Consensus opinion model in online social networks based on the impact of influential users

Amir Mohammadinejad

Doctor of Philosophy (PhD) Thesis
Institut Mines Telecom, Télécom SudParis

Specialization: COMPUTER SCIENCE

Doctoral School:
Computer Science, Telecommunications and Electronics of Paris

Presented by

Amir MOHAMMADINEJAD

Consensus Opinion Model in Online Social Networks based on the impact of Influential Users

4^{th} December 2018

Commitee:

Gabriela Pasi	Reviewer	Professor, University of Milano Bicocca - Italy
Abdelhamid Mellouk	Reviewer	Professor, UPEC - France
Ioan Marius Bilasco	Examiner	Professor, Université Lille 1 - France
Mai-Trang Nguyen	Examiner	Professor, UPMC - France
Reza Farahbakhsh	Advisor	Researcher, Télécom SudParis - France
Noel Crespi	Director	Professor, Télécom SudParis - France

Thesis Number : 2018TELE0018

Thèse de Doctorat de
l'Institut Mines Telecom, Télécom SudParis

Spécialité : INFORMATIQUE

École Doctorale :
Informatique, Télécommunications et Electronique de Paris

Présentée par

Amir MOHAMMADINEJAD

Modèle D'avis De Consensus Dans Les Réseaux Sociaux En Ligne Basé Sur L'impact Des Utilisateurs Influents

4^{th} Décembre 2018

Jury composé de :

Gabriela Pasi	Rapporteur	Professor, University of Milano Bicocca - Italy
Abdelhamid Mellouk	Rapporteur	Professor, UPEC - France
Ioan Marius Bilasco	Examinateur	Professor, Université Lille 1 - France
Mai-Trang Nguyen	Examinateur	Professor, UPMC - France
Reza Farahbakhsh	Encadrant	Chercheur, Télécom SudParis - France
Noel Crespi	Directeur	Professor, Télécom SudParis - France

Thèse numéro : 2018TELE0018

Dedication

To my Dearest Friends, Family
and
To my wife.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my Supervisor Prof. Noel Crespi for the continuous support of my Ph.D. study and related research and Reza Farahbakhsh for his motivation. Their guidance helped me in all the time of research and writing of this thesis. I could not have imagined having better mentors for my Ph.D. study. Besides them, I would like to thank the rest of my thesis committee : Prof. Gabriella Pasi, Prof. Mai-Trang Nguyen, and Prof. Ioan Marius Bilasco, for their insightful comments and encouragement which motivated me to widen my research from various perspectives. More importantly, I would like to thank my family, especially my wife Dr. Hajar Hiyadi for her immense knowledge and supporting throughout my Ph.D., writing this thesis and my life in general.

Amir Mohammadinejad
20^{th} November 2018

Abstract

Online Social Networks are increasing and piercing our lives such that almost every person in the world has a membership at least in one of them. Among famous social networks, there are online shopping websites such as Amazon, eBay and other ones which have members and the concepts of social networks apply to them. This thesis is particularly interested in the online shopping websites and their networks. According to the statistics, the attention of people to use these websites are growing due to their reliability. The consumers refer to these websites for their need (which could be a product, a place to stay, or home appliances) and become their customers. One of the challenging issues is providing useful information to help the customers in their shopping. Thus, an underlying question the thesis seeks to answer is how to provide a comprehensive information to the customers in order to help them in their shopping. This is important for the online shopping websites as it satisfies the customers by this useful information and as a result increases their customers and the benefits of both sides.

To overcome the problem, three specific connected studies are considered: (1) Finding the influential users, (2) Opinion Propagation and (3) Opinion Aggregation. In the first part, the thesis proposes a methodology to find the influential users in the network who are essential for an accurate opinion propagation. To do so, the users are ranked based on two scores namely optimist and pessimist. In the second part, a novel opinion propagation methodology is presented to reach an agreement and maintain the consistency among users which subsequently, makes the aggregation feasible. The propagation is conducted considering the impacts of the influential users and the neighbors. Ultimately, in the third part, the opinion aggregation is proposed to gather the existing opinions and present it as the valuable information to the customers regarding each product of the online shopping website. To this end, the weighted averaging operator and fuzzy techniques are used.

The thesis presents a consensus opinion model in signed and unsigned networks. This solution can be applied to any group who needs to find a plenary opinion among the opinions of its members. Consequently, the proposed model in the thesis provides an accurate and appropriate rate for each product of the online shopping websites that gives a precious information to their customers and helps them to have a better insight regarding the products.

Keywords

Online Social Networks, Influential Users, Opinion Propagation, Opinion Aggregation, Consensus Opinion Model, Optimist and Pessimist users, Fuzzy Majority Opinion, Link Analysis, Shopping Websites

Résumé

Cette thèse s'intéresse particulièrement aux sites de vente en ligne et à leurs réseaux sociaux. La propention des utilisteurs utiliser ces sites Web tels qu'eBay et Amazon est de plus en plus importante en raison de leur fiabilité. Les consommateurs se réfèrent à ces sites Web pour leurs besoins et en deviennent clients. L'un des défis à relever est de fournir les informations utiles pour aider les clients dans leurs achats. Ainsi, une question sous-jacente à la thèse cherche à répondre est de savoir comment fournir une information complète aux clients afin de les aider dans leurs achats. C'est important pour les sites d'achats en ligne car cela satisfait les clients par ces informations utiles.

Pour surmonter ce problème, trois études spécifiques ont été réalisées : (1) Trouver les utilisateurs influents, (2) Comprendre la propagation d'opinion et (3) Agréger les opinions. Dans la première partie, la thèse propose une méthodologie pour trouver les utilisateurs influents du réseau qui sont essentiels pour une propagation précise de l'opinion. Pour ce faire, les utilisateurs sont classés en fonction de deux scores : optimiste et pessimiste. Dans la deuxième partie, une nouvelle méthodologie de propagation de l'opinion est présentée pour parvenir à un accord et maintenir la cohérence entre les utilisateurs, ce qui rend l'agrégation possible. La propagation se fait en tenant compte des impacts des utilisateurs influents et des voisins. Enfin, dans la troisième partie, l'agrégation des avis est proposée pour rassembler les avis existants et les présenter comme des informations utiles pour les clients concernant chaque produit du site de vente en ligne. Pour ce faire, l'opérateur de calcul de la moyenne pondérée et les techniques floues sont utilisées.

La thèse présente un modèle d'opinion consensuelle dans les réseaux. Les travaux peuvent s'appliquer à tout groupe qui a besoin de trouver un avis parmi les avis de ses membres. Par conséquent, le modèle proposé dans la thèse fournit un taux précis et approprié pour chaque produit des sites d'achat en ligne.

Mots-clés

Réseaux Sociaux En Ligne, Utilisateurs Influents, Propagation D'avis, Agrégation D'avis, Modèle D'avis De Consensus, Utilisateurs Optimistes Et Pessimistes, Avis De La Majorité Fuzzy, Analyse De Liens, Sites D'achat En Ligne

Table of contents

Chapter 1

Introduction

Contents

1.1 Motivation and Research Problems

Online Social Networks (OSN) are increasing and piercing our lives such that almost every person in the world has a membership at least in one of them. Among famous social networks such as Facebook and Tweeter, there are other networks like Amazon, eBay and also small online shopping websites which have members and communities and the concepts of social networks applies to them. A user may refer to these websites to get information about TV, songs, movie tickets, jobs, or even mates. As these decisions and the fundamental financial processes move to the web, there is growing economic inspiration to propagate opinion through the web. Hence, these people take a lot of decisions toward such networks. In particular, the networks that we are talking about are online shopping centers that have their own users and the network of them. In recent years the demand of people for online shopping increased and it is predicted that this increment will continue. Because of the numerous advantages and benefits, more and more people prefer online shopping over conventional shopping these days. The reasons for this attraction are introduced by some factors such as shopping in their pajamas to convenience for the elderly and disabled. Also because of the wider choice, not subject to up-selling or impulse buying, better prices, good

for the environment, and more. In general, the reasons for online shopping are:

1) Convenience: The customers don't have to wait in a line or wait till the shop assistant helps her with her purchases. She can do her shopping in minutes even if she is busy, apart from saving time and avoiding crowds. Online shops give us the opportunity to shop $24x7$ and also reward us with 'no pollution' shopping.

2) Better Prices: People get cheap deals and better prices from online stores because products come to them directly from the manufacturer or seller without middlemen involved. Also, many online shops offer discount coupons and rebates.

3) Variety: One can get several brands and products from different sellers at one place. The consumers can get in on the latest international trends without spending money on travel; they can shop from retailers in other parts of the country or even the world without being limited by geographic area. These stores offer a far greater selection of colors and sizes than they will find locally.

4) Fewer Expenses: Many times when people opt for conventional shopping they tend to spend a lot more than the required shopping expenses, on things like eating out, traveling, impulsive shopping etc.

5) Comparison of Prices: Online shops make comparison and research of products and prices possible. Online stores also give the consumers the ability to share information and reviews with other shoppers who have firsthand experience with a product or retailer.

6) Crowds: People like to avoid the crowds when they do the shopping. Crowds force them to do a hurried shopping most of the time. Crowds also create a problem when it comes to finding a parking place nearby where they want to shop and going back to their vehicle later loaded with shopping bags.

7) Compulsive Shopping: Many times when people go out shopping they end up buying things which they do not require because of the shopkeepers' up-selling skills or they will compromise on their choices because of the lack of choices in those shops.

8) Discreet Purchases: Some things are better done in privacy. Online Shops enable the consumers to purchase undergarments and lingerie or adult toys without the embarrassment that there are several people watching them and their choices.

Furthermore, the statistics show the demands of people in shopping online are increasing. Figure 1.1 gives information on retail e-commerce sales worldwide from 2014 to 2021. In 2017, retail e-commerce sales worldwide amounted to 2.3 trillion US dollars and e-retail revenues are projected to grow to 4.88 trillion US dollars in 2021.

In order to recommend a product or prepare some information to a user who wants to buy a product, these websites gather the opinions of other users who already purchased the current product and at the end provide a rate or some information as an aggregated opinion of other users for that. This process which is named group decision making, recom-

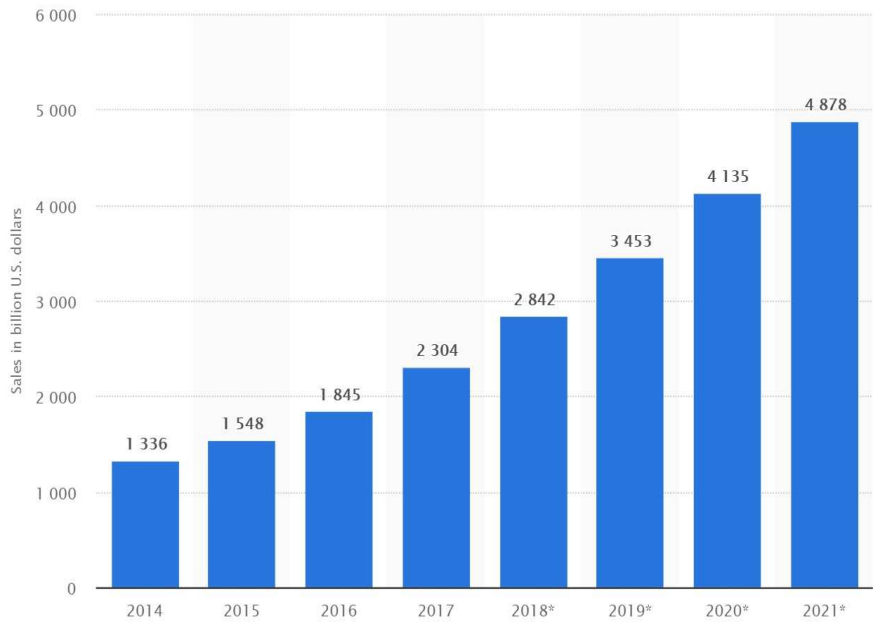


Figure 1.1 – Statistics of online demands in each year (* are predicted)

mendation system, opinion mining and etc. in other articles got a lot of attention in recent years. However, it is still difficult to determine the extent to which such users affect the opinions of the others. We claim that each user has a different impact on others which affect their general opinions and can be reflected in other links she may initiate. It is in this way that opinion propagates through a network. Of particular interest, and the focus of this study is to investigate an opinion consensus model in a networked social group based on link analysis. For example, in online shopping centers such as Amazon, eBay, Alibaba and etc. the most important issue is satisfying the customers (to have a better and successful company). These websites have products and also the networks of their users in which each user can review the products. While a user reviews the products, one of the most situations regarding these websites is when she wants to decide about buying a product and she has an inquiry about it. The websites usually put the aggregated rate and other users reviews. This aggregated rate gives a valuable information to the current user and can effectively form or change her opinion. In other words, the shopping websites provide a comprehensive opinion for each of their products and help their customers to decide about them (in their shopping), but how this rate can be aggregated in an accurate way so that the customers can be assured about the provided rate? The main problems in aggregating are reaching the agreement among the users to omit the inconsistency (see section 5.1) and dealing with incomplete information (it means that in some situations such as when a new user comes to

the network, we don't know what is her opinion and cannot participate her in aggregation). Both situations can be solved by propagating the opinion towards the network. Thus, these websites can spread the effective opinions through their users. In addition, in the propagation process, the users will affect and consequently change each others' opinions. However, as long as the impact of users is not equal to each other, to perform a good and precise propagation, it is needed to know which users have more impact. Hence, by finding these special users, the propagation can be performed perfectly and as a result, the accurate aggregation will be feasible. This rate will help the customers to choose the best decision in their shopping. Furthermore, if the online shopping websites know which users have the better effect on the current user, they can recommend and highlight these specific users' opinions who they know their opinions have a positive and constructive impact. Together with providing more accurate information through these websites, the more satisfaction will be reached among the customers and therefore, more benefits will be attained. This benefit includes both sides (customers and online websites) as it helps the customers to have better insight about the products and choose the right one and also the shopping websites to have more customer and more sells. It is worth mentioning that with propagating the opinions, these websites can predict and realize the current and future opinions of their customers regarding each product and adapt their products based on that.

