for $j \leftarrow 1, N$ **do**
        initialize an inactive user set $IS = \varnothing$, who might be activated at this step
        find all the edges $(u, v) \in E$ where $u \in AS$ and $v \notin AS$, and add v into IS
        for each user $u \in IS$ **do**
            generate a random value $r \in (0, 1)$
            if $r < p_u(S)$ **then**
                $AS \leftarrow AS \cup \{u\}$
                find the follower set FS of user u
                for each user $v \in FS$ **do**
                    $p_v(S) \leftarrow p_v(S) + (1 - p_v(S)) * p_{u,v}$
                end for
            end if
        end for
    end for
    $s \leftarrow s +$ number of users in AS
end for
$s \leftarrow s/r$

5.5 Approximation Algorithms

Kempe et al. [28] develop a greedy algorithm to solve their identified influence maximization problem and have obtained the best result for expected influence spread comparing with existing approximation algorithms. While dealing with a large real-world social network, this greedy algorithm is inefficient and it is infeasible to get the results in an acceptable period of time on a normal computer.

Kempe et al. [28] prove that the influence function $f(\cdot)$ has the properties of monotonicity and submodularity. The submodularity property means that the marginal gain from adding a user u to a set S is equal to or greater than the marginal gain from adding the same user to a superset of S. In this work, we use the similar idea in the CELF algorithm [24] and present an improved greedy algorithm (Algorithm 2) based on our problem definition in Section 5.2. The details of the algorithm is described in Algorithm 2. A data structure *slist* is used to store users' ids and their incremental influence spreads. The first round selection (Lines 2-5) is the same as the original greedy algorithm. The users in the *slist* are sorted in a descending order of the incremental influence spread. From the second round, users in the *slist* are explored one by one. If the evaluated user's incremental influence spread in the current round is bigger than the next user in *slist*, the evaluated user will be selected as a seed user (Lines 11-15). Otherwise, move to the next user and then repeat this process until the whole *slist* has been explored (Lines 17-31). In each round, the re-computations for the

unevaluated users are avoided if the incremental influence spreads of these users in previous rounds are smaller than the biggest incremental influence spread of the evaluated users in the current round. Here the submodularity is utilized to improve the efficiency.

Heuristic algorithms have been developed to tackle the efficiency issue in solving the influence maximization problems. The *high-degree* heuristic algorithm selects the seed nodes based on their degrees (in descending order), i.e. the number of followers on Twitter. Intuitively, the follower count of a user in a social network is considered as an important indicator for user influence. Experimental results in [28] show that the high-degree heuristic algorithm can achieve the performance close to that of the greedy algorithm, outperforming several existing algorithms.

The *distance centrality* is another commonly used influence measure in sociology. It has been evaluated in [28, 29]. The distances from one node to other nodes are measured. The node with shorter average distance to other nodes is regarded at more central position in the social network. Nodes at more central positions are more influential and they will be selected as seed nodes.

We propose an *influence index* heuristic algorithm. The algorithm aims to obtain the expected spread results close to the greedy algorithms with much less computation time. Different from the measures mentioned above, we use *influence index* to represent the overall influence power of a user. It is calculated as the sum of the influence probabilities from one user to others. For example,

Algorithm 2 ImprovedGreedy

Input: $G = (V, E)$, k, $RJCascade$

Output: SS, s

initialize $SS = \varnothing$
for each user $v \in V$ **do**
    $spread \leftarrow RJCascade(v)$
    Add the tuple <*user id, spread*> into *slist*
end for
Sort *slist* in descending order based on *spread* values
Add the first user in *slist* into SS
$s \leftarrow$ *spread* of the first user in *slist*
remove the first user in *slist*
for $i \leftarrow 2, k$ **do**
    $s' \leftarrow RJCascade(SS \cup$ the first user in *slist*$)$
    if $s' - s >=$ *spread* of the second user in *slist* **then**
        $SS \leftarrow SS \cup$ the first user in *slist*
        remove the first user in *slist*
        $s \leftarrow s'$
    else
        $\Delta \leftarrow s' - s$
        spread of the first user in *slist* $\leftarrow s' - s$
        $x \leftarrow 2$
        while $\Delta <$ *spread* of the x-th user in *slist* **do**
            $s' \leftarrow RJCascade(SS \cup$ the x-th user in *slist*$)$
            spread of the x-th user in *slist* $\leftarrow s' - s$
            if $s' - s > \Delta$ **then**
                $\Delta \leftarrow s' - s$
            end if
            $x \leftarrow x + 1$
        end while
        Sort *slist* in descending order based on *spread* values
        $SS \leftarrow SS \cup$ the first user in *slist*
        remove the first user in *slist*
        $s \leftarrow s + \Delta$
    end if
end for

user u has only one outgoing linkage with an influence probability 100%, user v has ten outgoing linkages with an influence probability 10% on each edge. In this case, user u and v are considered to have equivalent influence power. In solving the influence maximization problem, we give priority to the users with larger *influence index* value when selecting seed nodes.

5.6 Experiment and Analysis

We build up an experimental dataset collected from Twitter. This dataset includes the social network data associated with users from a real city *Darwin* in Australia. Firstly, we capture all the users whose location profiles include the word "*Darwin*" (5,276 users as of July 21, 2015). Then we check through the details of user profiles and filter out users who are not from the *Darwin* city in Australia. For example, in some cases, the word "*Darwin*" indicates a city or town in another country rather than Australia, or it actually means a person's name. Finally, the identified social network includes 3,292 users from the city of *Darwin* in Australia.

In the social network graph $G = (V, E)$, if user u has a follower v, there is a directed edge from u to v. There are 23,605 following relationships (i.e. directed edges) and 1,158 isolated users in the *Darwin* community. Furthermore, there are 39 connected components and maximum vertices in a connected component are 2,048. The real social graph is demonstrated in Fig.5.1.

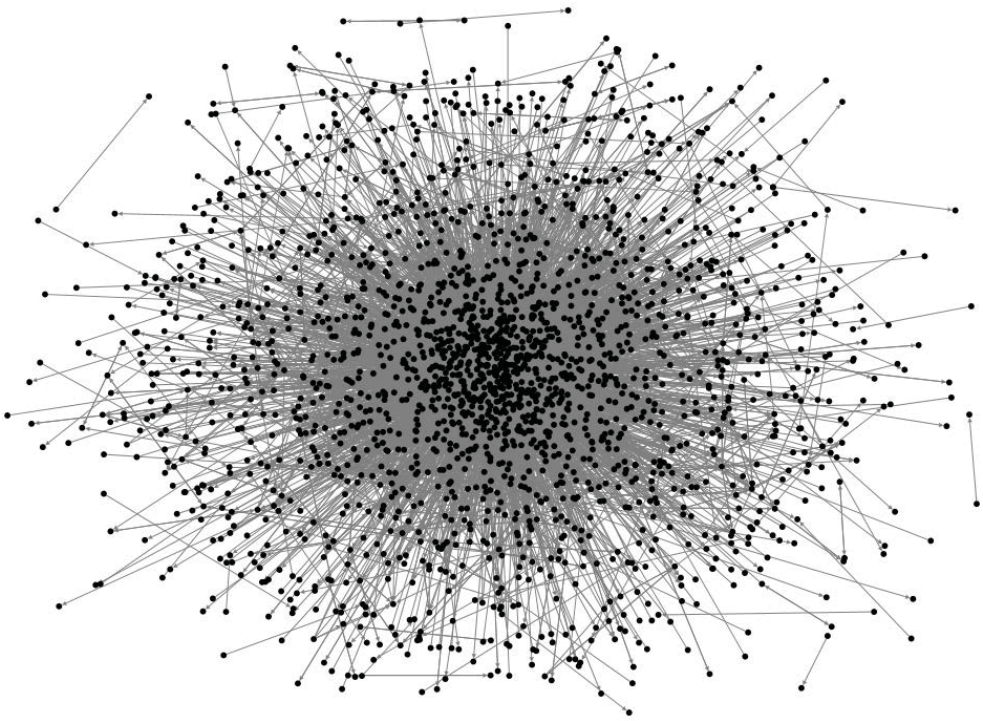

Figure 5.1: A real Twitter social network for Darwin city in Australia

The dataset includes tweets posted by the users from Darwin city in 30 days. According to our influence probability model, the accumulative influence probability of these 30 days is calculated based on the ratio of reactions to total actions. The probability of each day is calculated by Eq. 5.4 with $N = 30$. The influence propagation is simulated by setting one day as one step. The influence maximization goal is to find k seed users ($k = 1, 2, ..., 20$) in order to maximize the influence coverage after 30 days. To compare the performance of different algorithms, we run the following algorithms against a dataset from Twitter.

- **ImprovedGreedy:** The improved greedy algorithm proposed in Section 5.5.
- **InfluenceIndex:** A heuristic algorithm based on users' sum of influence probability to others, defined as *influence index* in Section 5.5.
- **HighDegree:** A simple heuristic algorithm based on users' follower count, which is known as "degree centrality" in sociology literature.
- **DistanceCentral:** A heuristic algorithm based on users' average distance to other users in the whole network.
- **Random:** Seed users are randomly selected.

In order to understand the characteristics of the social network of the Darwin community on Twitter, numbers of followers and influence probabilities (which

are the cumulative probabilities $P_{u,v}$, defined in Section 5.3) are analysed and the results are shown in Fig. 5.2.

Fig. 5.2(a) provides the distribution of the numbers of users over the numbers of followers. It follows a power-law distribution. Among the 2,134 users in Darwin community (excluding the 1,158 isolated users), 1740 users (81.5%) have less than 10 followers, and 3 users have more than 500 followers.

Fig. 5.2(b) gives the distribution of social ties based on influence probabilities. The influence probabilities on more than 95% of edges are equal to or smaller than 0.1, which means most of social ties are weak ones in terms of influence. Only around 1% of edges are strong ties, whose influence probabilities are greater than 0.5.