Generally, in a network, the users' opinions are based on the information they have and the ones shared by others. For instance, which city to travel to?, what kind of medicine may help? which candidate to vote for? and etc. which is characterized by users' and others opinion. In recent years, opinion formation in online social networks has become a widespread phenomenon that indicates its importance. It can be even said that almost all social interactions are shaped by users beliefs and opinions [1]. Thus it is of high value to study opinion dynamics, and up to now, many researchers with different background have proposed various models to analyze the evolution of the opinion dynamics, propagation, and aggregation from various aspects [2–4].

Given a product, a user or a group (social network is full of groups) such as a company wants to make a decision about purchasing it. Considering the network of users, each of them implies their decision - which we call opinion - by prior knowledge, experts or influential users' knowledge and her friends' knowledge toward the products. The user may review the other users' opinions and negotiate with them and therefore, get affected by their opinions. In this way, the product opinions will propagate through the users of the network. Finally, after this propagation, these opinions should be aggregated and presented as the comprehensive opinion of the users toward the product (1.2) to provide a helpful information for the current user who wants to decide about purchasing it. This user can be a manager of a company who needs to take the decision about the product or simply she

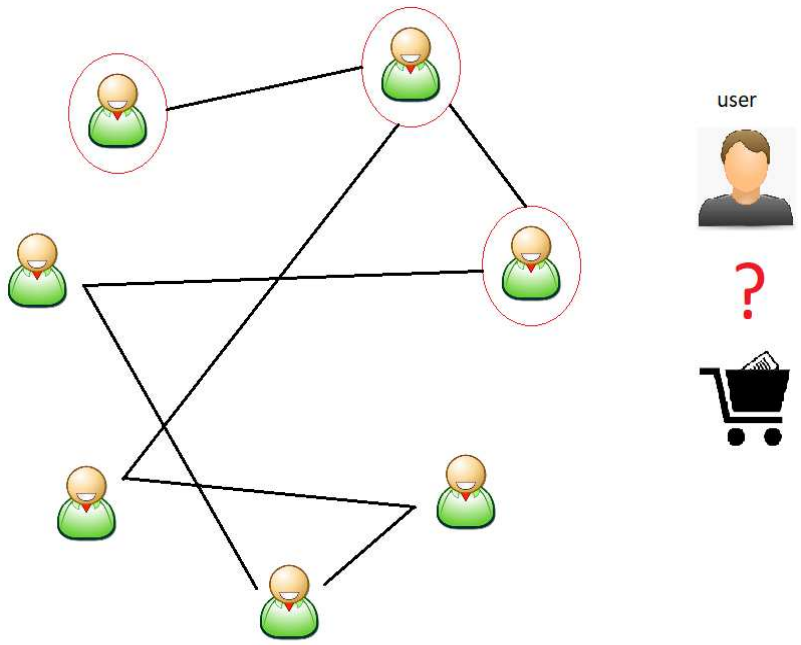


Figure 1.2 – Problem definition and propagating the opinion

is a user of an online shopping center who wants to buy the product. The user's opinion can be changed by the other users' opinions through the connections and the links she has. The aggregated opinion will be presented to the current user as the final rate or opinion of others who already rated this product (and in case of a company, it will be presented as the final decision of the members toward the product). This comprehensive rate will help the current users in her decision. In figure 1.2, the red circles show the users who have more information and can affect others more than usual users (see 3). These users who we call them influential users are not considered in current related studies. Considering them and the links of the users, we propagate the opinions in the network and then aggregate these opinions to provide the ultimate rate for each product. In another scenario, consider an officer who wants to take a decision regarding an attack. There is some information from the different sources regarding the enemy but which one is more important and how can he aggregate this information and adopt his decision? He needs to know which source is more important and how to connect this information and finally decide based on all of the provided information.

Social shopping is expanding significantly in internet business because of the advancement of social media applications. Consequently, website managers face numerous challenges in delivering their quality website experience in order to fulfill clients' needs and in

creating connections among members, and network. In brief, it is very important to offer a great quality website experience to support online clients. Hence, it is important to give hypothetical conceptualizations just as fundamental empirical proof for such phenomena in which social shopping is supported. The empirical results demonstrate that the apparent framework and administration quality are vital precursors of client fulfillment, but not for the impact of perceived information quality on him. Besides, it demonstrates that client fulfillment essentially impacts engagement, trust, and buy expectation, and trust in its turn fundamentally impacts engagement [5].

While investigation the affiliations that exist among clients and businesses, we find that the quality of relationships assumes a noteworthy role. Searching information about an item before purchasing it then sharing personal online after its use is a need that is satisfied by combining shopping with social networking. Here are some examples of social sites, blogs, and communities for shopping, exchanging and sharing opinions on products and recommending their favorites. These recommendations and opinions help clients to get interesting information about products and guide them to make decisions about their shopping. The relationship between the business and the client can be built or devastated depending to relationship quality before, during and after transactions. Therefore, to ensure that researchers and practitioners may better comprehend and hold connections even more proficiently, it is imperative to build up a proportion of relationship quality in social shopping context. In addition, e-Marketer reported that individuals are tend to trust information and recommendation provided by other clients more than that given by companies. Therefore, shopping sites may improve their business volume by means of purchasers' trust through the opinions or suggestions shared by other known or obscure customers. Albeit sees trust of online shopping is basic to help online buyers, it is essential to provide empirical proof for such phenomena in which social shopping is enabled.

Social shopping sites such as Kaboodle, ThisNext, Crowdstorm, Stylehive, Rakuten Ichiba Taiwan, LuxJoy, WooGii, and Pinkoi provide blogs or virtual communities for consumers to share shopping thoughts, exchange opinions on specific merchandises, and recommend their favorites. Therefore, influencing potential consumers through information sharing and interactions of the provided blogs or virtual communities is a significant benefit of social shopping sites. In addition, these websites employ a lot of data scientists to verify the data they have and provide it as a helpful information to their customers [1]. Website managers are developing social shopping functions or launching social shopping networks on their websites due to the increase in social shopping and online shops/auction. In a social shopping context, the firms often cooperate with social networking sites such as Facebook and Twitter to encourage their transactions and in contrast, consumers are able

[1]https://www.simplilearn.com/why-and-how-data-science-matters-to-business-article

to help sellers' products and services spontaneously through social networking sites [6].

Today, there are several websites providing such information to online shopping websites. For example, Crazy egg is a website, which provides optimization tools for social shopping shops such as Etsy, Zendesk, Dell etc. to improve the user experience (they believe happier customers make higher revenue!). In general, they record the entire users' sessions and following that make the network of users and record their activities such as buying the products and commenting them. Later, this information will be provided to the social shopping websites to help them to maximize the users' registrations, increasing the subscription and selling more products. On the other hand, the social shopping websites use this information to provide more insights for their customers. For instance, Wanelo, Etsy, Fancy etc. use the users' rates to compute the products' general reviews and provide them to other users. Charles schwab is another websites, which uses tweeter platform to help customers. They help their customers to have insights where to invest, how to manage money, how to use it etc. by providing useful information gathered from the experience of other users (consensus experience).

On the plus side of things, social networking is helping online enterprises become more focused on their customers than ever before. Businesses are finding out what people want and delivering those things much more quickly. Company executives believe they have more control over their businesses' reputations on the Internet because they can influence people's perceptions in cyberspace. This also applies to how an online company is viewed by suppliers and as a competitor to other companies. Social networking technology can help online businesses learn more about what their competition is doing and respond rapidly. In addition, there are three main factors in content marketing: 1) attracting the customers by putting influential users comments and the other users' reviews, 2) providing information such as consensus opinion method to help the users to decide, and 3) keeping the customers by recommending them what to buy, where to search, how to work with current product, customer service etc. which improve their satisfaction. In short, information quality, communication between users and sellers, and word of mouth (WOM) communication play crucial roles in the development of trustworthy social shopping sites [5].

The research questions regarding the above problem are:

1) How can we distinguish the influential users in a network? 2) How the opinion of influential users propagate in a network and affect others? 3) How can we present the final decision of a group as a consensus opinion toward the product?

1.2 Proposed Solutions

The solution of the above problem is proposing a methodology which helps to aggregate the opinions in a correct way and as a result provides the comprehensive opinion for each product. In order to be able to aggregate the opinions, we need to reach an agreement among the users. Some of the users have a different background which prevents us to simply average the opinions. To reach this agreement, the propagation method is proposed. With the propagation, the inconsistency problem will be dropped away and the aggregation will be feasible. Current studies on opinion dynamic are based on the links a user has in the network which means the users' opinion is affected by the opinion of her connections (neighbors). However, the impact of neighbors are different, i.e. some of them have a greater impact. We consider such neighbors as influential users who are more popular or trusted among others. To the best of our knowledge, there is no prior study which considers the impact of influential users in opinion formation. Current studies assume that the experts of the network are specified before and they just need to aggregate their opinions. However, in most situations, they are not defined.

The aim of this study is to present an accurate rate for each product which gives a valuable information to the customers. To do so, considering a product, we need to aggregate the rates of other users who rated this product. In addition, to aggregate the opinions, we need to propagate the opinions to omit the inconsistency and solve the incomplete information. Finally, the proper propagation can be achieved if we know which user can affect the opinion of others (influential users). In other words, we divided the solution to three main parts. Consider an online shopping website in which its users have some opinions toward some products. First, we find the influential users and then propagate their opinion in the network and later aggregate their opinion as final rate for the product. Hence, our proposed solution has three main phases:

1) Finding the Influential users in the online social networks. 2) Opinion Propagation in Online Social Networks considering the impact of influential users 3) Aggregate the opinions and present it as the consensus opinion for the product (opinion consensus model)

In this study, we present a consensus opinion model based on the impact of influential users in signed and unsigned networks in the context of link analysis. The links of a social network show the connection between users. In signed networks, the link between users has positive or negative values. These signs present trust/distrust relation between users. Consider a shopping center website that has its products and the network of its users. These users are partially connected with each other (the network of users) and rated some of the products as their opinions (each user has some opinions for a limited number of products). Thus, there are two kinds of links, the first one is the link between users and the second

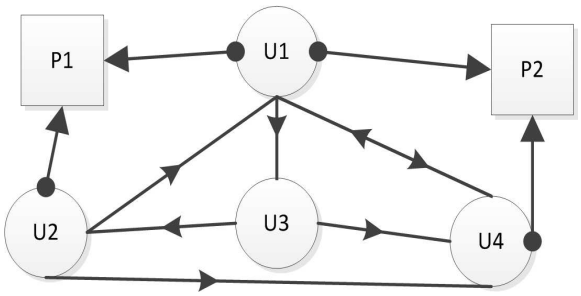


Figure 1.3 – simple network of users and products

one is the rates or opinions of users toward the products. Two users are neighbors if there is a link between them. In signed networks, the links have positive and negative values and in directed networks the links have direction. Here, we say $U1$ is neighbor of $U2$ if there is a link from $U2$ to $U1$. Figure 1.3 shows our network of users and products. In this figure, there are two products $P1$ and $P2$ and four users. There is an opinion from $U4$ toward $P2$ but $U1$ has opinion for both $P1$ and $P2$. Also, $U1$ and $U4$ are the neighbors of $U2$.

We address the problem of consensus opinion formation by considering today online shopping centers. We considered four websites (Epinion, Amazon, Booking, and eBay) and assumed their network and users opinion adoption as our base for the model. In reality, the neighbors of a user have a direct effect on her opinion. Thus, the neighbors can convince her to take the same opinion as they have regarding a product and in this way, they propagate their opinions in the network. As described before, the user opinion is influenced by the opinion of neighbors, influential users and the current users' knowledge. The impact of neighbors and influential users can be considered through the link analysis and the users' knowledge can be determined by analyzing the profile of her. Here, we utilize the link analysis and considering the neighbors, we propose a model for consensus opinion formation based on the impact of influential users through the signed and unsigned networks.

1.3 Contributions of the Thesis

The main contributions of this study can be summarized as follows: (1) Introducing a ranking method to distinguish the influential users and their importance in signed and unsigned networks as well as their opinion propagation impact. (2) Introducing influential user as a new feature for opinion propagation and propose an opinion propagation model based on the impact of this feature. (3) Investigating different aggregation methods and introducing the fuzzy aggregating model to provide the consensus opinion or rating score of products based on users' opinion.

1.4 Thesis Organization

A brief synopsis of each of the chapters of this thesis is included below: Chapter 2 presents the background and state-of-the-art of current studies. In this chapter, we describe studies regarding our solution (for each of the main phases). In chapter 3, we propose our method to find the influential users in the network which is based on the ranking of the users. We also investigate different ranking algorithm as our methodology and describe the personality feature used for influential users identification. Furthermore, we introduce the credibility as a measure to analyze our ranking. Chapter 4 discusses the opinion propagation in OSNs based on the impact of influential users. This chapter includes the detailed discussion on opinion propagation. Moreover, the fuzzy majority opinion is presented for the evaluating the proposed propagation model. Chapter 5 provides the consensus opinion model introducing the fuzzy aggregating method. Chapter 6 presents the conclusion and future work toward this study. Finally, chapter 7 presents the other research parallel to this study.

Chapter 2

Background and State-of-the-Art

Contents

2.1 Introduction

In the following chapter, we will describe the state-of-the-art for the main problem and the described phases. The online shopping portals (websites) are a great place to buy products for everybody. The products in online shopping websites can be a dress, a car, an electric device, home appliances, or searching for hotels or place to stay for a short/long period and etc. and as described in the previous chapter, due to their reliabilities, who does not like to benefit from them. The organization of this chapter is as follows: First, we discuss the basics of online shopping websites and the way they provide information to their customers. Then the state-of-the-art for each phase will be presented and finally, this chapter will be summarized in the last section.

2.2 Basics of Online shopping websites

There are two main data in the online shopping websites. The first one is the products and the second one is the users of the website. These portals usually put their products in which the users can see them and their descriptions. One of the most important description regarding each product is the overall rating of that based on the reviews of other users. The websites simply ask the other users who used the current product to rate it and then aggregate these rates to present it as a valuable information to other users. This rate has a great influence on the decision of the current user who needs to decide about buying this product. In this way, the websites give a great information to their customers and as a consequence, satisfy them. There are several studies tried to find a consensus opinion (see section 2.5). Our solution has three phases: 1) finding the influential users, 2) propagating the opinions in the network and 3) aggregating the opinions and presenting the consensus opinion. Below are the current studies for each of these phases.

2.3 Influential Users

In general, the researches for the problem of expert finding can be categorized into two main groups, (i) Authority ranking approaches [7–14] and (ii) non-authority ranking approaches [15–22].

The authority ranking approaches are based on link analysis for finding the influential users. These techniques which are based on web page rankings, evaluate the connections and relationships between users of a network and it is used in a situation in which there is no access to the profile of users. For example, Kardan et al. in [23] find the experts to solve the problem regarding whose knowledge in social networks should be shared, which is based on the PageRank method. Later in [24] they extend the experts detection in online communities. S. Chen et al. proposed an integrated PageRank method in [7] for the maximization problem to select the seeds in signed networks. Jurczyk et al. in [13] discover the users' authorities in question answer communities by adjusting the HITS method [25]. X. Kong et al. in [8] tried to calculate the authors' impacts on the author-paper network by a new algorithm based on PageRank scores. H. Zhu et al. in [9] proposed an expert finding framework using Topical Random Surfer (TRS) which is originally used for web page ranking. Bouguessa et al. [12] identify the experts in question-answering forums by ranking the users regarding the validity of their answers. On the other hand, the non-authority ranking approaches are based on information retrieval from the activity and profile of the users. This class of ranking methods aims to find the experts using the information included in the profile of the users as well as analyzing their activities and posts in a given social network. For example, H. Deng et al. in [18] tried to develop weighted language,

topical and hybrid model based on them for expert finding in DBLP bibliography and Google scholar dataset. Chen et al. [19] proposed a model for expert detection using the user activity analysis in rating the comments in question answering systems. D. mimno et al. [21] created a user model in order to determine the expertise level of reviewers based on papers. Also, J. Went et al. in [22] tried to find the influential users in Tweeter using the topical similarity among users. In addition, there are methods which utilizing both link and profile information to increase the accuracy of the detection task. For instance, J. Zhang et al. in [17] proposed a propagation based approach that takes into consideration both the person local information and the network information (e.g. relationships between persons). Zhiqiang et al. in [26] proposed a method to find the experts based on the gating method that dynamically combines structure aspect and text aspect representations in question answering networks. Their method combines the two users' features namely structure and text and based on that find the experts in a given network. Z. Zhao et al in [15] declared that some parts of the available information in question answering systems are missing and then they find the experts using matrix completion technique and users similarity to fill the gap. Balog et al. [27] introduced a probability model for expert identification based on users topical profile in multilingual systems. Guo et al. [20] presented a method to find the best related user regarding a specific question by constructing the user's profile by discovering latent topics and interests of users. Lu et al. in [28] used the question sessions and user profiles to build the network graph. Then using this graph, they proposed two expert detection method based on semantic propagation and semantic language model. Shahriari et al. In [16] proposed a new method to identify the experts using overlapping community detection. The presented personality feature that is used in ranking algorithms falls in the first category (link analysis).

In [29], the authors proposed a system which includes a memory storing a dataset representing a community of users of a social networking service.It also includes a processor coupled to the memory, the processor configured to determine a ranking of the service users in the dataset based on an initial influence score for at least one of the users. Then, The ranking is revised based at least in part on the calculated influence score, and information is rendered to a target user based on the revised ranking. In [30], a novel algorithm for discovering the most influential nodes based on neighborhood diversity was presented. They introduce two new influential node ranking algorithms that use diversity of the neighbors of each node in order to obtain its ranking value. In [31], the authors propose a node ranking method based on the social conformity theory and community feature based on VoteRank. Their proposed method calculates the node influence capability from two points of view, one is the individual, the other is the group. Also, in order to improve VoteRank, they combined node attractive power, initiating power and the node selection strategy.

The article [32] proposes a new method to identify influential users in a social network by considering those interactions that exist among the users. Since users tend to act within the frame of communities, the network is initially divided into different communities. Then the amount of interaction among users is used as a parameter to set the weight of relations existing within the network. Afterward, by determining the neighbors,Äô role for each user, a two-level method is proposed for both detecting users' influence and also ranking them. The proposed LAR methods use the links between users to rank them. However, each user has her own characteristic of making links that affect link analysis. Hence, this feature has a direct impact on the ranking which is not considered. In this thesis, we will take into the account the impact of users personality (based on their opinions) and try to rank them in order to identify the most influential ones among them.

2.4 Opinion Propagation

Social network analysis [33] studies the relationships between social entities like members of a group. Thus, it enables us to inspect their structural properties such as links, neighbors, centrality and etc. In recent years, opinion propagation in online social networks has become a widespread phenomenon that indicates its importance. It can be even said that almost all social interactions are shaped by users beliefs and opinions [1]. Thus it is of high value to study opinion dynamics, and up to now, many researchers have proposed various models to analyze the evolution of the opinion dynamics and propagation from various aspects [2–4]. Current studies on opinion propagation are based on the links a user has in the network which means the users' opinion is affected by the opinion of her neighbors. However, the impact of neighbors are different, i.e. some of them have a greater impact. We consider such neighbors as influential users who are more popular or trusted among others.

Kou et al. [34] studied the opinion dynamics with multilevel confidences in Hegselman and Krause (HK) model by defining three clusters for the users namely, close-mind, moderate-mind and open-mind based on social differentiation theory. They divided the network into three sub-group but they did not consider the impact of each user. In another work, Liang et al. [35] considered the impact of both the bounded confidence and influence radius of agents on the opinion dynamics and they found that heterogeneity did not always promote consensus and there is an optimal heterogeneity under which the relative size of the largest opinion cluster reaches its peak point. Zhang et al. in [36] focused on mining features namely double propagation. They used two improvements based on part-whole and 'no' patterns to increase the recall. They applied feature ranking to the extracted feature candidates to improve the precision of the top-ranked candidates. Yang et al. in [37]

developed a linear influence model where rather than requiring the knowledge of the social network and then modeling the diffusion by predicting which user will influence other users of the network, they focused on modeling the global influence of a user based on the rate of diffusion through the network. Ximeng et al. in [38] propose a mixed diffusion-based recommendation model to enhance the performance of recommendation by using and mixing the similarity of explicit and implicit feedback. They took into the account the users feedback, however, the personality and impact of each user which plays a crucial in diffusion, are not considered. Cha et al. in [39] analyzed the information diffusion in the Flickr social network. they found that even popular photos do not spread widely throughout the network. Also, they implied that the information exchanged between friends is likely counted for some of the favorite markings but with a remarkable delay at each hop. Shang et al. in [40] proposed an opinion formation model under bounded confidence over multiplex networks, consisting of edges at different topological and temporal scales. They found that the existence of multiplexity prevents the convergence and that working with the aggregated or summarized simplex network is inaccurate since it misses vital information. Durret et al. in [41] considered a simplified model of a social network in which individuals have one of two opinions (called 0 and 1) and their opinions and the network connections co-evolve. They concluded that there is a discontinuous transition in the rewire to the same model so a small change in the dynamics of the model results in a large change in the qualitative behavior. In [42], a trust based recommendation mechanism is developed to generate advices according to individual trust relationship, making recommendations more likeable to be implemented by the inconsistent experts to achieve higher levels of consensus. In [43], the authors propose a method which comprises calculating a first feature vector for a first user, calculating a second feature for a second user and comparing the first feature vector with the second feature vector to calculate a similarity value. Then, a determination is made as to whether the similarity value falls within a threshold. If the similarity value falls within the threshold, a relationship is recorded between the first user and the second user in a first user profile and a second user profile. In [44] a novel trust-based approach for recommendation in social networks is proposed. In particular, the authors attempt to leverage deep learning to determinate the initialization in Matrix Factorization for trust-aware social recommendations and to differentiate the community effect in user's trusted friendships. A two-phase recommendation process is proposed to utilize deep learning in initialization and to synthesize the users' interests and their trusted friends' interests together with the impact of community effect for recommendations. Yucheng et al. in [45] proposed a strategy to reach the consensus opinion based on the leadership users. They implied that the consensus opinion is a linear combination of the opinions of the leadership. This confirms the importance of the influential users, however, we need a way to find these

users as they are not defined in most situations.

These studies focused on different consensus opinion formation models in the network regardless of the impact of the users who make the opinions and propagation. A user may have a tough personality and thus won't change her opinion easily or may have an unsophisticated personality so adapt her opinion quickly. On the other hand, in real life, the users make their opinions based on their links in networks (neighbors). The above studies, as well as HK model, used the current users' links to her neighbors for users' opinion propagation. However, the neighbors may have a different impression. As mentioned above in [34] the users divided into three clusters and the proposed propagation model based on each cluster manipulated. Nevertheless, each of the users of these groups may have a different impact on propagation. In this thesis, the opinion propagation is proposed considering the influence and impact of each user in the network.

2.5 Aggregation and Consensus Opinion

A consensus opinion problem is characterized by a group of users or experts who have their own knowledge, ideas, experience and motivation, and express their preferences on a finite set of alternatives on a happened problem to achieve a common solution. One of the main problem in reaching the common solution in dealing with incomplete information. In other words, there are some situations (such as when a new user comes to the network or we don't have the user's information) that we don't know what is the opinion of a current user. In this situation, considering the user's connection in the network, we can estimate her opinion by propagating the opinion of others toward her. Then the aggregation part comes to present the consensus opinion of the group. The other problem is inconsistency meaning that due to the different backgrounds and knowledge of the users, their opinion cannot be simply aggregated. Some of them who are the experts should have more impact on aggregation. As a result, we find the influential users, then propagate the opinions and finally aggregate the opinions to present the consensus opinion.

In [46], the authors propose a new aggregation strategy on the basis of the standard collaborative filtering. They then present an heuristic algorithm based on learning automata, called DLATrust, for discovering reliable paths between two users and inferring the value of trust using the proposed aggregation strategy. In [47], a new crowd opinion aggregation model is proposed, namely CrowdIQ, that has a differential weighting mechanism and accounts for individual dependence. In [48], the authors introduce a novel weighted rank aggregation method considering position based score, while ranking is done depending on various features like specificity, accuracy, sensitivity, etc. In [49] the visual trust relationship is constructed, a trust induced recommendation mechanism is investigated and an

interval-valued trust decision making space is developed to model uncertainty.

Jian et al. in [50] proposed a visual consensus aggregation model for multiple criteria group decision making with incomplete linguistic information. They propagate the trust between the users to estimate the opinion or trust of each of them and then aggregate their trust scores by induced order weighted averaging. In their methodology, they assumed that the experts are defined by the network. Nicola et al. in [51] tried to find an aggregated opinion among the experts considering the social impact of them on each other. In their model the concept of social influence is strictly interconnected with that of interpersonal trust according to the intuition that the more an expert trusts in the capability of another expert, the more his opinion is influenced by the trusted expert, especially in presence of incomplete information i.e. when experts are unable to express an opinion on any of the alternatives. Again in this work, the authors assumed that the influential users are defined by the network. Hengjie et al. in [52] investigated the heterogeneous large-scale GDM problem with the individual concerns and satisfaction in consensus reaching process, and proposed its consensus reaching model. In this model, the inconsistency of the users is not considered. Moreover, the authors focused on the personality of the users which makes a more accurate consensus model, however, the consistency is another problem which needs to be solved in any network for consensus problem. Maria et al. in [53,54] did a comparative study on consensus measures in a group of users. They analyzed how the use of different OWA operators (maximum, minimum and average) affects the level of consensus achieved through five of the most commonly used distance functions, Manhattan, Euclidean, Cosine, Dice, and Jaccard, once the number of experts of the users' network has been established. They found that the consensus degrees are deduced at the level of the relation and according to the number of experts considered, the aggregation operators and distance functions produce significantly different results in most of the GDM problems carried out. Yucheng et al. in [55] developed a method for consensus reaching process considering two paradigms namely trust relationships and opinion evolution. The first one analyzes the impact of the trust relationships in the different aspects (incomplete preference values estimation, aggregation, and feedback mechanism) while the second one which is based on opinion evolution involves two key elements (opinion evolution and opinion management).

The above two main problems in aggregating are usually solved with propagation. Nevertheless, most of the current works did not consider the impact of influential users and some of them assumed that these users are defined by default. Furthermore, some of the current studies did not propagate the information to remove the inconsistency problem. In our study, we consider both the impact of influential users and the inconsistency and incomplete user information problem which is solved by propagating the opinion in the network.

2.6 Summary

In order to solve the discussed consensus opinion problem, we proposed a methodology with three phases. This chapter presented the current studies for each one of the phases. Each phase is related to the previous phase which ends by presenting a model for the problem. The aim of the thesis is presenting a consensus opinion for the online shopping products. The consensus model is usually a score which is aggregated among the users. This aggregation cannot be reached unless the opinions are propagated in the network. However, a good propagation needs to know which users have more effect on others in which the influential users' impact are employed. In the current studies, the impact of influential users is not considered in propagation which consequently ended up with inaccurate aggregation for consensus problem. We proposed a method to present a more accurate consensus opinion by propagating the opinion in networks and considering the impact of influential users based on the link analysis.

Chapter 3

Finding the Influential users in the online social networks

Contents

3.1 Introduction

In the following chapter, we will describe the method used to find the influential users in the network using link analysis. As the first solution to the main problem, we need to find the influential users in the network.