Fig. 5.2(c) shows the relationship between influence index and follower count. It is believed that there is a strong positive correlation between influence power and the number of followers in online social networks. That is why the high-degree heuristic algorithm has been widely used in the study of the influence maximization problem and it is effective in a wide range of applications. In Section 5.5, we have provided our proposed *influence index* heuristic algorithm. In the social network of the Darwin community on Twitter, the influence index and the follower count are positively correlated with each other (Pearson Correlation Coefficient is 0.732).

Fig. 5.2(d) demonstrates the distribution of the 260 strong ties over follower

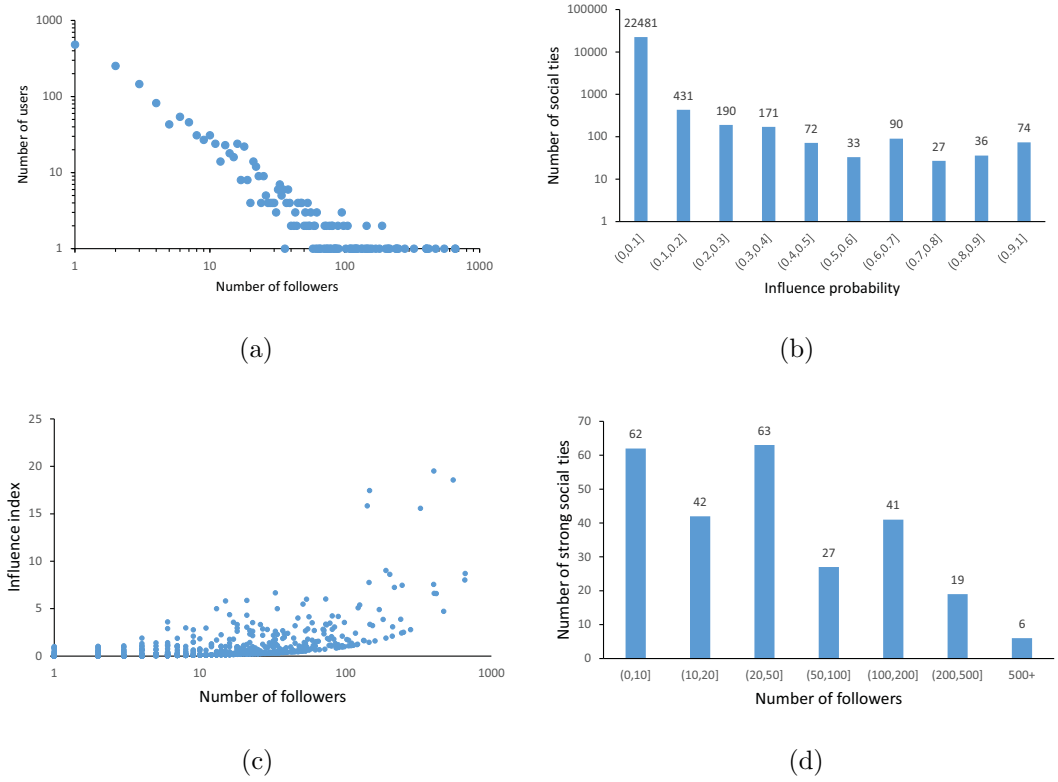

Figure 5.2: Social network analysis of the Darwin community on Twitter. (a)number of followers analysis; (b)social ties analysis based on influence probability; (c)relationship between influence index and number of followers; (d)strong social ties analysis based on number of followers.

counts. In the influence maximization problem, we focus on the strong social ties, since these linkages play a critical role in social influence propagation. The influential users may not have a large number of followers. Actually, most of the strong ties start from users with small numbers of followers. In other words, although those popular users have large population of followers, they are seldom involved in strong ties.

In the experiments for the five influence maximization algorithms, we simulate

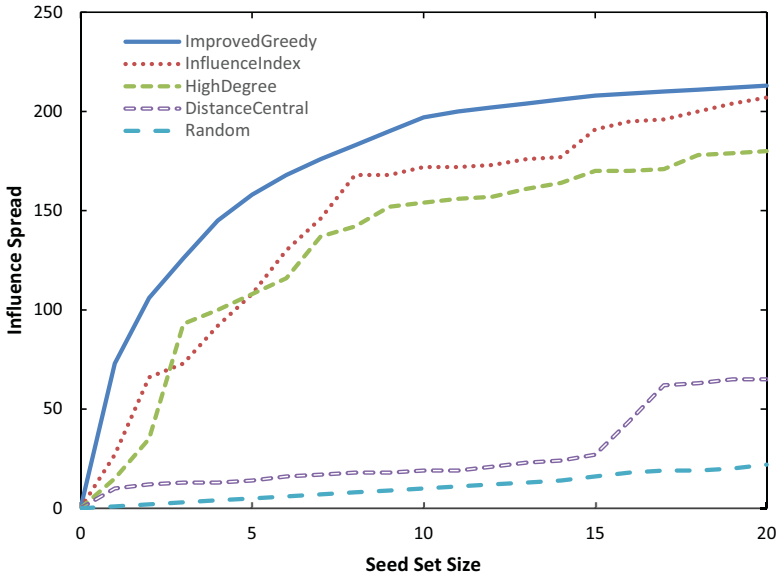

Figure 5.3: Influence spread achieved from the seed sets selected by different algorithms, with our proposed diffusion model

the diffusion process 100 times (i.e. set $r = 100$) based on the cascade model proposed in Section 5.4, and calculate the average of the results as the expected influence spread. Fig. 5.3 shows the influence spreads of these algorithms, with different seed set sizes ranging from 1 to 20. The performances of **HighDegree**, **DistanceCentral** and **Random** are similar to the experimental results in [28]. The simple **Random** algorithm is a baseline and performs quite poor. In the experiments of [28], the influence spread of **DistanceCentral** algorithm is close to that of **HighDegree** algorithm. **DistanceCentral** algorithm performs worse against our dataset. It is only slightly better than the **Random** algorithm. This means that the performance of **DistanceCentral** algorithm can change when different datasets and diffusion models are applied.

Our proposed **InfluenceIndex** algorithm and **HighDegree** algorithm can achieve significantly better influence spread than **DistanceCentral** algorithm. The overall performance of **InfluenceIndex** algorithm is better than that of **HighDegree** algorithm. The influence index is a more effective indicator for evaluating a user's influence than the number of followers. Compared with **ImprovedGreedy** algorithm, when seed set size is 20, the influence spreads of **InfluenceIndex** and **HighDegree** are 2.8% and 15.5% smaller respectively.

The curve for **ImprovedGreedy** algorithm becomes almost horizontal when the seed set size is bigger than 15. From the 16th seed node, the influence spread only increases by 1 when adding a new seed node. If we assume the cost of targeting a new seed user is equal to the profit of obtaining a new influenced user, it makes no sense to expand the seed set size after it reaches 15. The number 15 is approximately 0.5% of the total number of users (3,292) in our dataset. This percentage (0.5%) can be used as a benchmark for determining the seed set size in a social network.

Fig. 5.4 shows the influence spreads of five implemented algorithms based on the classic independent cascade model. In this cascade model, the independent probability on the directed edge (u, v) takes the same value of the accumulative influence probability $P_{u,v}$ defined in Section 5.3. Each activated user has only one chance to activate his inactive neighbours at the step right after himself is activated. Comparing Fig. 5.3 with Fig. 5.4, we can see that the influence

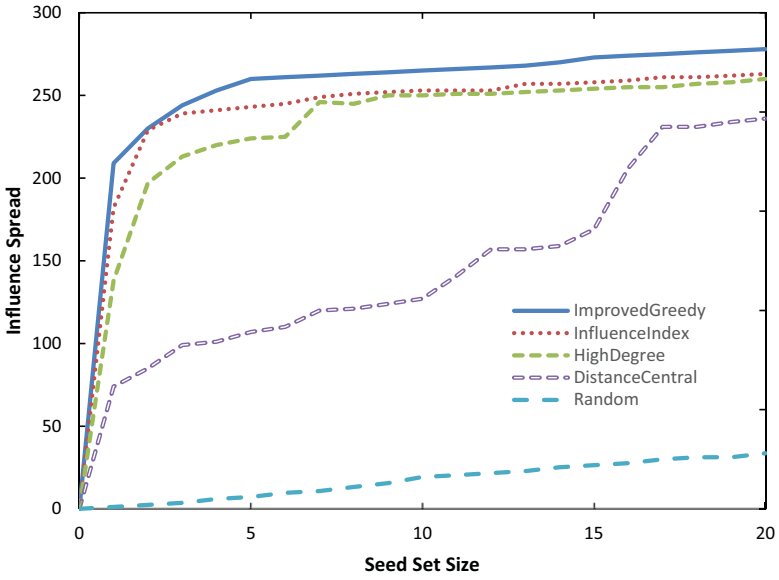

Figure 5.4: Influence spread achieved from the seed sets selected by different algorithms, with the classic independent cascade diffusion model

spreads with the classic independent cascade model are generally more than those with our proposed diffusion model. The reason is that, in the classic independent cascade model, there is a bigger probability for an active user to activate his neighbours with a single chance; in our proposed model, there is a time constraint (i.e. information validity period) in the diffusion process. Our proposed diffusion model reflects the real situations on Twitter better than the classic cascade model. We also observe that the effectiveness of the algorithms might be different between the two diffusion models. For example, **HighDegree** and **DistanceCentral** algorithms perform better with the classic independent cascade model. Our proposed **InfluenceIndex** algorithm achieves good results which are very close to those of **ImprovedGreedy** algorithm in both diffusion models.