In this era of the digital world, lots of people are registered in different social networks and take tremendous decisions based on their knowledge and information such as buying a product from online shopping websites or booking a hotel or restaurant. However, there is a tremendous amount of shared information and there is no mechanism yet to distinguish the validity of them in an accurate way. Thus, the knowledge shared by users on social networks could not be fully trusted. Recently, with the popularity of knowledge sharing in the social networks, this problem attracted lots of attention. One of the main direction in this domain is identifying influential users or experts and rely more on their opinion. By identifying the influential or expert users in social networks and determining their level of knowledge, the reliability of their provided information could be identified. Also in recommending systems, finding these users is important due to the fact that the preferable choices of them can be recommended to other users. The expert finding is one of the most important subjects for mining from (web-based) social networks. The task of expert finding is aimed at detecting the most influential and useful users in a network. These influential users defined as users who are more popular and more trusted among others. The problem of expert finding emerged many years ago to achieve reduced processing by selecting only influential users, achieve fast marketing query results, to address these users directly by 'targeted advertising' (so as to create public opinions or market awareness quickly and efficiently while spending much less on ineffective general advertising approach) and to improve the accuracy of the statistical results by avoiding the outliers and odd opinions contaminating the aggregated totals.

There are two approaches to find the influential users in the social networks. The first approach is analyzing the users' profile and the second one is users' link analysis. The links show the connection of users and the profiles show their personal information such as age, city, gender, the area of interests and etc. Most of the social networks include both link and profile information but with limitation to access them publicly. Link analysis is one of the common methods to analyze the users' connections and extract the needed information. Link analysis ranking (LAR) [56] is a method which ranks objects based on their links and the sign of links with each other. There are many studies which aim to rank the users considering their links and neighbors [25, 56–59]. However, to the best of our knowledge they don't consider the personality of the neighbors and most of them focused on unsigned networks and there are few studies on signed networks with positive and negative links

which are important to study the interactions in social media because the richness of a social network in most cases generally consists of a mixture of both positive and negative ones. The users are ranked based on their neighbors but how can we distinguish which neighbor link has more strength in current user ranking. Moreover, in signed networks, the link between users has positive or negative values. These signs present trust/distrust relation between users. The personality of each user has a direct impact on creating the signs of the links which affect the ranking calculations. In other words, the task of influential user detection by LAR method may greatly be affected by the personality of each user.

Here, considering the signs of the link, we first review most of the link ranking methods and then try to use the users' personality in ranking to find the most influential users. There are tremendous types of personalities in social science [60] which we use two main ones, namely optimism and pessimism (as user personality metrics) which can be calculated based on users propensity in relations (links) [61]. The optimism of a user shows how optimistic she thinks and in contrast, pessimism shows how pessimistic she thinks about the environment [62]. This personality feature is applicable to different ranking algorithms and we use these features in a sample ranking algorithm (PageRank) in order to verify the impact of them in ranking users and identifying the most influential ones. We call the new extended ranking algorithm as POPRank (Personality based on Optimist and Pessimist as a new feature for the ranking algorithm). In order to evaluate the rankings, we used the credibility criterion which relies upon the fact that better rankings should have more credibility values. The results showed that the added personality feature can effectively improve the ranking scores and has a meaningful impact on detecting the influential users. As described before, the researches for the problem of influential detection can be categorized into two main groups, (i) Authority ranking approaches and (ii) non-authority ranking approaches.

3.2 Ranking Algorithms

In social networks, there are two main approaches regarding the influential user detection problem as mentioned above. We introduce two measures as the personality of each user that can be added to any ranking algorithm in order to improve the performance of the ranking. First, we review most of the existing algorithms of ranking users including their shortcoming and then we will try to apply and utilize the personality measures on them. To this end, we take into account the sign of the links in signed networks and used Optimist and Pessimist scores of each user as their personality.

3.2.1 Existing Ranking Algorithms

According to the link analysis in social network, we first describe the baselines approaches and algorithms for the link analysis which rank the users in order to find the influential ones and then we will describe the proposed method that can be applied to ranking algorithm and effectively rank the users in order to identify the most influential ones.

1) ***In degree***: The most common and simple way to find the influential users is verifying the number of coming links in a particular domain network and label the users with the most in-degree as expert [25]. More positive links in a trust\distrust relations mean more expertise a user has on that network. This method is used when there is only the information about the connection between users. However, this method is not very accurate because it only considers the positive in-links without considering the users who made the links.

2) ***Popularity or Prestige***: This method is based on positive and negative links received by a user [63]. The main idea of Prestige is that the users who have received plenty of positive links should be ranked high and the ones who have received many negative links should be ranked low.

$$popularity_i = \frac{|IN_i^{(+)}| - |IN_i^{(-)}|}{|IN_i^{(+)}| + |IN_i^{(-)}|} \tag{3.1}$$

where $IN_i^{(+)}$ and $IN_i^{(-)}$ are positive and negative links received by user i respectively. Considering the signs of the links is a positive point of this method, yet it lacks utilizing the personality of the users who make the links in order to define a weight for links which indicate the importance of user's votes toward others.

3) ***Exponential ranking***: In this probabilistic algorithm, the negative links are taken into consideration [64]. The idea behind this ranking algorithm is to decrease the rank of the users if they receive negative links. Also, it relies on that the user links should not be distrusted if she has a negative reputation and in fact, they just need to be trusted less. Particularly, the users with negative reputation should not be assumed completely trust-less (as if she point negative to another user, we assume it as positive) instead, her judgment should be considered less. The expected reputation is calculated as $a = A^T P$ where a is a pillar vector, A is adjacency matrix, P is a positive definite pillar probability vector with $|P|_1 = 1$ which is calculated recursively as follows:

$$P(t+1) = \frac{exp(\frac{1}{\mu} A^T P(t))}{|exp(\frac{1}{\mu} A^T P(t))|} \tag{3.2}$$

where μ specifies the amount of noise in selecting the highest reputable judge. This algorithm emphasizes the importance of negative links and the fact that the enemy of a user enemy should not be considered as a friend. Indeed, their assumption which is based

on social balance theory does not consider the importance of both positive and negative labeled users who make the links.

4) ***HITS***: The HITS algorithm mainly relies on the fact that the way the links go has more information than just shared content [25]. This algorithm has two update rules namely authority and hub to rank the web pages. It assumes that each user has its own hub and authority value. Hubs are users which links to other users and authorities receive incoming links. First, an initial weight is assigned to hub and authority. Then in a specified repetitive iteration, the authority and hub will be updated until they converge as follows:

$$hub(i) = \sum_{j \in E_{ji}} authority(j) \tag{3.3}$$

$$authority(i) = \sum_{j \in E_{ij}} hub(j) \tag{3.4}$$

At the end of each iteration, weights are normalized under a norm such as In-degree, Salsa, Max-norm and etc. However, if a page makes several links to many good authorities, the hub score of it will be enhanced (so it will be ranked high).

Furthermore, it is worth mentioning that the HITS algorithm has two properties. It is symmetric, in the sense that both hub and authority weights are defined in the same way. Secondly, it is egalitarian, in the sense that when computing the authority weight of some page p, the hub weights of the pages that point to page p are all treated equally (same with computing the hubs weights). However, these two properties may sometimes lead to non-intuitive results. If the number of white authorities is larger than the number of black hubs, the HITS algorithm will allocate all authority weight to the white authorities, while giving little weight to the black authority and easily cause topic drift. However, intuition suggests that the black authority is better than the white authorities and should be ranked higher. Similarly, after computing, the middle black authority will have higher authority weight than the white authority, but actually, they should be equally good. Therefore, its method should seek to change the symmetric and the egalitarian of the HITS algorithm, and aim at treating links differently.

5) ***Bias and Deserve***: In this algorithm which is similar to HITS, the bias of a user is its tendency to trust/distrust other users and deserve of a user reflects the true trust a user deserves [65]. A user is biased if her tendency of making trust/distrust connection to other users is high. The algorithm can work for both signed and unsigned networks. The update rules of the Deserve and Bias are as follows respectively:

$$Deserve_i(t+1) = \frac{1}{|d^{in}(i)|} \sum_{k \in d^{in}(i)} [w_{ki}(1 - X_{ki}(t))] \tag{3.5}$$

$$Bias_i(t+1) = \frac{1}{2|d^{out}(i)|} \sum_{k \in d^{out}(i)} [w_{ki} - Deserve_k(k)] \tag{3.6}$$

where $d^{in}(i)$ is the set of all receiving links by user i and $d^{out}(i)$ is the set of all outgoing links from user i, w_{ki} is the trust score from user k to user i (the weight of the links between users which is 1 for positive links and -1 for negative links). $X_{ki}(t)$ represents the effect of bias of user k on its outgoing link to user i at time t and is computed as $X_{ki}(t) = max\{0, Bias_k \times w_{ki}\}$. This method suffers the same problem as HITS that is a user can show herself trustful if she rate users that deserve high positive values negatively and users that deserve high negative values positively which make her bias almost zero (trusted user).

6) ***PageRank***: The PageRank algorithm performs a random walk in a given network to rank the nodes based on their connections [57]. The PageRank algorithm was proposed in order to rank the web pages regarding their hyperlinks to each other. Consider we have $P_i, P_2, ..., P_N$ pages that should be ranked. The update rule of the algorithm is as follows:

$$PR(P_i) = \alpha \sum_{P_j \in M(P_i)} \frac{PR(P_j)}{L(P_j)} + (1-\alpha)\frac{1}{N} \tag{3.7}$$

Where $M(P_i)$ is the set of pages that link to P_i, $L(P_j)$ is the number of outgoing links from page P_j, N is the total number of pages and α is a damping factor. α is added as a coefficient to the formula to guarantee that the algorithm does not accidentally end up with an infinite series of PageRanks. For implementation, an initial ranking will perform to the nodes and then they will be updated until convergence. The original PageRank algorithm does not consider the negative links and in fact, it is created for unsigned networks ranking. Also, nature (personality) of the nodes are not considered which can effectively change the ranking scores.

7) ***PageTrust***: PageTrust is an extension of the PageRank algorithm which considers both positive and negative links. The idea behind this algorithm is to decrease the random walk encounters to the pages which have negative incoming links [66].

$$\begin{aligned} PageTrust_i(t+1) &= (1 - Z_{ii}(t)) \\ &.[\alpha \sum_{j,(j.i) \in G^+} \frac{PageTrust_j(t)}{|d_j^{(+)}|} + (1-\alpha)\frac{1}{N}] \end{aligned} \tag{3.8}$$

where α is damping factor as PageRank, G^+ is sub-graph of positive links, $d_j^{(+)}$ is outgoing links in positive sub-graph from node j and Z is a matrix which is calculated as $Z(t+1) = T(t)P(t)$, where T is the transition matrix at time t which is calculated as the row-normalized version of the sub-graph with positive links. P is the distrust matrix that

considering the negative links is calculated iteratively as follows:

$$P_{ij}(t+1) = \begin{cases} 1 & if(i \neq j; (i,j) \in G^- \\ 0 & if(i = j; (i,j) \in G^- \\ Z_{ij}(t+1) & otherwise \end{cases} \tag{3.9}$$

where G^- is sub-graph with negative links. The algorithm is promised to improve the PageRank accuracy by enabling it for both signed and unsigned networks yet it sustains the problem of PageRank to involve the personality of users.

8) ***Distance Algorithm***: This simple algorithm ranks the web pages based on their shortest logarithmic distance from each other [67]. The distance algorithm between two pages i and j is $Distance_{ij} = -log \prod_{S \in path(i,j)} \frac{1}{O(S)}$ where $O(S)$ refers to out degree of user S. Then the ranking score of page j is equal to $Rank_j = \frac{\sum_{i=1} Distance_{ij}}{N}$. This algorithm is as simple as in-degree which ranks the web pages based on their distance (number of edges between them). It can be used to rank the users based on their distance as well, yet it suffers the same problems as the in-degree method.

9) ***Ontology Ranking Algorithms***: This is the other branches of the ranking algorithm which is usually used in semantic web and tries to decrease the amount of overloaded data [68]. The main idea behind this algorithm is providing relevant information regarding a user query and rank the related information as high as possible so the searcher can easily access it. The problem of these algorithms is satisfactory of the users which are not guaranteed.

Considering all of the mentioned ranking algorithms, we noticed that the PageRank is the most common algorithm for ranking the users and observed that many existing ranking methods used this algorithm as their baseline for comparison. Hence, in this study, we consider PageRank as the base ranking algorithm and add the personality feature to it to verify its effect. As long as the Optimist and Pessimist score of each user is defined based on their in and out links, we consider the Prestige algorithm as another evaluation ranking algorithm which uses the in-links for the ranking calculation.

3.3 Personality as a ranking feature

We add personality as a new feature to the PageRank algorithm and propose a new ranking namely POPRank, in order to see how much this feature can improve the ranking. The PageRank has been originally proposed for networks with only positive links which is unable to be used directly for signed networks. We modified it to perform better and also can be used for signed networks. The added personality is consist of two social science features, namely Optimism and Pessimism which are added to the algorithm in order to improve

the ranking accuracy. In order to predict the links between users, we used Optimism and Pessimism concepts from social science.

3.4 Optimism and Pessimism

Optimist users are those who think positive about everything around them and make more positive (trust) links to other users. This personality makes the other users establish positive links to her as well. Therefore, an optimist user usually has both trust links to others and trusted links from others. In contrast, pessimist users are those who think negative about their environment and make more negative (distrust) links. We say that a pessimist user usually has both distrust links to others and distrusted links from others. We try to calculate the optimism and pessimism scores of the users from their rates (votes) toward external items (e.g. Epinion dataset). The optimist and pessimist scores of the users are defined and calculated in [69] which are used to rank the users in the sign prediction problem. We will use this definition and add them as a feature in the ranking algorithms (in the case of this study to the PageRank) to verify its impact on link ranking methods and influential user detection. The optimist and pessimist scores are defined as follows:

Consider there are N items $I_1, I_2, ..., I_N$, the set of items with low average rating scores rated by user u_i are:

$$OptLow_i = \left\{ I_k | r_{ik} \neq 0 \wedge \bar{r}_k \leq \frac{(1+z)}{2} \right\} \tag{3.10}$$

where r_{ik} indicates the rating score from user u_i to item I_k and $\bar{r}_k$ denotes the users average rating score toward I_k. If the rates are in the range of $[1, z]$, we consider scores in $[1, (1+z)/2]$ as low and $[((1+z)/2)+1, z]$ as high scores. The set of items which have low average scores and are scored high by user u_i are as follows:

$$OptHigh_i = \left\{ I_k | I_k \in OptLow_i \wedge r_{ik} > \frac{(1+z)}{2} \right\} \tag{3.11}$$

Likewise, the set of items with high average rating scores rated by user u_i are:

$$PessHigh_i = \left\{ I_k | r_{ik} \neq 0 \wedge \bar{r}_k > \frac{(1+z)}{2} \right\} \tag{3.12}$$

And the set of items which have high average scores and are scored low by user u_i are as follows:

$$PessLow_i = \left\{ I_k | I_k \in PessHigh_i \wedge r_{ik} \leq \frac{(1+z)}{2} \right\} \tag{3.13}$$

If the user u_i has rated above the average then she is more optimistic. Hence, the optimism score of user u_i is $Optimism_i = \frac{|OptHigh_i|}{|OptLow_i|}$. Accordingly, the pessimism score of

user u_i is $Pessimism_i = \frac{|PessLow_i|}{|PessHigh_i|}$. These two quantities will be used as a coefficient in ranking algorithms, therefore they will be normalized to the range of [0,1] in order to adjust the values and prevent diverge.

3.5 POPRank

The original PageRank algorithm is a vote by all the other pages to show how important a page is (a link to a page counts as a vote). In fact, it does not consider the users who make the connections. Using this algorithm, we consider each page as a user and take into account the validity of users who make a connection with a specific user. In other words, to calculate the rank score of a user, we consider the coming links (same as PageRank) and the personality of users making them in the POPRank algorithm. As mentioned above, optimism and pessimism are two quantities that provide us the possibility to measure the personality. The idea of using personality is that when an optimist user makes a positive link, her vote should be considered less (we will decrease her vote impact) and in contrast when she makes a negative link, her vote should be considered more (we will increase her vote impact). A similar theory is used for the pessimist user, meaning that, her negative votes will be decreased and her positive ones will be increased. In this ranking, we will apply PageRank separately on sub-graph with positive links G^+ and sub-graph with negative links G^-. The update rules of POPRank are as follows:

$$POPRank^+(P_i) = (1-\alpha)\frac{1}{N} + \alpha \sum_{P_j \in M(P_i)} \frac{PR^+(P_j)}{L^+(P_j)} \times Per_j \tag{3.14}$$

$$POPRank^-(P_i) = (1-\alpha)\frac{1}{N} + \alpha \sum_{P_j \in M(P_i)} \frac{PR^-(P_j)}{L^-(P_j)} \times Per_j \tag{3.15}$$

where $L^+(P_j)$ and $L^-(P_j)$ are the number of positive and negative outgoing links from node j, respectively. Similar to PageRank algorithm, it starts with some initial condition for both positive and negative PageRanks vectors and after enough iterations, it converges to the final rank vectors. In social science, a person can be an optimist or a pessimist. Taken this into account, we consider personality as follows:

$$Per_j = max\{Optimism_j, Pessimism_j\} \tag{3.16}$$

where the $Optimism_j$ and $Pessimism_j$ are the optimism and pessimism scores of the user u_j and are calculated as mentioned above. The final rank vector POPRank is calculated by:

$$POPRank(P_i) = POPRank^+(P_i) - POPRank^-(P_i) \tag{3.17}$$

The convergence of this algorithm is assured since it is the same as the standard PageRank algorithm with the same computational complexity.

3.6 Credibility as a measure to analyze ranking

The previous studies [70] indicate that the trust can emerge among users with two main factors: the first one is familiarity and the second one is the similarity. That is when the users know (familiarity) or resemble (similarity) each other, they trust each other more. Hence, these two measures can calculate the trust score toward the users in social networks which shows the credibility of them.In simple words, the credibility is the quality of being trusted and believed in by others. Credibility of the user in a network is more like a linear scale on which other users of the network give her a rating. It is a perceived quality that the users assign to her based on their interaction with her. It can be said that the credible users are someones who are more believable than the other ones. In addition, it is a measure that creates the definition of users' reputation. By comprehending its definition, we can understand how are the opinion leaders created, and why are role model adopted. The scientists found that there is a high similarity between the belief of the credibility and properties of the users' most admired leaders [71]. Hence, we can say that the credibility is trust value, or reputation of the users in the network which shows their leaderships. In the social networks, the credibility of a user is defined by familiarity and similarity.

$$Credibility = trust\ value = leadership = familiarity\ or\ similarity\ in\ social\ networks \tag{3.18}$$

In this study, we use similarity to calculate the credibility. W. hu et al. in [72, 73] used the similarity of neighbors to calculate the credibility and concluded that popular ranked users have more credibility. They also showed that users credibility of a network has a direct relationship with its ranking so it can be used to compare the rankings. We will use the credibility of users as the evaluation criteria which can confirm and verify the ranking outcomes. The credibility indicates the votes of a user's neighbor towards her. In other words, the credibility of a user reflects her expected trust value in the network. The value of the credibility does not consider only the number of coming links instead, it depends on their quality. The credibility of user u_i is calculated as follows:

$$\begin{aligned} Credibility(u_i) = \frac{1}{|M^i(u_i)|} \sum_{u_p \in M^i(u_i)} W_{u_p u_i} \\ .Sim(u_p, u_i).Credibility(u_p) \end{aligned} \tag{3.19}$$

where $M^i(u_i)$ denotes the set of all incoming links to node u_i and $W_{u_p u_i}$ presents the link weight from user u_p to user u_i. There are several methods such as the correlation coefficient,

the cosine similarity measure, and the euclidean distance that can be used to calculate the distance of two endpoints and return a quantitative value to represent the similarity between users. In the trust network, a user's similarity depends on its neighbors [74], while user tends to trust similar users like her. According to this, in this context, we use the *Jaccard Distance* to model the similarity between u_p and u_i, which is $Sim(u_p, u_i) = \frac{|F_p \cap F_i|}{|F_p \cup F_i|}$. $F_p \cap F_i$ is the set of two users common neighbors and $F_p \cup F_i$ is the set of two users total neighbors. Note that in signed networks, the weights $W_{u_p u_i}$ are -1 or +1 and the credibility value lies in the range of [-1,1] for such networks.

In order to calculate the credibility for a signed network, we divide the network to positive and negative sub-graphs, then we calculate the credibility for each of them and at the end, we subtract them to reach the credibility of each node of the network:

$$credibility(u_i) = credibility(u_i)^+ - credibility(u_i)^- \tag{3.20}$$

The original PageRank algorithm presents non-convergence issues for some topologies. For example, consider there are two nodes a and b that point to each other but not to other nodes, and there is a third node c which points to one of them. This loop will accumulate rank, but never distribute any rank to the first two nodes, as there are no outgoing links. The loop will form a sort of trap, also known as 'rank sink'. To handle this problem, we approximate the Weighted PageRank value $wpr(b)$ for a node $b \in V$ via an iterative process. The computation of $wpr(b)$ requires several iterations to adjust the approximation to the theoretical true value. In each iteration, the $wpr(b)$ value of each node $b \in V$ is computed as follows:

$$wpr(b) = (1-\alpha) + \alpha \sum_{a \in R(b)} wpr(a) w^{in}_{<a,b>} w^{out}_{<a,b>} \tag{3.21}$$

$$w^{in}_{<a,b>} = \frac{i_b}{\sum_{c \in R_a} i_c} \tag{3.22}$$

$$w^{out}_{<a,b>} = \frac{o_b}{\sum_{c \in R_a} o_c} \tag{3.23}$$

where α is a damping factor that is usually set to 0.85 [57], i_b is the number of incoming links of node b, i_c the number of incoming links of node c and R_a is the reference node set of node a. Accordingly, o_b is the number of outgoing links of node b and o_c is the number of outgoing links of node c. In each iteration, the wpr values for all nodes are reduced. Following the implementation of Gephi, a widely used toolbox for graphs, the iterative process stops when the following convergence criterion is satisfied for all nodes $b \in V$:

$$\frac{wpr(b)_{iter-1} - wpr(b)_{iter}}{wpr(b)_{iter}} \leq \xi \tag{3.24}$$

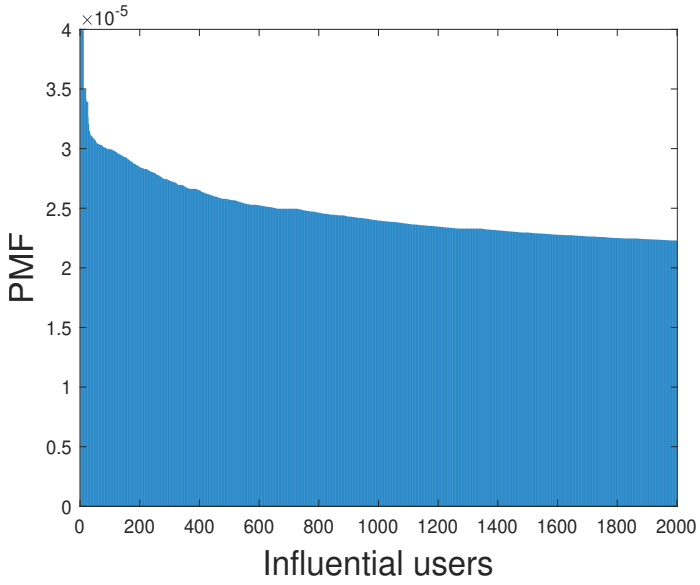


Figure 3.1 – The probability mass function of credibility in Epinions data-set

where the fraction is the normalized difference between the previous and the current iteration, and ξ is a predefined convergence threshold. Finally, after the algorithm converges, when $\forall b \in V$ above equation holds true, the p nodes with the highest wpr values are selected.

3.6.1 Credibility Mass Function

In order to generalize the proposed method, we also added the probability mass function (PMF) of the data-set. Figure 3.1 shows the PMF of the Epinions data-set. Note that here we depicted the PMF for 2000 users as they are influential (these users include the influential users who are introduced by the proposed method). We can see that the more influential users have better PMF (credibility). The slope of the PMF is steep at the beginning, which means that the first ranked users have a higher level of credibility and they are different from other users. Then, the slope becomes gentle for lower ranked users, which means that their ranks are same (there is still differences between their credibilities but they are not that important as the high ranked users). In other words, the differences between the credibility of first and 5th users are more than the one between 200th and 205th, because the differences of 200th and 205th users to be influential is not sensible compared with the 1st and 5th ones.

3.6.2 Real world data-sets

Our experiments are conducted in real world data-set. Consider that in these experiments, we used the data from actual real shopping website environments namely Epinions data-set (and later in section 4.4 Etsy data-set), however, it does not have a property to indicate the influential users. Therefore, we used the credibility to measure the validity of the found influential users by the algorithm. Our next experiment is gathering another data-set from online shopping websites which has the ground truth for influential users in which the network of users and products are accessible. Note that in these data-sets the nodes are actual customers of the real shopping environment. These users made the connection to other users of this online shopping website and there for build the network of the users. In addition these users rated the products and built the other network (network of users product). In the following section we will explain in details what are these networks and how we can conduct our experiments from each of them.

3.7 Experimental Results

In this section, we evaluate the proposed algorithm by using a real-world signed network in the context of influential user detection using a ranking algorithm involving each nodes personality. We also used personality (optimist and pessimist) of each node which is application dependent meaning that the definition and calculation of it can be varied in different data-sets. The evaluations have two parts. The aim of the first one is to show that the algorithm works correctly. In other words, the aim of first evaluation is not showing that our algorithm always, or in most of the cases, produces better rankings when compared to the baselines. Instead, we demonstrate that our algorithm produces such rankings that are useful in the sense that they produce rankings that are distinct and competitive with the ones produced by baseline and high quality of link analysis. The second evaluation compares the performance of the proposed algorithm with baseline ones using credibility. As we discussed, credibility is a criterion which can verify and show the validity of rankings.

3.7.1 Data-set

As we discussed, to evaluate our work we used Epinions data-set gathered from Stanford Large Network dataset Collection (SNAP)[1]. The Epinions website is a general consumer review site and its data-set consists of two types of ratings, trust relationship among users (members of the site can decide whether to trust each other or not) and users rating on items (the rate of users regarding the items of the website). This data-set includes 131,828

[1]https://snap.stanford.edu/data/

Table 3.1 – The main characteristic of the Epinion data-set

Total number of users	131,828
Total trust ratings	841,372
Number of filtered users	49,289
Trust ratings	507,592
Positive trust ratings	434,694
Negative trust ratings	72,898
Number of Items	139,738
Number of Items' ratings	664,824

users and 841,372 trust ratings. The main characteristic of the data-set is presented in table 3.1. In our evaluations, we did a filtering step and omitted who has no links and only considered 49,289 users with links with 507,592 trust ratings as links for the input of each algorithm (table 3.1).

Furthermore, please note that we evaluate our approach on two different kinds of real data. The first one is signed data-set (Epinions) and we discussed it in this chapter and the second one is unsigned data-set (Etsy) which we will present and conduct the experiments on the next chapter (please see 4.4.1). In addition, we tried to cover the real existing data-sets and the corresponding connection exist among their users in social shopping websites. Today, most of the online shopping websites such as Wanelo, Etsy, Fancy and etc. are unsigned and undirected meaning that the connection between users does not have a value nor direction. In fact, these connections shows a link between users and implies that these two users know each other. That is the reason why we added the Etsy data-set to our evaluations. In this section we try to depict and discuss the results toward the Epinions deta-set and later in 5.4 we present the experiment and result of the method on Etsy data-set (we will find the influential users of it) and present and discuss about the credibility of Etsy found influential users.