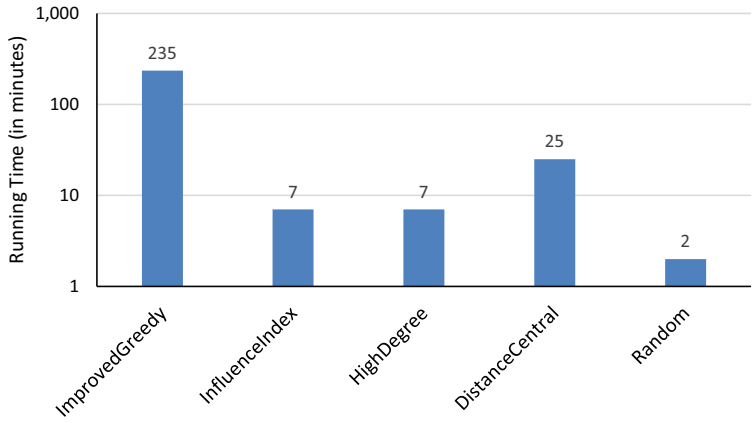

Figure 5.5: Running time (in minutes) for different algorithms, with our proposed diffusion model

Fig. 5.5 reports the running times of different algorithms with our proposed diffusion model when the seed set size is 20. Although **ImprovedGreedy** algorithm achieves the best influence spread, its running time is very long (nearly 4 hours). It is impractical to use **ImprovedGreedy** algorithm when dealing with large-scale social networks (such as a network with millions of user nodes). Comparing with **ImprovedGreedy** algorithm, our proposed **InfluenceIndex** algorithm can reduce the running time significantly (more than 30 times faster) and achieve a quite close influence spread. **DistanceCentral** algorithm takes a longer time because it is time-consuming to calculate users' average distance to others in the whole network.

Fig. 5.6 shows the running times of the five algorithms with the classic independent cascade model when the seed set size is 20. The results are very similar to Fig. 5.5, except that the running time of **ImprovedGreedy** and **Dis-**

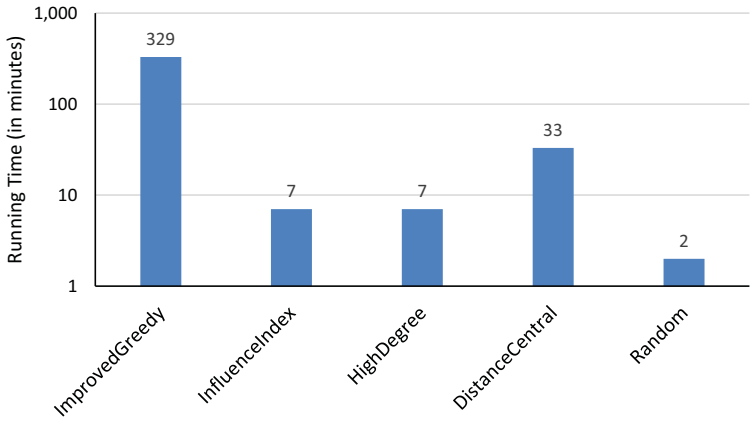

Figure 5.6: Running time (in minutes) for different algorithms, with the classic independent cascade diffusion model

tanceCentral algorithms is a little bit longer. Among various diffusion models, the seed selection algorithms themselves have no difference; the running time of the function that estimates the expected influence spread (such as Algorithm 1) varies.

5.7 Summary and Discussion

This chapter proposes an influence maximization approach with detailed description of influence probabilities, diffusion model, and heuristic algorithm in the context of Twitter. The proposed approach can cover a wide range of marketing campaign scenarios on Twitter. Influence probabilities are calculated based on users' action history. An information diffusion model is proposed to simulate the information spread on Twitter. The model covers the specific situation that a

user can have multiple chances to be influenced by others in a considered time period. A concise algorithm is developed based on heuristic principles.

A set of experiments are carried out with real Twitter data in Darwin (a city of Australia). We implement various algorithms with both our proposed *R-J cascade model* and classic *independent cascade model*. The influence spreads and running times of these algorithms are compared using the collected Twitter dataset. Experimental results show the effectiveness of the proposed influence maximization approach. The developed heuristic algorithm has the capability to achieve much better influence spread than existing heuristic-based solutions. Comparing with the well-known improved greedy algorithm, our algorithm can obtain close influence coverage but save about 97% running time.

In the future, we will develop methods to specify communities in online social networks and study influence maximization by utilising the specific features of these communities. We will recruit more large-sized datasets to analyse the characteristics of real-world network structures and investigate the scalability of seed selection algorithms.

6

Maximizing the Effectiveness of Advertising Campaigns on Twitter

6.1 Introduction

Nowadays, Online Social Networks (OSNs) play an important role in our daily life. People create, share, and exchange various information on these platforms. More and more companies start to utilize OSNs to spread their product/service information for marketing purpose. Social media advertising has made a great progress and development in a relatively short period of time. When Facebook launched its first advertising option in May 2005, no one could have predicted that social media advertising revenue reached 17.08 billion US dollars in 2015 [1], only ten years later.

There are mainly two ways for social media advertising as: (1) advertisers can

[1] http://www.statista.com/statistics/271258/facebooks-advertising-revenue-worldwide/

take the advantage of various users' information, including their interests, demographics (such as gender, age, race, level of education, etc.), and behaviours in the social networks, to deliver the advertisements directly to the target audiences; (2) advertisers can identify some influencers in the social networks as the seeds and propagate the advertisements through these seed users' social circles. In OSNs, comments and recommendations from friends, relatives, colleagues, schoolmates are normally trustworthy and they can affect people's feelings about advertised products and services. In this work, we focus on the second approach for advertising in social networks and study how to maximize the effectiveness of advertising campaigns on Twitter.

Motivated by the marketing applications, the problem of *Influence Maximization* was firstly proposed by Kempe et al. [28], and has attracted a lot of interest in the research field of online social networks. The goal of influence maximization is to find a set of most influential users in the social network so that the spread of influence (defined as the excepted number of influenced users) is maximized.

There are three important components when studying an influence maximization problem: (1) the influence probability model, which determines how influence probabilities between users are calculated; (2) the influence diffusion model, which reflects how influence propagates in the networks; (3) the seed nodes selection algorithm, which is used to select the influential users from the social network in order to maximize the expected influence spread.

Generally speaking, the diffusion model is more crucial in comparison with the algorithm for seed nodes selection in real applications. In many situations, the information propagation in the social network has specific characteristics that have not been studied in the existing diffusion models. For example, we consider an active user has multiple chances to influence his neighbors on Twitter. Different from most existing works that focus on the improvement of algorithm efficiency, this work develops a new diffusion model and utilizes the influence maximization techniques to support advertising campaigns on Twitter. Our main contributions can be summarized as follows:

- Utilizing the advertising theory from the marketing area, a specific influence maximization problem is identified for maximizing the effectiveness of advertising campaigns on Twitter.

- An influence probability model is proposed. The cumulative probabilities are calculated based on users' action history including tweet, favorite, reply and retweet. The probability at each step decays over time and the decay function is modelled based on the analysis of a real dataset from Twitter.

- An influence diffusion model is developed according to the major characteristics of advertising campaigns on Twitter. More specifically, this model inherits the classic *independent cascade model* and adopts two new assumptions: 1) a user can have multiple chances to influence his inactive neighbors;

2) a user can be influenced for multiple times, based on which the concept of *advertising effectiveness* is introduced.

The remainder of the chapter proceeds as follows. Section 6.2 describes the identified influence maximization problem. In Section 6.3, we propose a new influence probability model and diffusion model in the context of advertising on Twitter; discuss several popular algorithms and develop a heuristic algorithm for the selection of seed users. Section 6.4 provides the experimental results and discussion. Section 6.5 concludes the chapter and discusses the future work.

6.2 Problem Definition

The classical influence maximization problem is described as follows. A social network is represented by a directed graph $G = (V, E)$, where the nodes V represent users, and the directed edges E represent links between users. Each directed edge has a weight, representing the influence probability. There is also a budget k, which is an integer. The goal of influence maximization is to find k users (seed nodes) in the social network so that the expected influence spread (defined as the excepted number of influenced users) is maximized.

In this work, we identify a specific influence maximization problem based on real-life scenarios on Twitter. Suppose a company has released a new product or service. This company wants to launch a new advertising campaign on Twitter. They need to select some influential users (i.e. seed users) to help the propaga-

tion of the advertising information. The company expects that these seed users influence their followers, and then these followers influence their own followers, and so on. The goal of the advertising campaign is to maximize the expected influence spread through the online word-of-mouth effect within a budget.

For the identified influence maximization problem, we have the following assumptions:

- We interpret *influence* as "the ability to let someone know something, or pass information to others". A user u successfully influences user v if v gets the advertising information from u.

- The probability of u influencing v, which is denoted as $p_{u,v} \in (0, 1)$, will be calculated based on the *action history* on Twitter, including individual user's actions and the reactions from other users. We assume that the historical interactions between users are positively correlated with the influence probability.

- The information diffusion can be simulated as a process with discrete steps. An active user can make multiple attempts to influence his inactive followers, and the influence probability decays with time. This assumption reflects the real situation on Twitter: If a user posts an advertising message, his followers might get this information any time afterwards. But as time goes on, the likelihood of seeing this message decreases, since the tweet

moves down in user's timeline when new tweets come in. More details will be provided in Section 6.3.1.

- We adopt the concept of *effective frequency* from marketing theory, and assume that a user can be influenced for multiple times. We believe that the effect of influence is different when a user get the same message on Twitter for multiple times. A new metric *advertising effectiveness* is introduced. The expected influence spread is not the number of activated users but the sum of advertising effectiveness for all activated users. More details will be discussed in Section 6.3.2.

In this work, we utilize influence maximization techniques to support the development of advertising campaigns on Twitter. New models for influence probability and information diffusion will be proposed based on the above assumptions.

6.3 Influence Maximization Method

6.3.1 Influence Probability Model

We present a Twitter social network as a directed graph $G = (V, E)$. If the user v follows the user u, there is a directed edge $(u, v) \in E$, which is labelled with a value $p_{u,v}$ $(0 \leq p_{u,v} \leq 1)$, representing the influence probability with which u will succeed in activating his follower v. Here we propose an improved model based on the work in Section 5.3, by incorporating important temporal features in the dynamics of influence diffusion.

The action history contains the records of individual users' actions and reactions from other users. Due to capacity constraints of Twitter APIs [1], not all the historical data are available. We collect data in a recent period of time and assume that the actions/reactions happened during this period can reflect the pairwise user influencing situation in the future.

Let A_u denote the total number of actions the user u performs, i.e. the number of tweets the user u posts. T denotes the set of reaction types. $R(u, v, a)$ is the number of reaction $a \in T$ with which user v reacts to user u. We consider three kinds of reactions (denoted by T) as *favorite*, *reply*, and *retweet*.