3.7.2 Evaluation

For the first part of the evaluation, we implemented the PageRank and Prestige algorithms to obtain our performance benchmark. We used 664,824 item rating by users in order to calculate the optimist and pessimist score of each user. In order to compare POPRank with PageRank and Prestige we use Spearman's rank correlation which measures the similarity of two rankings:

$$Similarity = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{3.25}$$

Here, x and y are rankings by two algorithms and $\bar{x}$ and $\bar{y}$ are average ranks. We compare the effectiveness of our proposed rank algorithm with the benchmark algorithms.

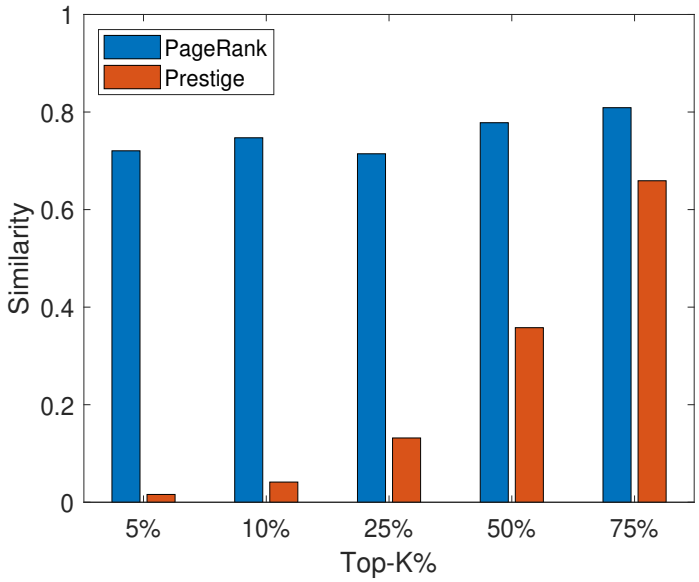


Figure 3.2 – Similarity of POPRank with each approach in found influential users. X axis represents different percentages of top found influential users and Y axis shows the similarity with POPRank

To compute the rank coefficient, a portion of the highest ranked nodes in the merged graph according to x are considered. As a default, we considered 10% highest ranked nodes but we also varied the target percentage (5%, 10%, 25%, 50%, and 75%) to observe how the accuracy varies with result size. For damping factor, we used $\alpha = 0.85$ as a parameter of the ranking algorithm. Also, it is worth mentioning that the runtime of the PageRank algorithm is $O(m + n)$ in which n is the number of nodes and m is the number of edges. POPRank has similar complexity to calculate the users' ranks. In particular, the proposed algorithm has more complexity to calculate the personality which is O(m+n+k) in which k is the number of item ratings.

Figure 3.2 compares POPRank with PageRank and Prestige in different percentages of top found influential users. To compare the algorithms we ignored the users which have no links to others and the ones whom item rating is not available because they have no impact on ranking algorithms. This Figure presents the percentage of common found influential users in different percentages of data between POPRank and the other two algorithms. The result is shown in table 3.2. This, confirms that POPRank performance is near to PageRank but far from Prestige. The similarity of POPRank and PageRank is maintained with different percentages of data.

We expect that the commonly found users should be increased if we consider more and larger percentage of data. In the Figure 3.2 the similarity of POPRank and Prestige

Table 3.2 – Common Found Influential users with POPRank

Top-K%	5%	10%	25%	50%	75%
PageRank	72.05	74.72	71.44	77.81	80.89
Prestige	1.60	04.14	13.18	35.79	65.91

increased when we added more data. However, in top-25% the similarity between POPRank and PageRank decreased. This can happen if we are comparing the similarity in the beginning or middle of the x-axis of this figure because for each next step of comparison (top-N%) the newfound users could be different, but it can not happen at the end of the x-axis because the users who are added are same. This decrement, indicate that the users found by POPRank and PageRank are different in top-25% of the found users.

The other perception of this experiment was the difference between POPRank and Prestige. Prestige is based on coming positive and negative links and the personality is based on the user votes (links) to the items. Nevertheless, the similarity of these two concepts did not affect the POPRank ranking. Particularly, the personality of the users involved in POPRank will not force it to be dependent only on the received links.

For the second experiment, we verify the ranking of nodes by all the three algorithms using the credibility values. The credibility of nodes is used as the criterion to evaluate and analyze the performance of algorithms. We say that nodes with more credibility should be ranked higher than those with less one. Taking it into account, we compare the top found nodes in different algorithms with nodes with more credibility. The evaluation was conducted with different percentages of top found nodes. We partitioned the result of each ranking algorithm in different percentages. For each percentage of found nodes, we sum up their credibility and compare it with different algorithms. Figure 3.3 shows the normalized credibility values of each algorithm for different percentages of top found influential users.

The Prestige algorithm is based on positive and negative links received by a user and PageRank is based on a random walk to rank the nodes based on their connections while POPRank considers the personality of each user as an added value to rank them. As is shown, the nodes that identified and ranked high by POPRank have more credibility for top-5% and top-10% in comparison to the others. In contrast for top-25% and rest, the PageRank has better credibility. This shows that for more influential users (top-5% and top-10%) POPRank has better performance. Also, as we observed in Figure 3.2, the similarity of POPRank and PageRank decreased in top-25% so we expect a meaningful difference in the credibility of them. The credibility increment of the PageRank in top-25% is beheld in Figure 3.3 (as we expected), showing that there is a meaningful difference between the found users by these algorithms here. The POPRank algorithm found the most influential users based on credibility in top-25% of its ranking whereas PageRank and

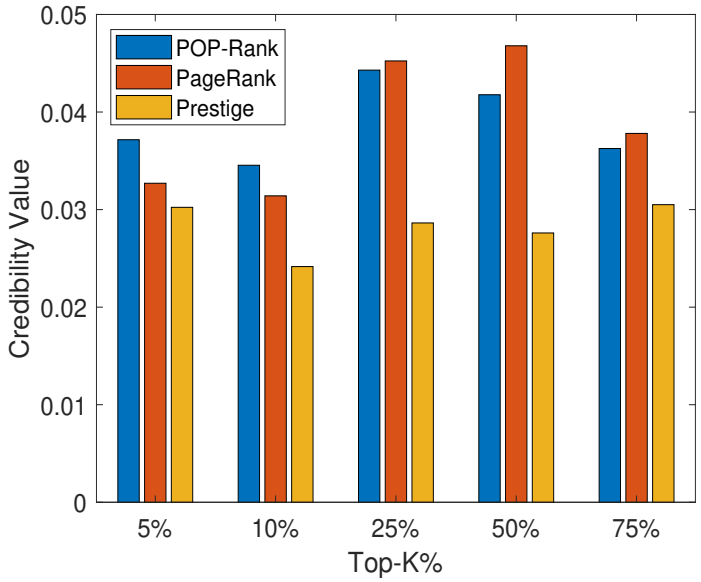


Figure 3.3 – Normalized Credibility of each ranking algorithm regarding different percentages of found influential users in Epinions. X axis represents different percentages of top found users and Y axis shows the normalized credibility value of found users

Table 3.3 – Comparison of PageRank and POPRank

Top-N	10	20	50	100	500
Common found	0	3	8	23	304
PageRank Credibility	0.017	0.024	0.028	0.031	0.035
POPRank Credibility	0.066	0.069	0.070	0.068	0.048

Prestige found it in top-50% and top-75% respectively. Overall, the POPRank algorithm has better performance in identifying the influential users within top-5% and top-10% and can find the most influential users within top-25% which is better than other algorithms. In other words, POPRank outperformed the baseline ranking algorithms such as PageRank and Prestige.

We noticed that the percentage of found influential users in POPRank have more credibility in comparison with PageRank and Prestige. This shows that leveraging the power of personality of each node can further improve the performance of expert finding. We also verified the Spearman correlation of more credible nodes with all algorithms. We found that the correlation for Prestige, Pagerank, and POPRank are 5.73%, 16.09%, and 19.50% respectively. This again confirms that the users found by POPRank algorithm have more credibility in comparison with others. Furthermore, in Table 3.3, we compared the performance of PageRank and POPRank in terms of the top ranked users.

In addition, although the similarity of them is high (72.05%) for top-5% of the ranked

users, it is different for the most top found ones. As we can see in this table, the number of commonly found users are quite low which indicates that POPRank makes a distinct ranking. To have a better understanding, Table 3.3 presents the normalized credibility of PageRank and POPRank as well. The comparison of the credibility values demonstrates that POPRank rankings always have higher credibility which indicates that it provides a better ranking. In the nutshell, the results show the positive impact of users personality in the rankings algorithm in order to find the influential users.

3.8 Conclusion

There are a huge amount of information and opinion on different topics and products shared on different social networks. However, one way to find the useful and trustful ones in relying on the influential users or experts whom can provide valuable shared knowledge. Toward this end, we first need to identify this set of influential users and to this end, the links between users and their profiles can be used. In this research, we used two features from social science namely Optimism and Pessimism to add the personality of each user in the ranking algorithms. We applied the user's personality to PageRank algorithm and created a new ranking POPRank for signed networks. Next, we compared the influential users found by POPRank with two baseline approaches of ranking. The result demonstrates the efficiency of the proposed algorithm.

3.9 Improving the method

This chapter proposed an algorithm for detecting influential users in signed networks. Two future directions can be taken into account: *(i) Trust propagation:* In the application of group decision making, when a problem occurs the group will discuss to find the solution. The fact is that the influential users are more trusted and has more effect on the final decision. We plan to investigate how a decision is made in a group by identifying the expert users related to the problem and propagating their information based on their trusted links. *(ii) Using profile information:* As a future guideline, we plan to use the profile of users in addition to their connections. In other words, the link analysis can identify the expert users but we plan to investigate the effect of each user profile on the accuracy of expert detection.

Chapter 4

Opinion Propagation in Online Social Networks

Contents

4.1 Introduction

In the following chapter, we will present our opinion propagation model which considers the impact of influential users. Following the second phase, we will investigate the propagation of the opinions considering the impact of influential users.

The web increasingly impacts the processes used by individuals to access preferences across items. A user may refer to the web to get information about TV, songs, movie tickets, jobs, or even mates. As these decisions and the fundamental financial processes move to the web, there is growing economic inspiration to propagate the information through the web. Open standards and a low barrier to publication demand novel mechanisms for validating information. Thus, we see unscrupulous exploitations of the holes in the social fabric of the web: successful manipulation of stocks by teenagers posting on investment boards under assumed personas; posts by product marketers pretending to be customers extolling the virtues of their product; online relationships that turn sour when one partner uncovers dramatic misinformation with respect to age or gender; link spamming of search engines to simulate popularity; and so forth.

Social networking websites have facilitated a new style of communication and information propagation through the links between their users. Today, millions of users participate in different social networks such as Facebook, Twitter, Amazon etc. and make a lot of social links with other users of that network. However, it is still difficult to determine the extent to which such users affect the opinions of the others. We claim that each user has a different impact on others which affect a given users' general opinion and be reflected in other links she may initiate. For example, in online shopping centers such as Amazon, eBay, Alibaba and etc. the most important issue is satisfying the customers to have a better and successful company. These websites have products and also the networks of their users in which each user can review them. While a user reviews the products, one of the most situations regarding these websites is when she wants to buy a product and she has an inquiry about it. The websites usually put the other users' reviews and opinions for the products to help the current user to decide. Thus, these websites can spread the effective opinions through their users. However, in order to convince the current user by others opinion, they should know which opinions have the better effect in order to assure her decision in shopping. In this case, these websites can recommend and highlight the specific users' opinions who they know their opinions have a positive and constructive impact on the products reviewers which will persuade the users to choose the product. In other words, the online shopping centers need to propagate the effective and positive opinions through the network of their users to have more benefit. Furthermore, with propagating the opinion, these websites can predict and realize the current and future opinions of their

customers regarding each product and adapt their products based on that.

4.2 Problem Definition and Solution

Consider an online shopping website and its users. A person may turn to these websites to buy a product. She will review the other users' opinions and take her decision regarding that product. In other words, the opinion of the user is based on other opinions and this will propagate and spread the other opinions in the network. However, the impact of influential users is not considered in propagation. Current studies on opinion propagation are based on the links of the user and her neighbors so that each users opinion will change based on the opinion of her neighbors. However, as we discussed before, the impact of neighbors are not the same, which means some of them have less impact while others have more impact on opinion propagation. We considered those users as influential users who have more effect on the opinion and decision of other users. Most of the current studies assumed that the experts are defined by the network and did not consider the impact of influential users in opinion propagation. Moreover, the links of a social network show the connection between users. In signed networks, the link between users has positive or negative values. These signs present trust/distrust relation between users.

In this chapter, we present an opinion propagation model based on the impact of influential users for signed and unsigned networks. Consider a shopping website that has products and users. These users are partially connected with each other (the network of users) and rated some of the products as their opinions. Thus, there are two kinds of links: 1) the link between users and 2) the rates or opinions of users toward the products. Two users are neighbors if there is a link between them. In signed networks, the links have positive and negative values and in directed networks the links have direction. In reality, the user decision is influenced by the opinion of neighbors, influential users and current users' knowledge. The first two can be considered through the link analysis and the third one can be determined by analyzing user' profiles. Here, we utilize the link analysis and considering the neighbors, we propose a model for opinion propagation based on the impact of influential users.

4.3 Propagation Methodology

This section consists of two distinct parts. In the first part, we review the base propagation model which leads us to a model appropriate for our case and in the second part we describe the proposed model.

4.3.1 Opinion Propagation Method

Opinion propagation is consist of several methods. There are several models and among them, the Voter model [75,76] is one of the promising ones which attracted a lot of attention. The Voter model is a stochastic process and assumes that there is an interaction between a pair of voters (users). The opinions of any given user on the same issue change at random times under the influence of the opinions of her neighbors. The model starts with an initial set of active users and for each user at time step t, one of her neighbors will be chosen at random and the user will assume the opinion of that neighbor. There are three opinion formation methods for Voter model:

Sznajd (S): It is used for discrete opinions, e.g. $+1$ and -1. In each time step t, two randomly selected users transfer their opinion to their neighbors if and only if they share the same opinion.

Deffuant (D): It is used for continuous opinions, e.g. in the range of $[0, 1]$. In each time step, one neighbor of the current user will be met and these two interact. The interaction will update the two users' opinions if the differences of their opinions are near to each other (confidence bound).

Karause and Hegselman (KH): The KH is used for continuous opinions. In each time step, one user is chose randomly and changes her opinion into the arithmetic average of her neighbor's opinions who are within her confidence bound. The user U_i will update her opinion x_i as follows

$$x_i(n+1) = x_i(n) + \frac{\mu}{N_i} \sum [x_j(n) - x_i(n)] \tag{4.1}$$

where μ is convergence parameter and is in the interval of $[0, 1]$ and N_i is the set of user U_i's neighbors.

Algorithm 1 is the pseudo code of the whole process of proposed opinion propagation. It has three main processes. First, we find the influential user's scores (ranks). Then, we propagate the opinions of users considering the impact and rank of users and at the end, we analyze the propagation with the Fuzzy Majority Opinion.

In this section, we will propose our opinion propagation model inspired by Voter model. The proposed model is based on link analysis which considers the connection of users in a network. Furthermore, the networks of interest are both signed (with positive and negative links) and unsigned ones. These properties convinced us to use Voter model as our baseline for opinion propagation since it uses the links between users and their neighbors. Moreover, we will use the results of Voter model in order to compare the performance of the proposed model.

Algorithm 1 Opinion propagation model based on influential users

Finding influential users in the network
  Personality definition
  Rank the users using SPRank
    Using Credibility to rank and find the influential users
Propagating the opinion
for each user i $[U_i]$ **do**
    Consider all the $neighbors_i$ and their ranks
    Set the neighbors whose opinions are in confidence bound U_i
    Updating the U_i opinion based on these neighbors
      $p_i^{t+1} = p_i^t + \frac{1}{|N_i^t|} \sum \mu \frac{SP_j}{SP_i + SP_j} [p_j^t - p_i^t]$
end for
Analyzing the propagation using Fuzzy Majority Opinion
for each product P_i **do**
    Let $A = a_1, .., a_n$ as the users' opinions toward P_i
    Let E as all the subsets of A
    for each subset E_i **do**
        Compute the Fuzzy Majority Opinion (FMO)
    end for
    Assign the dominant opinion as the opinion of P_i
end for
for each user U_i **do**
    for each opinion of U_i that changed **do**
        Consider the product P_j that U_i has opinion
        Consider the FMO of P_j
        Compare the new opinions with the FMO
    end for
end for

4.3.2 Opinion propagation using influential users

Usually the users' opinion changes by her prior opinion (initial opinion), the opinion of experts and friends (neighbors in the network). In our case, because the initial opinion is hard to access, we just consider the links and connections of the user to propagate the opinion. The problem of current propagation methods is that they don't consider the impact of influential users and the impact of neighbors (some of them have more and some have less impact on users' opinion formation). In OPIU, we consider the influential users and update the users' opinion with the fact that each neighbor has a different influence on current user final opinion in the network. First, we discuss about the social influence then we introduce the method of finding influential users in the network and at the end, we present the proposed model (OPIU). It is worth mentioning that, one way to enhance the influential users finding, is using the users' profiles. It is possible that users put some of their information and expertise in their profiles. This information can give us an additional value to find the influential users. For instance, in Facebook, the users put their favorites,

expertise, occupation and etc. and from this information the profile of the users will be created. Accordingly, a user who is working in the communities related to homes and buildings has more potential to be influential in hotels.com online accommodation website. Here we are not using them, and instead we use the links between users to build and find the influential users. The reason of considering just the links is that first, as mentioned before the link approach and the dynamics of networks based on the links attracted a lots of attention and second, today most of the introduced online shopping sites such as Wanelo, Etsy and etc. have the network of their users (the users' connections and links) but not their profiles. Hence, this encouraged us to use the link approach to find the influential users. Furthermore, we predict that the profile of the users will be added to online shopping websites and following that, we introduce their usage as the future work of this study. There are two approaches for using the users' profile: 1) using the already existed profiles by the users, 2) create the users profiles based on their activities and then use it. The first one will add another value which needs to be considered and computed based on its impact, and the second one needs to reviewing the users activities (if it exist) and based on that make an appropriate profile which can be used for the model. For instance, there are some studies as [77] in question answering websites such as Yahoo answers that use the statistical models based on topic analysis to create the profile of the users and find the most influential ones regarding a proposed question. This study is in the domain of information retrieve and far from what we are doing here, yet its method can be used in future study of this manuscript.

4.3.3 Social influence opinion propagation

Social influence is the process that users adjust/modify their opinions or change them because of their social interactions with others [78]. A simple process model acquired from the observations implied how opinions in a group of interacting users can shape or spread over repeated interactions. In particular, the studies in this domain identified two major attractors of opinion: (i) the influential user's effect, derived from the presence of highly confident users in the network, and (ii) the majority effect, induced by the presence of a group of users sharing homological opinions.

Indeed, it is difficult to measure how opinions alter under experimental situations, as it depends on many social factors such as the personality of the users, their bounded confidence level, their social status, their credibility, or their social power [79]. The present work draws upon experimental methods motivated by the concepts of opinion propagation in sociology and psychology. In the proposed model, we utilize the first major opinion attractor namely influential users as an effective factor on propagation. Also, we use the second one as a criterion for evaluating the opinion propagation (see section 4.3.6).

4.3.4 Influence impact on opinion propagation

Based on social influence studies, the people who are connected can change each other's opinion if their opinions are close enough. For example, the studies showed that the people sharing similar opinions have a strong tendency to amplify their confidence after interacting with each other [80]. Therefore, in our proposed model, for each user U_i we considered a set of neighbors whose opinions are not more than a certain confidence and then update the current users' opinion based on these neighbors. In reality, not all the users have the same influence on each other and some have social power e.g. have greater influence. One way to calculate the users' social power is their rank 3. The OPIU model comes from the previous chapter in which we employed the ranking of each user as their social power. This means that the users with more ranks have more social power and effect on others (we called them influential users). Here, for a signed network we employ that ranking algorithm (which is based on PageRank) to compute the ranking of the users based on their links:

$$Rank^{+}(U_i) = (1-\alpha)\frac{1}{N} + \alpha \sum_{U_j \in M(U_i)} \frac{PR^{+}(U_j)}{L^{+}(U_j)} \times Per_j \tag{4.2}$$

where $L^{+}(U_j)$ is the number of user j's positive outgoing links (similar equation is used for $Rank^{-}(U_i)$ 3.5). Similarly, the ranking of the users in an unsigned network is:

$$Rank(U_i) = (1-\alpha)\frac{1}{N} + \alpha \sum_{U_j \in M(U_i)} \frac{PR(U_j)}{L(U_j)} \times Per_j \tag{4.3}$$

The ranking algorithm starts with some initial conditions and it converges to the final rank vectors after enough iterations. In the formula, the Per_j is the personality of user u_j based on optimist and pessimist scores of the users defined in [69, 81]. The final social power rank vector (SPRank) for the signed network is computed as $SPRank(P_i) = Rank^{+}(U_i) - Rank^{-}(U_i)$ and for unsigned networks is as $SPRank(P_i) = Rank(U_i)$. These formulas compute the social power score (rank) of each user of the network. The users who have higher scores are more influential in the network.

4.3.5 OPIU model

We gave a score to each user of the network using their links (which is used to detect the influential users). Now we formulate the OPIU as follows:

Given a directed network G, we observe the decision of users toward a particular product over it. The user U's decision toward the product P is $Decision_{U \to P} = Function\{PK, C, R\}$ which PK is U's prior knowledge, C is the U's connection in the network and R is the review of others toward the product. There are two approaches to formulate a propagation

model through a network: 1) information effects [82], 2) direct-benefit effects [83]. Network models based on direct-benefit effects involve the following significant consideration: The user has certain social network neighbors and her benefits in adopting a new opinion increase when more and more of these neighbors pursue it. We consider this on the users' decision which consists of the users' connection. The connections consist of two kinds of impacts: 1) the impact of neighbors and 2) the impact of influential users.

Consider a weighted graph $G = (V, E, A)$ where V is the set of vertices with n users, E is the set of directed links, and A is the adjacency matrix. A neighborhood matrix G^t is used to represent the social relationships on A at time t. For all i, $j \in A$, $G^t_{ij} \in \{0, 1\}$ shows if there is a directed link from user i to j at time t. So the nxn matrix G^t is specified as:

$$G^t_{ij} = \begin{cases} 1 & if\ i\ pays\ its\ attention\ to\ j \\ 0 & otherwise \end{cases} \tag{4.4}$$

where $G^t_{ij} = 1$ denotes user i can receive an opinion from a supplier user j. In fact, we assume that each user is always connected with itself, i.e. $G^t_{ij} = 1$, for all $i \in A$, all t. G^t is asymmetric to describe a directed network, so that $G^t_{ij} \neq G^t_{ji}$, for some i, j. A user $i \in A$ only observes herself and her neighbors, including the users in the set of $j|G^t_{ij} = 1$ for all j, at time t. The opinions of n users at time t are appeared by an $1 \times n$ vector $P^t = (p^t_1, p^t_2, ..., p^t_n)$, where p^t_i is the user i's opinion at time t, $p^t_i \in (0, 1), i \in A$. We define dif^t_{ij} as the difference between opinion p^t_i and p^t_j: $dif^t_{ij} = |p^t_i - p^t_j|$ where $|p^t_i - p^t_j|$ is the absolute value of $p^t_i - p^t_j$. Obviously, we have $dif^t_{ii} = 0$ and $dif^t_{ij} = dif^t_{ji}$. Furthermore, we define w^t_{ij} as the weight of the influence of j on i.

$$W^t_{ij} = \begin{cases} 1 & if\ dif^t_{ij} \leq \epsilon and G^t_{ij} = 1 \\ 0 & otherwise \end{cases} \tag{4.5}$$

where ϵ is the confidence level (CL) and $w^t_{ii} = 1$ for $\epsilon \geq 0$ for all t and i. Each user will update her opinion by taking the average of all opinions which lie in her CL including her opinion at each time step t. The element p^{t+1}_i of new opinion vector P^{t+1} is calculated as:

$$P^{t+1}_i = \sum_{j=1}^{n} \frac{w^t_{ij}}{\sum_{a \in A} w^t_{ia}} p^t_j \tag{4.6}$$

The P vector will keep updating until it converges. The convergence criteria is

$$\sum_{i=1}^{n} (p^{t+1}_i - p^t_i)^2 \leq \xi \tag{4.7}$$

where ξ is a very small positive number (e.g. 10^{-4}). Also, the influential users have great influence on other individuals in the society but, their opinions are hardly influenced.

Let us suppose that there are M users and K of them are influential ones. We consider social power scores so the update rules for user i with p_i opinion will be:

$$p_i^{t+1} = \begin{cases} p_i^t + \frac{1}{|N_i^t|} \sum \mu \frac{SP_j}{SP_i + SP_j} [p_j^t - p_i^t], & N_i^t \neq 0 \\ p_i^t, & Otherwise \end{cases} \tag{4.8}$$

Where $N_i^t = j||p_i^t - p_j^t| \leq \epsilon_i$ is the opinion neighbor set of user i at time t and $|N_i^t|$ is the cardinality of N_i^t. The other consideration regarding the impact of neighbors is that we should consider only the neighbors who have greater weight in the connection because in real life, a user will be impacted by close friends.

4.3.6 Fuzzy Majority Opinion

In order to evaluate our experiments, we used the concept of the majority opinion (section 4.4.3). First, we indicate to some extent the Fuzzy Majority Opinion can be computed and then we present its use to evaluate the opinion propagation.