If user v reacts to user u, it means u has successfully passed the advertising information to v (u has influenced v). The *influence factor* ($infl(u, v) \geq 0$) from user u to v is defined as the ratio of v's reactions to u's actions.

$$infl(u, v) = \sum_{a \in T} R(u, v, a) \ / \ A_u \tag{6.1}$$

The cumulative influence probability $P_{u,v}$ is calculated from the *influence factor* with the following empirical formula.

$$P_{u,v} = 1 - \exp(-infl(u, v)) \tag{6.2}$$

Here $0 \leq P_{u,v} < 1$. If user v has no historical reaction towards u, it means u has no influence on v ($P_{u,v} = 0$). If the ratio of v's reactions to u's actions is

[1] https://dev.twitter.com/overview/api

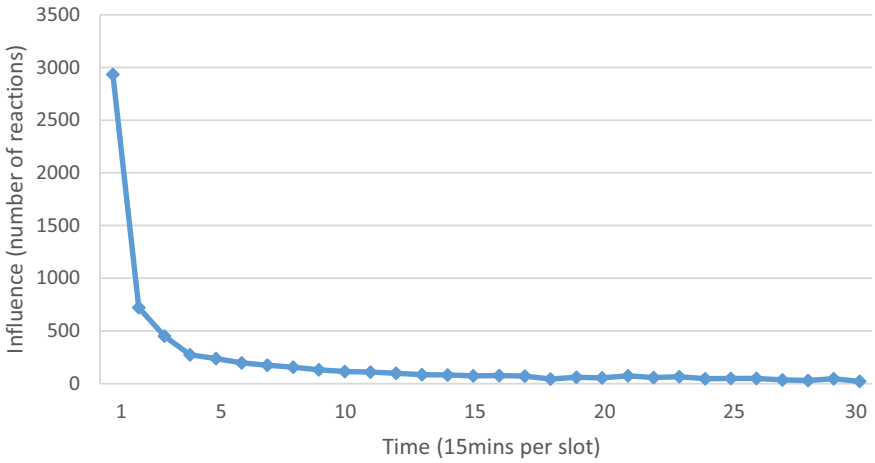

Figure 6.1: The distribution of users' reactions over time

very high, it means u has great influence on v ($P_{u,v}$ is close to 1). Based on the interactions between users, this formula captures the major characteristics of the cumulative influence probability.

Considering the characteristic of timeline on Twitter, we can imagine that the influence of a piece of advertising information (i.e. a tweet) will decay with time. In order to get an empirical influence decay function $f(\cdot)$, we analyze the time attributes of users' reactions in a real Twitter dataset, which includes the data of the *Darwin* social network during one month (more explanation in Section 6.4.1). We take every 15 minutes as one timeslot, and Fig.6.1 shows the distribution of users' reactions over time. Only the first 30 timeslots are listed because over 85% of reactions happen during this period. The numbers of reactions in timeslots decreases rapidly over time, so we ignore the reactions after the 30th timeslot.

We utilize the curve fitting toolbox in MATLAB to get the influence decay

function as:

$$f(t) = \exp(\alpha t), \alpha = -1.15 \tag{6.3}$$

Exponential function is chosen since the RMSE (Root Mean Square Error) reaches the minimum comparing with other models, and it also accords with our intuition. We assume that there is at least a small influence probability $P_{u,v} = 0.01$ when user v follows user u. This means $P_{u,v} \geq 0.01$ if user v follows user u. This constant value 0.01 has been used in [28, 29]. We consider the influence diffusion as a process with discrete steps. An active user makes one attempt to activate his inactive neighbors at each step. The influence probability of the active user to his neighbors will decay with time. The user can only have effective influence to his neighbors within a limited period of time (we consider N steps in this work). $P_{u,v}$ is calculated as follows.

$$P_{u,v} = 1 - \prod_{t=1}^{N}(1 - p_{u,v}(t)) \tag{6.4}$$

where $p_{u,v}(t)$ is the probability of u influencing v at the time t; when u is activated, $t = 0$; and afterwards $t \in \{1, 2, 3, ..., N\}$. Based on Eq. 6.3, $p_{u,v}$ is:

$$p_{u,v}(t) = \beta_{u,v} \exp(\alpha t) \tag{6.5}$$

where $\beta_{u,v}$ is the tuning parameter for the user pair (u, v). Based on the discussion above, $\beta_{u,v}$ can be computed from the cumulative influence probability

$P_{u,v}$, and then the probability function of user u influencing v, $p_{u,v}(t)$, can be obtained.

6.3.2 Influence Diffusion Model

This subsection proposes an Advertising Independent Cascade Diffusion Model (referred to as *Ad-ICDM* in the following part of this chapter) to capture the major characteristics of advertising information spread on Twitter.

The classic Independent Cascade diffusion model assumes that an active user can make only one attempt to activate his neighbors and an activated user will not accept any further activation/influence from other users. The diffusion process stops until every active node has tried its single chance and there are no more activations. The *Ad-ICDM* inherits the classic independent cascade model and has two major modifications based on the assumptions discussed in Section 6.2.

Firstly, a user u has the multiple chances to activate his inactive neighbor v. In the context of Twitter, once u posts a tweet, his follower v can obtain this information any time, which could be after 1 minute or after 10 days. Based on our data analysis in Section 6.3.1, we assume a user will attempt to influence his neighbors within N steps after he is activated, and the influence probabilities decay with time.

Secondly, the metric *advertising effectiveness* is introduced in this work. Suppose both u_1 and u_2 have a follower v. After v is activated (or to say, successfully influenced) by u_1, v can still be influenced by u_2 in the future time. In the mar-

keting theory, it is widely believed that messages have more effective influence if they have been repeated multiple times. The concept *effective frequency* was introduced by Naples [120]. AMA (American Marketing Association) Dictionary defines it as "An advertiser's determination of the optimum number of exposure opportunities required to effectively convey the advertising message to the desired audience or target market." [1]

Sawyer proposed the habituation-tedium theory [121], which is based on Berlyne's two-factor model [122]. This theory suggests a two-stage process that governs response to repeated messages. The first stage (wear-in) is related to habituation, and the second stage (wear-out) is connected to tedium. When consumers are first exposed to novel advertising stimuli, they experience uncertainty and tension. Repeated exposure reduces the apprehension through habituation, which initially leads to more positive response. However, as the number of exposures exceeds a certain level, boredom and resentment set in, and attitude toward the advertisement as well as response diminish; the two forces lead to an inverted U-shaped relationship between the number of exposures and advertisement response.

There are also numerous studies with their own theories or models as to what the correct number is for effective frequency. Thomas Smith, a London businessman, wrote a guide called *Successful Advertising* in 1885 [123]. The sayings

[1] https://www.ama.org/resources/Pages/Dictionary.aspx

Thomas Smith, Successful Advertising. 1885.

The first time people look at any given ad, they don't even see it.
The second time, they don't notice it.
The third time, they are aware that it is there.
The fourth time, they have a fleeting sense that they've seen it somewhere before.
The fifth time, they actually read the ad.
The sixth time they thumb their nose at it.
The seventh time, they start to get a little irritated with it.
The eighth time, they start to think, "Here's that confounded ad again."
The ninth time, they start to wonder if they're missing out on something.
The tenth time, they ask their friends and neighbors if they've tried it.
The eleventh time, they wonder how the company is paying for all these ads.
The twelfth time, they start to think that it must be a good product.
The thirteenth time, they start to feel the product has value.
The fourteenth time, they start to remember wanting a product exactly like this for a long time.
The fifteenth time, they start to yearn for it because they can't afford to buy it.
The sixteenth time, they accept the fact that they will buy it sometime in the future.
The seventeenth time, they make a note to buy the product.
The eighteenth time, they curse their poverty for not allowing them to buy this terrific product.
The nineteenth time, they count their money very carefully.
The twentieth time prospects see the ad, they buy what is offering.

Figure 6.2: Successful Advertising theory by Thomas Smith

he used are still being used today and form the foundation for the theory of frequency in advertising and marketing, which has essentially become the authoritative guide for generating top-of-mind awareness. As can be seen from Fig. 6.2, it takes a minimum of 20 impressions to develop top-of-mind awareness and to generate a sale.

According to the famous Krugman's three-exposure theory [124], there are only three levels of exposure in psychological terms: curiosity, recognition and decision. On first exposure, consumers respond to the advertisement by asking: "What is it?" The second exposure triggers a response: "What of it?" The

third exposure is a reminder and evokes a decision: "I will (or will not) buy it." Krugman argues that all subsequent exposures are just reminders, similar to the third response. In this work, we adopt Krugman's three-exposure theory in our model, and define *advertising effectiveness factor* which has three values: 0.5, 0.8, 1.0. When a user is activated (i.e. influenced for the first time), the *advertising effectiveness* on this user is 0.5. When this user is influenced for the second time, the *advertising effectiveness* increases to 0.8. Then for the third time, it will reach to 1.0. The further additional influence (or exposure of the advertisement) on this user will be ignored, since we assume the effectiveness has reached its maximum and will not increase again after that.

For convenience, important variables used in this work are listed in Table 6.1. And Table 6.2 provides the comparison between the classic *IC* model and the proposed *Ad-ICDM* model. The two models differ in the number of attempts allowed, the times one can be influenced, the maximization objectives, and the termination conditions for the diffusion process.

The algorithm for estimating the expected influence spread with *Ad-ICDM* model is provided as Algorithm 3. The influence diffusion process terminates until there are no more activations after N consecutive steps (Line 7). A matrix records the number of attempts between two connected users (Lines 5, 12-14). This number is used to calculate the influence probability at each step. The value of *advertising effectiveness* is updated based on the *advertising effectiveness factor*

Table 6.1: Variables used in this work

Variables	Description
SS	set of seed users
s	expected influence spread
N	number of attempts an activated user can make to activate his neighbors
r	rounds of simulation for diffusion
k	number of seed users to be selected
$P_{u,v}$	cumulative probability of user u influencing v
$p_{u,v}(t)$	probability of user u influencing v at the time t

Table 6.2: Classic IC Model vs Ad-ICDM Model

	Classic IC model	Ad-ICDM model
How many attempts can one user make to activate the neighbors?	Only one time	Multiple times
How many times can one user be activated / influenced?	Only one time	Multiple times
What is the objective to be maximized in the problem?	The number of the activated users	The sum of advertising effectiveness for all activated users
When does the diffusion process terminate?	There are no activations at one step	There are no activations in N consecutive steps

mentioned before (Lines 16-28). Finally, we compute the average of the sum of *advertising effectiveness* in the r simulations as the expected influence spread

result.

6.3.3 Algorithms for Influence Maximization

Influence Maximization is the problem of selecting a set of seed users to spread the influence as much as possible. Kempe et al. [28] prove the influence maximization problem is *NP-hard* and develop a greedy algorithm for the selection of seed users. This greedy algorithm has obtained a good result for expected influence spread comparing with existing approximation algorithms, but it is computationally expensive and it is unrealistic to apply the algorithm when solving problems in large social networks. In order to address this issue, a lot of efforts have been made to improve its efficiency.

Leskovec et al. [24] develop an efficient approximation algorithm CELF (Cost-Effective Lazy Forward selection) by exploiting the submodularity property of the influence function. The submodularity property means that incremental influence spread from adding a user u to a set S is equal to or greater than the incremental influence spread from adding the same user to a superset of S, which can be denoted as: $f(S \cup \{u\}) - f(S) \geq f(W \cup \{u\}) - f(W)$, when $S \subseteq W$. In our proposed *Ad-ICDM* diffusion model, a user can be influenced for multiple times and the objective is to maximize the overall advertising effectiveness. The defined advertising effectiveness satisfies the "diminishing returns" property: the marginal gain of influencing a user for the second time is 0.3 and the one for the third time is 0.2. This "diminishing returns" property guarantees the influence

Algorithm 3 Estimate the expected influence spread with Ad-ICDM model

Input: SS, r, N, $p_{u,v}(t)$
Output: s

$s \leftarrow 0$
for $i \leftarrow 1, r$ **do**
    reset the activated set $AS \leftarrow SS$
    reset the array of advertising effectiveness $ae \leftarrow 0.0$
    reset the matrix of numbers of tries $nt \leftarrow 0$
    set the number of steps without activation $ns \leftarrow 0$
    while $ns < N$ **do**
        set a flag of successful activation $f \leftarrow 0$
        for each user $u \in AS$ **do**
            find the follower set FS of user u
            for each user $v \in FS$ **do**
                if $nt[u, v] < N$ **then**
                $nt[u, v] \leftarrow nt[u, v] + 1$
                $t \leftarrow nt[u, v]$
                generate a random value $x \in (0, 1)$
                if $x < p_{u,v}(t)$ **then**
                if $ae[v] == 0$ **then**
                $f \leftarrow 1$
                $AS \leftarrow AS \cup \{v\}$
                $ae[v] \leftarrow 0.5$
                end if
                if $ae[v] == 0.5$ **then**
                $ae[v] \leftarrow 0.8$
                end if
                if $ae[v] == 0.8$ **then**
                $ae[v] \leftarrow 1.0$
                end if
                end if
                end if
            end for
        end for
        if $f == 0$ **then**
            $ns \leftarrow ns + 1$
        else
            $ns \leftarrow 0$
        end if
    end while
    calculate the sum sae in the array ae
    $s \leftarrow s + sae$
end for
$s \leftarrow s/r$

function is submodular. With the similar way to utilize the submodularity in CELF, we develop an improved greedy algorithm based our probability model and diffusion model.

Various of heuristic algorithms have been developed to improve the efficiency when solving the influence maximization problems. The *high-degree* and *distance centrality* are two popular ones. The *high-degree* heuristic algorithm selects the seed users based on their out-degrees, i.e. the numbers of followers on Twitter. Intuitively, how many followers a user has in a social network is an important indicator to evaluate the user's influence. Experimental results provided in [28] show that the high-degree heuristic algorithm can achieve the influence spread close to the greedy algorithm, outperforming several other algorithms. The *distance centrality* is another commonly used influence measure. A set of concepts associated with the centrality were discussed in [125]. This heuristic algorithm [28, 29] selects the most influential nodes based on their network position. The more central a node is, the shorter its total distance from all other nodes. Nodes at more central positions are considered more influential and they are selected as seed nodes.

In the proposed *Ad-ICDM* diffusion model, influence probability is the indicator which reflects a user's capability to influence others. The influence probability at each step of propagation is calculated based on the cumulative influence probability between each pair of users. We define a user's *influence index* as the sum

of the cumulative influence probabilities from the user to others, and believe that *influence index* can effectively represent the overall influence power of a user. For example, user u has only one outgoing linkage with an influence probability 50%, user v has five outgoing linkages with an influence probability 10% on each edge. In this case, user u and v are considered to have equivalent influence power. Users with bigger *influence index* values will be selected as seed users.

6.4 Experiments and Analysis

6.4.1 Experiments Setup

We choose a social community on Twitter based on geographic location for our experiments. We capture all the users who claim *Darwin* as the location in their profiles. *Darwin* is the capital city of the Northern Territory in Australia. The selected dataset of the Twitter social network at *Darwin* includes 3,292 users from this city on July 21, 2015. There are totally 23,605 following relationships between users (i.e. directed edges in the network graph) in the *Darwin* community.

The calculation of the influence probability uses 30 steps as the parameter ($N = 30$) when activating his inactive neighbors (see the discussion in Section 6.3.1). A user can attempt to activate his inactive neighbors within the following 30 steps after the user has been activated. Monte Carlo simulation is the widely used way to estimate the influence spread [28, 29, 94]. We simulate the diffusion process for 100 times (i.e. set $r = 100$) and use the average value of the simulation

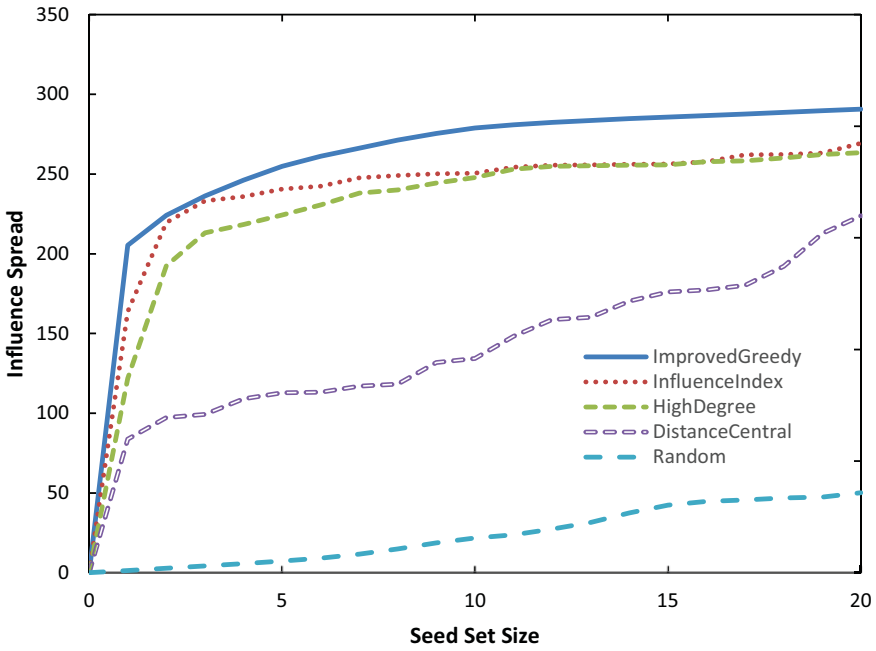

Figure 6.3: Expected influence spread by different algorithms

results as the expected influence spread. The influence maximization goal is to find k seed users ($k = 1, 2, ..., 20$) in order to maximize the expected influence spread.

We compare the performance of different seed selection algorithms mentioned in Section 6.3.3 with the proposed *Ad-ICDM* diffusion model. Furthermore, we use the *Random* algorithm as one of baselines. This algorithm simply selects the seed users randomly from the social network.

6.4.2 Results and Discussion

Fig. 6.3 shows the influence spreads of the five seed selection algorithms (seed set sizes from 1 to 20) with our proposed *Ad-ICDM* diffusion model. As expected,

the performance of **Random** algorithm is very poor. The expected influence spread is just a little bigger than the number of seed users, which means most of the users have little influence. **DistanceCentral** algorithm performs much better than **Random** algorithm but still obviously worse than the other three algorithms. The performance of **InfluenceIndex** algorithm is better than that of **HighDegree** algorithm, but they are quite close, especially when the seed set size is bigger than 10. When the seed set size is 20, **ImprovedGreedy** algorithm outperforms **InfluenceIndex** algorithm by 7.9% and **HighDegree** algorithm by 10.4%. The performance ranking of **Greedy**, **HighDegree**, **DistanceCentral** and **Random** algorithms in the experimental results of [28] and [29] is the same as our results. This implies that the performance of these algorithms is relatively stable across different datasets and diffusion models. Comparing with Fig. 5.3, the expected influence spread with *Ad-ICDM* model is generally more than the influence spread with *R-J cascade* model, because a user can be influenced for multiple times in *Ad-ICDM* model.

Fig. 6.4 reports the running times of different algorithms with the proposed *Ad-ICDM* diffusion model when the seed set size is 20. Although **ImprovedGreedy** algorithm can achieve the best result for the influence spread, its computational cost is very high (nearly 17 hours). It is impractical to use **ImprovedGreedy** algorithm in large-scale social networks. Comparing with **ImprovedGreedy** algorithm, our proposed **InfluenceIndex** algorithm can reduce

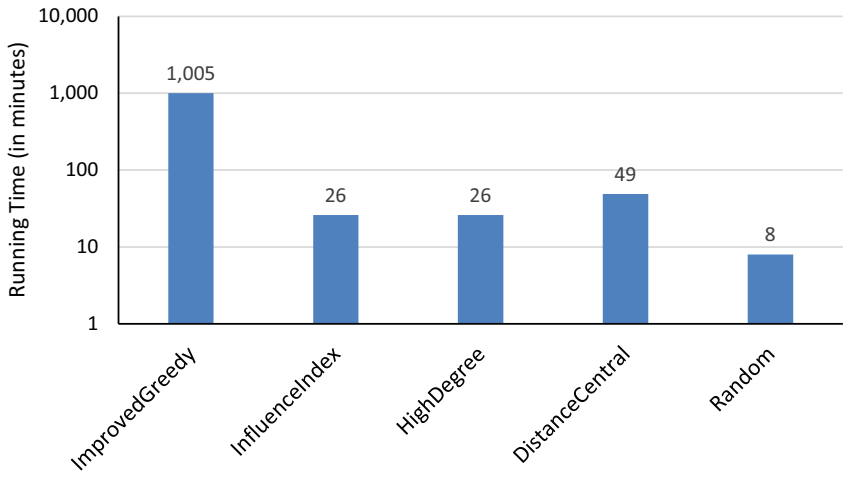

Figure 6.4: Running time (in minutes) for different algorithms

the running time significantly (about 40 times faster) and obtain a quite close influence spread. **DistanceCentral** algorithm takes a longer time because it is time-consuming to calculate each user's distance to others in the whole network.

6.5 Summary and Discussion

This chapter proposes a specific influence maximization problem for maximizing the effectiveness of advertising campaigns on Twitter. A new influence probability model and diffusion model have been proposed. Comparing with existing works, these models can better reflect the real situations of advertising information spread on Twitter. The cumulative probabilities are calculated according to users' action history. The probability at each step in the propagation decays exponentially as the experiment indicates. The proposed diffusion model inherits the classic independent cascade model and modifies two assumptions according

to the major characteristics of information propagation on Twitter. We introduce the *advertising effectiveness* as the maximization objective in the research problem. Several different algorithms are evaluated with the *Ad-ICDM* diffusion model.

In the future, more large-sized datasets will be used to analyze how information propagates in other real social networks. How to utilize influence maximization techniques to solve the practical problems (e.g. viral marketing) based on the characteristics of the network structure and user interactions in different social networks is a real challenge. Furthermore, we will develop methods to identify the communities in terms of interests or topics, and study the influence maximization problem in these communities.

7

Conclusion

7.1 Summary

The rapid growth of online social networks and social media has attracted much attention in online social influence research. Many applications like viral marketing, recommender systems, events detection, community detection, expert finding, link prediction and epidemics on networks can benefit from social influence analytics. In the past decade, considerable research has been conducted on the measurement of online social influence. However, most existing studies directly utilize their own pre-defined features to build the model without a pre-evaluation process for these selected features. There is a lack of comprehensive analysis regarding the effectiveness of the principal features for measuring user influence. One of the fundamental problems in the study of social influence is Influence Maximization. When we are solving a specific influence maximization problem, there is still a gap between the traditional solution framework and the real world

situation. The objective of this dissertation is to: (1) address the issue of feature selection for measuring social influence on Twitter; (2) specify an influence maximization problem on Twitter and propose our approach to tackle this problem.

In Chapter 2, we provide the background knowledge and review the related work in the research area of social influence. We start from the definition of influence and the types of influencers. Classic methods of feature selection are reviewed and some existing models for measuring user influence on Twitter are discussed. Furthermore, we discuss the hot research topic - *Influence Maximization*, and specifically review several popular influence diffusion models and seed selection algorithms.

In Chapter 3, we analyze the principal features for measuring user influence on Twitter. Both manifest features and social attributes (hidden features) are investigated. We employ *Entropy* method and *Spearman's Rank Correlation Analysis* to identify the major manifest features for measuring user influence on Twitter. We extract the latent features by *Principal Component Analysis (PCA)*, map these features to social attributes, and identify the principal social attributes by *Stepwise Multiple Linear Regression (SMLR)*. Our study reveals a number of novel findings as follows: (i) Firstly, besides *mention* and *retweet* actions that have been widely used to measure user influence in literature, we find that *number of public lists*, *new tweets*, *follower to friends ratio* are also fairly effective indicators for user influence; (ii) We further discover that *popularity*, *engagement* and

authority are the three most important social attributes to drive user influence in Twitter environment; (iii) Finally, we compare four popular influence scoring services, and find that *new mentions* and *number of public lists* are the two most effective manifest features for their influence ranking, and *popularity* is commonly considered as the first key social attribute of the influencers on Twitter.

In Chapter 4, we propose a hybrid feature selection method for predicting user influence on Twitter. A set of candidate features from Twitter is identified based on the five attributes of influencers defined in sociology. Firstly, less relevant features are filtered out with a feature-weighting algorithm. Then the *Sequential Backward Floating Selection (SBFS)* is utilized as the search strategy, a *Back Propagation Neural Network (BPNN)* is employed to evaluate the feature subset at each step of searching. Finally, an optimal feature set is obtained for predicting user influence with a high degree of accuracy. Experimental results are provided based on a real world Twitter dataset including seven million tweets associated with 200 popular users in Australia. The proposed method can provide a set of features that could be used as a solid foundation for studying complicated user influence evaluation and prediction.

In Chapter 5, we study an influence maximization problem on Twitter, i.e., selecting a set of seeds to maximize the information propagation, which can be used for information reaching out in marketing campaigns. The proposed approach is taking into the consideration of social ties, user interactions, and in-

formation propagation on Twitter. The influence probability is calculated according to users' action history including tweet, favourite, mention/reply, and retweet. An information diffusion model is proposed with the capability to simulate the dynamic process of information spread on Twitter. A concise heuristic algorithm (*influence index*) is developed for influence maximization accordingly. Experimental results and analysis are provided based on a real Twitter network including 3,292 users in Darwin city in Australia.

In Chapter 6, we focus on a specific influence maximization problem, selecting a set of seed users to maximize the effectiveness of advertising campaigns on Twitter. With this problem, the information diffusion model must have the capability to support: (a) an active user can make multiple attempts to activate his neighbors; and (b) a user can accept an advertising message many times. There are two major modifications comparing with the research work in Chapter 5. Firstly, when calculating influence probabilities, we incorporate important temporal features in the dynamics of influence diffusion. Secondly, we adopt the concept of *effective frequency* from marketing theory, and assume that a user can be influenced for multiple times. An influence diffusion model (*Ad-ICDM*) is developed and a new metric *advertising effectiveness* is defined as the maximization objective. Several existing seed selection algorithms are analyzed based on the proposed diffusion model against a real dataset from Twitter. Experimental results are provided to show the soundness of the proposed model.

7.2 Future Work

This dissertation presents our research progress in the study of social influence on Twitter. There are still a lot of opportunities for the future research.

Firstly, due to the limits of Twitter APIs and the time constraints, the datasets used in our experiments are relatively small. We should recruit more large-sized datasets, in order to analyse the characteristics of real-world social networks and evaluate the scalability of our proposed approach.

Secondly, when we are talking about *influence* in real life, usually it is with regards to a specific context. For example, a person probably won't ask a university professor for advice on how to plan a trip. If a famous doctor recommends both a new medicine and a new song in the online social network, which message seems more convincing? It is unlikely that one person has a great influence in all areas. It is an interesting work to develop methods to identify the communities in terms of interests or topics, and study the influence maximization problem within these communities.

Finally, this work studies the influence maximization problem in a specific online social network, i.e. Twitter. We are curious about how to apply our approach in other online social networks, such as Facebook, Google+, etc. What are the different characteristics of influence propagation in other networks? What kind of diffusion models are able to capture these characteristics and to better reflect the real situations? It is exciting to study some well-defined influence

maximization problems which can truly support the marketing decisions in real life.

References

[1] Brian Winston. *Media technology and society: a history: from the telegraph to the Internet*. Psychology Press, 1998. 1

[2] Douglas E Comer. *The Internet book: everything you need to know about computer networking and how the Internet works*. Prentice-Hall, Inc., 2000. 1

[3] Duncan Brown and Nick Hayes. *Influencer marketing: Who really influences your customers?* Routledge, 2008. 3

[4] Walter J Carl. **What's all the buzz about? Everyday communication and the relational basis of word-of-mouth and buzz marketing practices**. *Management Communication Quarterly*, **19**(4):601–634, 2006. 4

[5] Bertram H Raven. **Social influence and power**. Technical report, DTIC Document, 1964. 4, 17, 105

[6] Morton Deutsch and Harold B Gerard. **A study of normative and informational social influences upon individual judgment.** *The journal of abnormal and social psychology*, **51**(3):629, 1955. 4

[7] John RP French Jr. **A formal theory of social power.** *Psychological review*, **63**(3):181, 1956. 4

[8] Herbert C Kelman. **Compliance, identification, and internalization: Three processes of attitude change**. *Journal of conflict resolution*, pages 51–60, 1958. 4

[9] John RP French, Bertram Raven, and D Cartwright. **The bases of social power**. *Classics of organization theory*, pages 311–320, 1959. 4

[10] Peter Michael Blau. *Exchange and power in social life*. Transaction Publishers, 1964. 4

[11] Carl I Hovland, Irving L Janis, and Harold H Kelley. **Communication and persuasion; psychological studies of opinion change.** *Yale University Press*, 1953. 4

REFERENCES

[12] Al Mamunur Rashid, George Karypis, and John Riedl. **Influence in ratings-based recommender systems: An algorithm-independent approach**. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 556–560. SIAM, 2005. 6

[13] Xiaodan Song, Belle L Tseng, Ching-Yung Lin, and Ming-Ting Sun. **Personalized recommendation driven by information flow**. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 509–516. ACM, 2006. 6

[14] Shang Shang, Pan Hui, Sanjeev R Kulkarni, and Paul W Cuff. **Wisdom of the crowd: Incorporating social influence in recommendation models**. In *2011 IEEE 17th International Conference on Parallel and Distributed Systems (ICPADS)*, pages 835–840. IEEE, 2011. 6

[15] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. **Earthquake shakes Twitter users: real-time event detection by social sensors**. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010. 6

[16] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. **Emerging topic detection on twitter based on temporal and social terms evaluation**. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, page 4. ACM, 2010. 6

[17] Lei Tang and Huan Liu. **Community detection and mining in social media**. *Synthesis Lectures on Data Mining and Knowledge Discovery*, **2**(1):1–137, 2010. 7

[18] Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. **Cascade-based community detection**. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 33–42. ACM, 2013. 7

[19] Wen Dong and Alex Pentland. **Modeling influence between experts**. In *Artifical Intelligence for Human Computing*, pages 170–189. Springer, 2007. 7

[20] Jie Tang, Jimeng Sun, Chi Wang, and Zi Yang. **Social influence analysis in large-scale networks**. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 807–816. ACM, 2009. 7

[21] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. **Group formation in large social networks: membership, growth, and evolution**. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54. ACM, 2006. 7

[22] David Crandall, Dan Cosley, Daniel Huttenlocher, Jon Kleinberg, and Siddharth Suri. **Feedback effects between similarity and social influence in**

online communities. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 160–168. ACM, 2008. 7

[23] Neil Zhenqiang Gong, Lester Mackey, Ameet Talwalkar, Ling Huang, Dawn Song, Eui Chul Richard Shin, Emil Stefanov, et al. **Predicting links and inferring attributes using a social-attribute network (san)**. Technical report, 2011. 7

[24] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. **Cost-effective outbreak detection in networks**. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 420–429. ACM, 2007. 7, 41, 111, 139

[25] Ed De Quincey and Patty Kostkova. **Early warning and outbreak detection using social networking websites: The potential of Twitter**. In *International Conference on Electronic Healthcare*, pages 21–24. Springer, 2009. 7

[26] Eben Kenah and James M Robins. **Second look at the spread of epidemics on networks**. *Physical Review E*, **76**(3):036113, 2007. 7

[27] Pieter Trapman. **On analytical approaches to epidemics on networks**. *Theoretical population biology*, **71**(2):160–173, 2007. 7

[28] David Kempe, Jon Kleinberg, and Éva Tardos. **Maximizing the spread of influence through a social network**. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003. 7, 8, 9, 10, 11, 31, 32, 39, 40, 44, 107, 111, 112, 119, 126, 133, 139, 141, 142, 144

[29] Wei Chen, Yajun Wang, and Siyu Yang. **Efficient influence maximization in social networks**. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208. ACM, 2009. 8, 31, 107, 112, 133, 141, 142, 144

[30] Kazumi Saito, Ryohei Nakano, and Masahiro Kimura. **Prediction of information diffusion probabilities for independent cascade model**. In *Knowledge-based intelligent information and engineering systems*, pages 67–75. Springer, 2008. 8, 102

[31] Amit Goyal, Francesco Bonchi, and Laks VS Lakshmanan. **Learning influence probabilities in social networks**. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 241–250. ACM, 2010. 8, 102

[32] Wei Chen, Chi Wang, and Yajun Wang. **Scalable influence maximization for prevalent viral marketing in large-scale social networks**. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1029–1038. ACM, 2010. 9, 31, 42

[33] Wei Chen, Yifei Yuan, and Li Zhang. **Scalable influence maximization in social networks under the linear threshold model**. In *2010 IEEE 10th International Conference on Data Mining (ICDM)*, pages 88–97. IEEE, 2010. 9

[34] Wei Chen, Alex Collins, Rachel Cummings, Te Ke, Zhenming Liu, David Rincon, Xiaorui Sun, Yajun Wang, Wei Wei, and Yifei Yuan. **Influence Maximization in Social Networks When Negative Opinions May Emerge and Propagate.** In *SDM*, **11**, pages 379–390. SIAM, 2011. 9, 32

[35] Wei Lu and Laks VS Lakshmanan. **Profit maximization over social networks**. In *2012 IEEE 12th International Conference on Data Mining*, pages 479–488. IEEE, 2012. 9, 32

[36] Bo Liu, Gao Cong, Dong Xu, and Yifeng Zeng. **Time constrained influence maximization in social networks**. In *2012 IEEE 12th International Conference on Data Mining (ICDM)*, pages 439–448. IEEE, 2012. 9, 36

[37] Wei Chen, Wei Lu, and Ning Zhang. **Time-critical influence maximization in social networks with time-delayed diffusion process.** In *AAAI*, **2012**, pages 1–5, 2012. 9, 38

[38] Huiyuan Zhang, Thang N Dinh, and My T Thai. **Maximizing the spread of positive influence in online social networks**. In *2013 IEEE 33rd International Conference on Distributed Computing Systems (ICDCS)*, pages 317–326. IEEE, 2013. 9

[39] Wei Lu, Francesco Bonchi, Amit Goyal, and Laks VS Lakshmanan. **The bang for the buck: fair competitive viral marketing from the host perspective**. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 928–936. ACM, 2013. 9

[40] Wei Lu, Wei Chen, and Laks VS Lakshmanan. **From competition to complementarity: comparative influence diffusion and maximization**. *Proceedings of the VLDB Endowment*, **9**(2):60–71, 2015. 9, 38

[41] Guangmo Tong, Weili Wu, Shaojie Tang, and Ding-Zhu Du. **Adaptive influence maximization in dynamic social networks**. *IEEE/ACM Transactions on Networking (TON)*, **25**(1):112–125, 2017. 9, 31, 38

[42] Yadong Qin, Jun Ma, and Shuai Gao. **Efficient influence maximization under TSCM: a suitable diffusion model in online social networks**. *Soft Computing*, **21**(4):827–838, 2017. 9

[43] Yan Mei, Youliang Zhong, and Jian Yang. **Finding and analyzing principal features for measuring user influence on Twitter**. In *2015 IEEE First International*

Conference on Big Data Computing Service and Applications (BigDataService), pages 478–486. IEEE, 2015. 15

[44] Yan Mei, Zizhu Zhang, Weiliang Zhao, Jian Yang, and Robertus Nugroho. **A Hybrid Feature Selection Method for Predicting User Influence on Twitter**. In *International Conference on Web Information Systems Engineering*, pages 478–492. Springer, 2015. 15

[45] Yan Mei, Weiliang Zhao, and Jian Yang. **Influence maximization on twitter: a mechanism for effective marketing campaign**. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE, 2017. 15

[46] Yan Mei, Weiliang Zhao, and Jian Yang. **Maximizing the effectiveness of advertising campaigns on twitter**. In *2017 IEEE International Congress on Big Data*. IEEE, 2017. 15

[47] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. **Measuring User Influence in Twitter: The Million Follower Fallacy.** *ICWSM*, **10**:10–17, 2010. 17, 28, 51, 52, 57, 61, 69, 75, 80, 81, 105

[48] Alex Leavitt, Evan Burchard, David Fisher, and Sam Gilbert. **The influentials: New approaches for analyzing influence on Twitter**. *Web Ecology Project*, **4**(2):1–18, 2009. 17, 28, 51, 52, 55, 57, 69, 75, 80, 81

[49] Evan TR Rosenman. *Retweets - but not just retweets: Quantifying and predicting influence on Twitter*. PhD thesis, Bachelors thesis, applied mathematics. Harvard College, Cambridge, 2012. 17, 57, 69, 75, 105

[50] Malcolm Gladwell. *The tipping point: How little things can make a big difference*. Hachette Digital, Inc., 2006. 18, 70

[51] Girish Chandrashekar and Ferat Sahin. **A survey on feature selection methods**. *Computers & Electrical Engineering*, **40**(1):16–28, 2014. 23

[52] George H John, Ron Kohavi, Karl Pfleger, et al. **Irrelevant features and the subset selection problem**. In *Machine Learning: Proceedings of the Eleventh International Conference*, pages 121–129, 1994. 25

[53] Ron Kohavi and George H John. **Wrappers for feature subset selection**. *Artificial intelligence*, **97**(1):273–324, 1997. 25

[54] Isabelle Guyon and André Elisseeff. **An introduction to variable and feature selection**. *Journal of machine learning research*, **3**(Mar):1157–1182, 2003. 25

[55] Avrim L Blum and Pat Langley. **Selection of relevant features and examples in machine learning**. *Artificial intelligence*, **97**(1):245–271, 1997. 25

[56] Pat Langley et al. **Selection of relevant features in machine learning**. In *Proceedings of the AAAI Fall symposium on relevance*, **184**, pages 245–271, 1994. 25

[57] Mark A Hall and Lloyd A Smith. **Feature subset selection: a correlation based filter approach**. *Springer*, 1997. 25

[58] Mark A Hall and Lloyd A Smith. **Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper.** In *FLAIRS conference*, **1999**, pages 235–239, 1999. 25

[59] Kenji Kira and Larry A Rendell. **The feature selection problem: Traditional methods and a new algorithm**. In *AAAI*, **2**, pages 129–134, 1992. 25

[60] Hussein Almuallim and Thomas G Dietterich. **Learning with Many Irrelevant Features.** In *AAAI*, **91**, pages 547–552, 1991. 25

[61] Pavel Pudil, Jana Novovičová, and Josef Kittler. **Floating search methods in feature selection**. *Pattern recognition letters*, **15**(11):1119–1125, 1994. 26

[62] Juha Reunanen. **Overfitting in making comparisons between variable selection methods**. *Journal of Machine Learning Research*, **3**(Mar):1371–1382, 2003. 26

[63] Larry J Eshelman. **The CHC adaptive search algorithm: How to have safe search when engaging**. *Foundations of Genetic Algorithms 1991 (FOGA 1)*, **1**:265, 2014. 26

[64] Luiz S Oliveira, Robert Sabourin, Flavio Bortolozzi, and Ching Y Suen. **A methodology for feature selection using multiobjective genetic algorithms for handwritten digit string recognition**. *International Journal of Pattern Recognition and Artificial Intelligence*, **17**(06):903–929, 2003. 26

[65] Oscar Cordón, Sergio Damas, and Jose Santamaría. **Feature-based image registration by means of the CHC evolutionary algorithm**. *Image and Vision Computing*, **24**(5):525–533, 2006. 26

[66] David E Goldberg and John H Holland. **Genetic algorithms and machine learning**. *Machine learning*, **3**(2):95–99, 1988. 26

[67] Michael L Raymer, William F. Punch, Erik D Goodman, Leslie A Kuhn, and Anil K Jain. **Dimensionality reduction using genetic algorithms**. *IEEE transactions on evolutionary computation*, **4**(2):164–171, 2000. 26

[68] RC Eberchart and J Kennedy. **Particle swarm optimization**. In *IEEE International Conference on Neural Networks, Perth, Australia*, 1995. 26

[69] Li-Yeh Chuang, Hsueh-Wei Chang, Chung-Jui Tu, and Cheng-Hong Yang. **Improved binary PSO for feature selection using gene expression data**. *Computational Biology and Chemistry*, **32**(1):29–38, 2008. 26

[70] Fabián Riquelme and Pablo González-Cantergiani. **Measuring user influence on Twitter: A survey**. *Information Processing & Management*, **52**(5):949–975, 2016. 27

[71] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. **The PageRank citation ranking: Bringing order to the web.** 1999. 27, 51

[72] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. **What is Twitter, a social network or a news media?** In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010. 27, 28, 46

[73] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. **Twitterrank: finding topic-sensitive influential twitterers**. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM, 2010. 27, 51, 81

[74] D Tunkelang. **A Twitter analog to PageRank**. *The Noisy Channel*, 2009. 27, 51, 81

[75] Aron Yu, C Vic Hu, and Ann Kilzer. **Khyrank: Using retweets and mentions to predict influential users**, 2011. 28, 51, 57, 69, 75, 81

[76] Wenlong Chen, Shaoyin Cheng, Xing He, and Fan Jiang. **Influencerank: An efficient social influence measurement for millions of users in microblog**. In *2012 Second International Conference on Cloud and Green Computing (CGC)*, pages 563–570. IEEE, 2012. 28, 51, 81

[77] Phil E Brown and Junlan Feng. **Measuring user influence on twitter using modified k-shell decomposition**. In *Fifth international AAAI conference on weblogs and social media*, 2011. 28

[78] Rafael Cappelletti and Nishanth Sastry. **IARank: Ranking Users on Twitter in Near Real-time, Based on their Information Amplification Potential**. In *2012 International Conference on Social Informatics (SocialInformatics)*, pages 70–77. IEEE, 2012. 28, 51, 81

[79] Daniel M Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A Huberman. **Influence and passivity in social media**. In *Proceedings of the 20th international conference companion on World wide web*, pages 113–114. ACM, 2011. 29

[80] Menno Luiten, Walter A Kosters, and Frank W Takes. **Topical Influence on Twitter: A Feature Construction Approach**. In *Proceedings of 24th Benelux Conference on Artificial Intelligence (BNAIC 2012)*, pages 139–146, 2012. 29

[81] Yaxi He, Chunhong Zhang, and Yang Ji. **Principle Features for Tie Strength Estimation in Micro-blog Social Network**. In *2012 IEEE 12th International Conference on Computer and Information Technology (CIT)*, pages 359–367. IEEE, 2012. 29

[82] Shaomei Wu, Jake M Hofman, Winter A Mason, and Duncan J Watts. **Who says what to whom on Twitter**. In *Proceedings of the 20th international conference on World wide web*, pages 705–714. ACM, 2011. 29, 58

[83] Shishir Bharathi, David Kempe, and Mahyar Salek. **Competitive influence maximization in social networks**. *Internet and Network Economics*, pages 306–311, 2007. 31

[84] Flaviano Morone and Hernán A Makse. **Influence maximization in complex networks through optimal percolation**. *Nature*, **524**(7563):65–68, 2015. 31

[85] Shuo Chen, Ju Fan, Guoliang Li, Jianhua Feng, Kian-lee Tan, and Jinhui Tang. **Online topic-aware influence maximization**. *Proceedings of the VLDB Endowment*, **8**(6):666–677, 2015. 31

[86] Maoguo Gong, Chao Song, Chao Duan, Lijia Ma, and Bo Shen. **An Efficient Memetic Algorithm for Influence Maximization in Social Networks**. *IEEE Computational Intelligence Magazine*, **11**(3):22–33, 2016. 31

[87] Nan Du, Yingyu Liang, Maria-Florina Balcan, Manuel Gomez-Rodriguez, Hongyuan Zha, and Le Song. **Scalable Influence Maximization for Multiple Products in Continuous-Time Diffusion Networks**. *Journal of Machine Learning Research*, **18**(2):1–45, 2017. 31

[88] Mark Granovetter. **Threshold models of collective behavior**. *American journal of sociology*, pages 1420–1443, 1978. 32

[89] Ning Chen. **On the approximability of influence in social networks**. *SIAM Journal on Discrete Mathematics*, **23**(3):1400–1415, 2009. 32

[90] David Kempe, Jon Kleinberg, and Éva Tardos. **Influential nodes in a diffusion model for social networks**. In *Automata, languages and programming*, pages 1127–1138. Springer, 2005. 32

[91] Zaobo He, Zhipeng Cai, and Xiaoming Wang. **Modeling propagation dynamics and developing optimized countermeasures for rumor spreading in online social networks**. In *2015 IEEE 35th International Conference on Distributed Computing Systems (ICDCS)*, pages 205–214. IEEE, 2015. 32

[92] Jacob Goldenberg, Barak Libai, and Eitan Muller. **Talk of the network: A complex systems look at the underlying process of word-of-mouth**. *Marketing letters*, **12**(3):211–223, 2001. 34

[93] Ali Zarezade, Ali Khodadadi, Mehrdad Farajtabar, Hamid R Rabiee, and Hongyuan Zha. **Correlated Cascades: Compete or Cooperate.** In *AAAI*, pages 238–244, 2017. 39

[94] Amit Goyal, Wei Lu, and Laks VS Lakshmanan. **Celf++: optimizing the greedy algorithm for influence maximization in social networks**. In *Proceedings of the 20th international conference companion on World wide web*, pages 47–48. ACM, 2011. 41, 142

[95] Masahiro Kimura and Kazumi Saito. **Approximate solutions for the influence maximization problem in a social network**. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 937–944. Springer, 2006. 41

[96] Masahiro Kimura and Kazumi Saito. **Tractable models for information diffusion in social networks**. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 259–271. Springer, 2006. 41

[97] Amit Goyal, Wei Lu, and Laks VS Lakshmanan. **Simpath: An efficient algorithm for influence maximization under the linear threshold model**. In *2011 IEEE 11th International Conference on Data Mining (ICDM)*, pages 211–220. IEEE, 2011. 43

[98] Edward Keller and Jonathan Berry. *The influentials: One American in ten tells the other nine how to vote, where to eat, and what to buy.* Simon and Schuster, 2003. 58, 71, 83

[99] K. Pearson. **Note on regression and inheritance in the case of two parents**. *Proceedings of the Royal Society of London*, **58**(347-352):240–242, 1895. 61

[100] Alfréd Rényi et al. **On measures of entropy and information**. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961. 62

[101] J Machta. **Entropy, information, and computation**. *American Journal of Physics*, **67**(12):1074–1077, 1999. 62

[102] Wanhua Qiu. **Management decision and applied entropy**, 2002. 62

[103] Adithya Rao, Nemanja Spasojevic, Zhisheng Li, and Trevor DSouza. **Klout score: Measuring influence across multiple social networks**. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2282–2289. IEEE, 2015. 65

[104] Charles Spearman. **The proof and measurement of association between two things**. *The American journal of psychology*, **15**(1):72–101, 1904. 67

[105] Shaozhi Ye and Shyhtsun Felix Wu. **Measuring message propagation and social influence on twitter. com.** *SocInfo*, **10**:216–231, 2010. 69

[106] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2005. 71

[107] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*, **821**. John Wiley & Sons, 2012. 71

[108] Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. **Is it really about me?: message content in social awareness streams**. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 189–192. ACM, 2010. 88

[109] Marko Robnik-Šikonja and Igor Kononenko. **Theoretical and empirical analysis of ReliefF and RReliefF**. *Machine learning*, **53**(1-2):23–69, 2003. 92, 93

[110] Marko Robnik-Šikonja and Igor Kononenko. **An adaptation of Relief for attribute estimation in regression**. In *Machine Learning: Proceedings of the Fourteenth International Conference (ICML97)*, pages 296–304, 1997. 92

[111] Anil Jain and Douglas Zongker. **Feature selection: Evaluation, application, and small sample performance**. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **19**(2):153–158, 1997. 93

[112] Stephen Grossberg. **Nonlinear neural networks: Principles, mechanisms, and architectures**. *Neural networks*, **1**(1):17–61, 1988. 94

[113] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994. 94

[114] ATC Goh. **Back-propagation neural networks for modeling complex systems**. *Artificial Intelligence in Engineering*, **9**(3):143–151, 1995. 94

[115] Steve Jurvetson. **What exactly is viral marketing**. *Red Herring*, **78**:110–112, 2000. 101

[116] Sabrina Helm. **Viral marketing-establishing customer relationships by'word-of-mouse'**. *Electronic markets*, **10**(3):158–161, 2000. 101

[117] Matthew Richardson and Pedro Domingos. **Mining knowledge-sharing sites for viral marketing**. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 61–70. ACM, 2002. 101

[118] Rohan Miller and Natalie Lammas. **Social media and its implications for viral marketing**. *Asia Pacific Public Relations Journal*, **11**(1):1–9, 2010. 101

[119] Andreas M Kaplan and Michael Haenlein. **Two hearts in three-quarter time: How to waltz the social media/viral marketing dance**. *Business Horizons*, **54**(3):253–263, 2011. 101

[120] Michael J Naples. *Effective frequency: the relationship between frequency and advertising effectiveness*. Association of National Advertisers, 1979. 135

[121] Alan G Sawyer. **Repetition, cognitive responses, and persuasion**. *Cognitive responses in persuasion*, pages 237–261, 1981. 135

[122] Daniel E Berlyne. **Novelty, complexity, and hedonic value**. *Attention, Perception, & Psychophysics*, **8**(5):279–286, 1970. 135

[123] Thomas Smith. *Successful Advertising. Its Secrets Explained...* Smith's Press, 1885. 135

[124] Herbert E Krugman. **Why three exposures may be enough**. *Journal of advertising research*, **12**(6):11–14, 1972. 136

[125] Gert Sabidussi. **The centrality index of a graph**. *Psychometrika*, **31**(4):581–603, 1966. 141