There are two common ways to compute majority opinion [84], namely aggregation operators and fuzzy method. Here we used the fuzzy method which provides in addition to a value for the majority opinion a sign of the strength of that value as a delegate of the majority opinion. To do so, consider $A = a_1, ..., a_n$ be a set of values which establish the opinions of the users.Let E be a crisp subset of A. The first step is to specify the degree to which this is a subset carrying a majority opinion. A subset E holds a majority opinion if all the elements in E are similar and the cardinality of E satisfies the idea of being a majority of elements from A. Let $MOP(E)$ implies the degree to which the elements in E, form a majority opinion, are a majority of elements from A with similar values. Thus, $MOP(E) = Q(\frac{|E|}{n}) \wedge Sim(E)$ where $\wedge$ shows the min operator and $Sim(E)$ is equal to $Min_{a_i, a_j \in E}[Sim(a_i, a_j)]$. Then, $Opi(E) = Average(E) = \frac{\sum_{a_i \in E} a_i}{|E|}$ is the opinion of the elements in E which is the mean value of the elements involved in E. Using the above concepts, the fuzzy majority opinion FMO indicating the majority opinion of the set of elements in A is defined as:

$$FMO = \bigcup_{E \subseteq A} \left\{ \frac{MOP(E)}{Opi(E)} \right\} \tag{4.9}$$

So for each subset E, the value $MOP(E)$ indicates the degree to which the quantity $Opi(E)$ is a majority opinion. Also, following similarity relation is assumed:

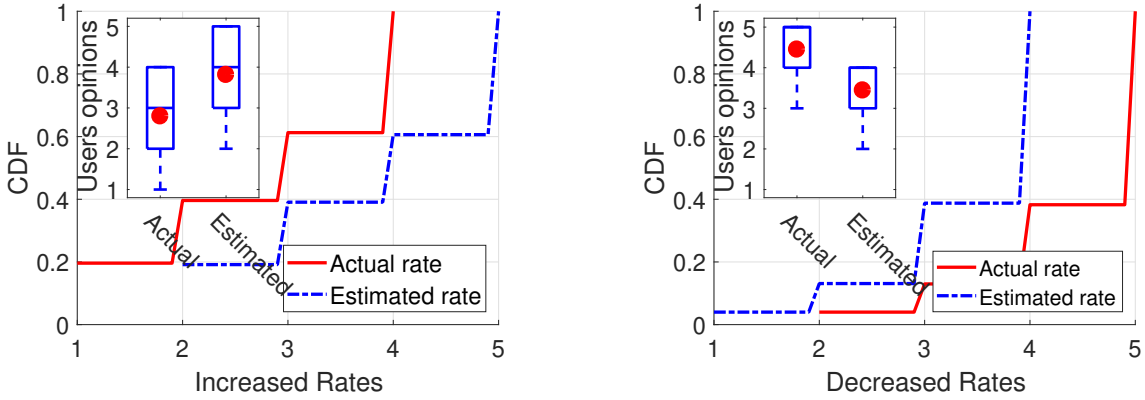
$$Sim(a_i, a_j) = \begin{cases} 1 & if |a_i - a_j| < \sigma \\ \frac{2\sigma - |a_i - a_j|}{\sigma} & if \sigma < |a_i - a_j| < 2\sigma \\ 0 & otherwise \end{cases} \tag{4.10}$$

where σ is the standard deviation of $a_1, ..., a_n$. Furthermore, for the formal definition of the quantity (Q), a definition of a majority in terms of a fuzzy subset Q is defined on the unit interval. In particular, $Q : [0, 1] \rightarrow [0, 1]$ such that $Q(0) = 0$, $Q(1) = 1$ and $Q(x) \geq Q(y)$ if $x > y$. $Q(x)$ is defined as below:

$$Q(x) = \begin{cases} 0 \quad if x \leq 0.4 \\ 5(x - 0.4) \quad if 0.4 < x \leq 0.6 \\ 1 \quad otherwise \end{cases} \tag{4.11}$$

4.4 Experimental and Results

In this section, we evaluate the OPIU using real-world networks within opinion propagation. The evaluation consists of two main parts. First, we present the details of the datasets and discuss our observation on OPIU and then we evaluate the OPIU performance using FMO.



(a) CDF of increased rates

(b) CDF of decreased rates

Figure 4.1 – CDF of Epinions dataset

4.4.1 Data-sets

To evaluate our work we used two directed datasets namely Epinions (signed) and ETSY (unsigned). The first one is the same dataset used in Stanford collection [1] and the second dataset is crawled by our crawler.

Etsy Data-set: In addition to the Epinions, we use the Etsy for our experiments. Etsy is a peer-to-peer e-commerce website covering a wide range of products on handmade or vintage items and supplies, as well as unique factory-manufactured items. Etsy's top three competitors according to Hoovers Online are Amazon Handmade, Craigslist, and

[1] https://snap.stanford.edu/data/

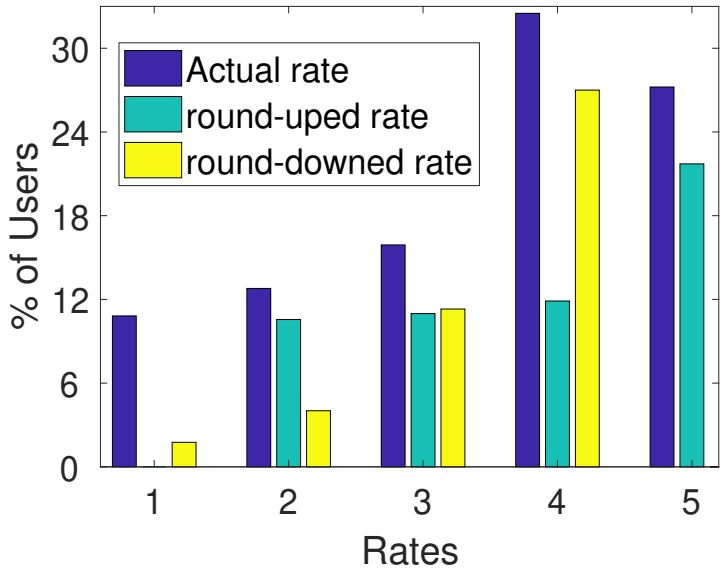


Figure 4.2 – Actual rates, round-up and round-down rates in Epinion Dataset

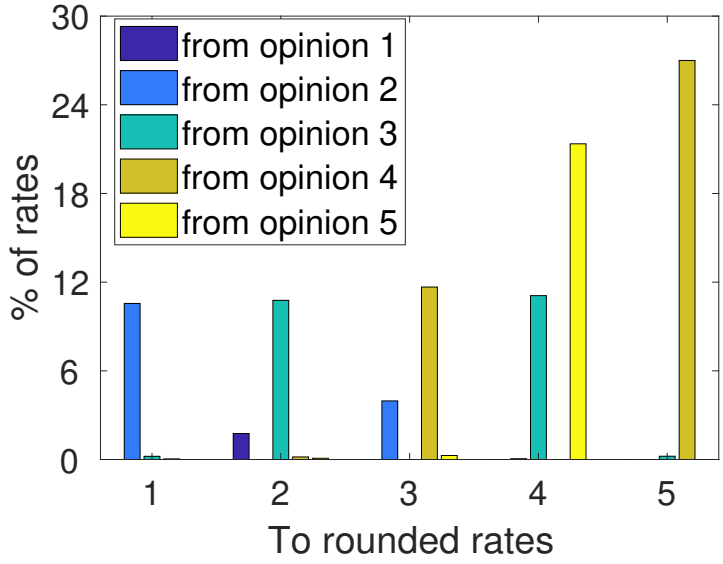
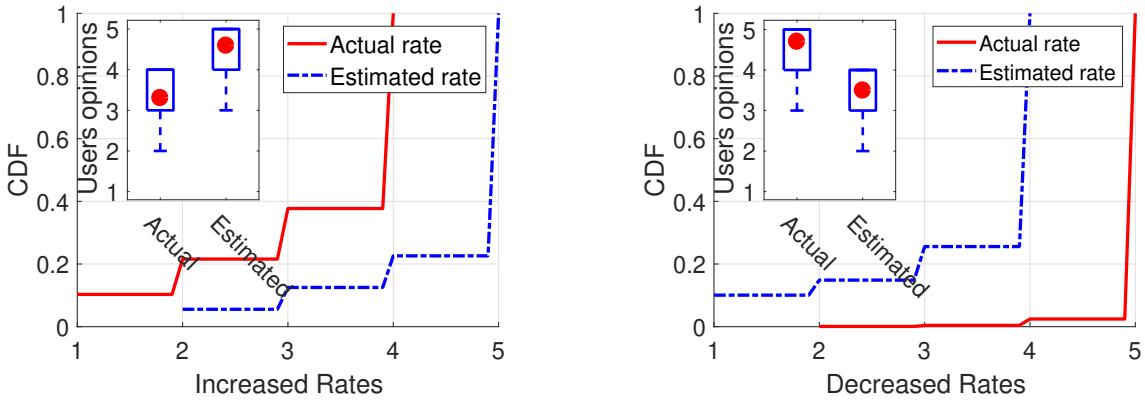


Figure 4.3 – Users rates changed to other rates (rounded) in Epinion Dataset

eBay. This website includes the network of users (unsigned links between users) and the user's opinions toward the products. Same as Epinions, the users' opinions toward the products contains the integer values between 1 and 5. This dataset is crawled from "https://www.Etsy.com". In general, Etsy has six main categories and due to the huge



(a) CDF of increased rates

(b) CDF of decreased rates

Figure 4.4 – CDF of Etsy dataset

number of users, we selected one category (namely Home) for our experiments. The crawler is programmed in C# which goes to the product pages one by one and collect the user's opinions for them. Then, for each user of the product, it collects the followers (who have a link to the current user) and following (who the current user has a link to them) in order to establish the network of the users. The most challenging part was the time consumed for crawling the users due to the fact that there is a huge number of links which took a month with a core i7 CPU and 16GB RAM computer. We crawled 239,237 users with 4,618,783 links. Same as Epinions, we did a filtering step to omit the users who had a few numbers of links. The main characteristic of Etsy and Epinions data-sets are presented in table 4.1.

Table 4.1 – Epinions [85] and Etsy Data-sets Characteristics

Characteristics	**Epinions**	**Etsy**
Total number of users (crawled)	131,828	239,237
Total links (edges)	841,372	4,618,783
Number of filtered users	49,289	72,528
Filtered links	507,592	1,914,852
Positive links	434,694	-
Negative links	72,898	-
Number of Products	139,738	24,362
Number of Products' ratings	664,824	200,148

4.4.2 Observations

We evaluated the proposed method from three different levels, namely opinions, users and products. In case of the opinions, we consider each opinion from users toward the products

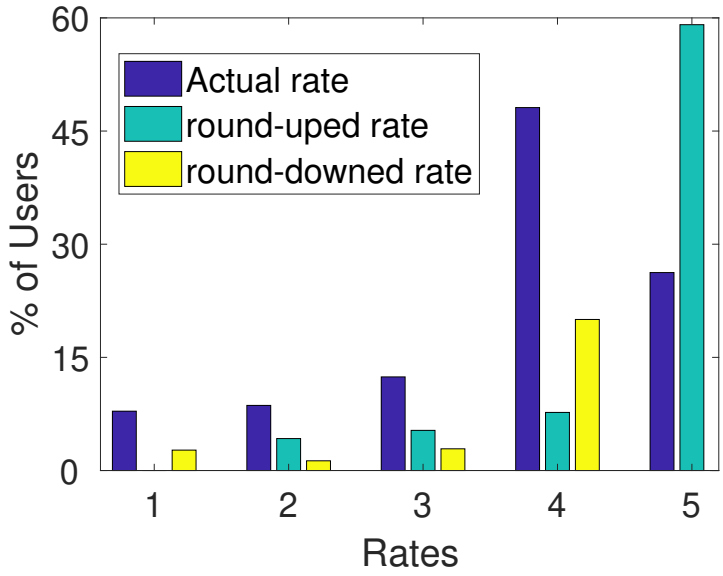


Figure 4.5 – Actual rates, round-up and round-down rates in Etsy Dataset

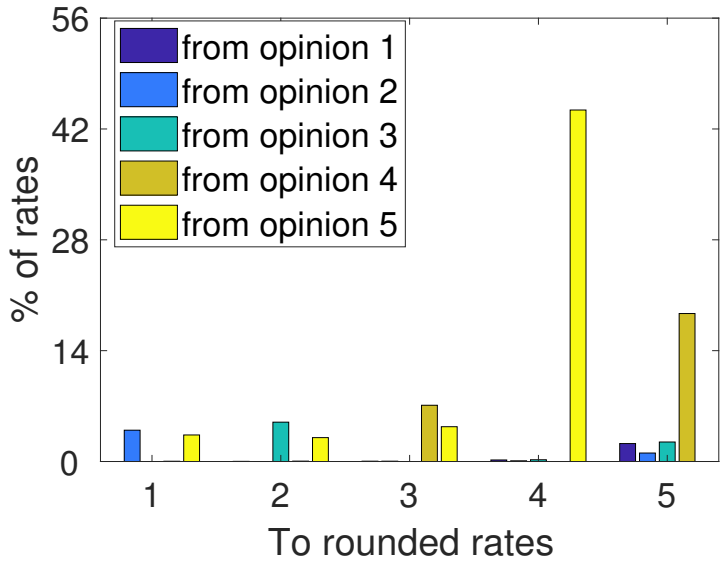


Figure 4.6 – Users rates changed to other rates (rounded) in Etsy Dataset

and analyze the differences between actual and estimated ones. In the case of the users, we analyze the users and their opinion changes. And, in the case of the products, we analyze the rates of products which changed significant and also analyze the users who made this changes. These levels provide us the different visions and understanding of the impact of

influential users on the opinion propagation.

In the formula, when $\mu = 0$, it means user i will never consider other users' opinions (we can treat it as a leader). Without loss of generality, we assume $\mu = 0.5$ in our experiments and in order to find the neighbors of a user, we considered the outgoing links of her (those users who have a link from the current user are the neighbors of her).

4.4.2.1 In the level of rates

The rates of users are the users' opinion toward the products. The number of rates from users to products has the mean value of 13.49 in Epinions and 19.53 in Etsy. We observed from Epinions that 2% of users have no rate, 3.5% have 1 rate, 65% have 1 to 10 rates and 97% have 1 to 100 rates toward the products. Also, for Etsy dataset 5% of users have no rate, 11% have 1 rate, 45% have 1 to 10 rates and 91% have 1 to 100 rates.

In our experiments, some of the opinions changed with OPIU while others remained unchanged (OPIU succeed to change 25.57% of the Epinions and 21.43% of Etsy rates). The rates whether increased or decreased. Hence, we separated the rates into two subgroups i.e. increased and decreased rates to analyze them. Figures 4.1a, 4.1b, 4.4a and 4.4b show the increased and decreased rates to compare the actual and estimated rates for datasets. These figures indicate that the users who gave low and high rates to the product tend to make their rates lower and higher respectively.

Figures 4.2 and 4.5 illustrate how much percentage of users have different rates toward the products. Note that the rates are in the range of 1 to 5 and we compared the actual rates with estimated rounded ones. We observed that the changes are mostly ascending i.e. from lower rates to upper ones and it means that users often tend to be positive rather than negative.

Figures 4.3 and 4.6 illustrate how many rates of users are changed to other rates (considering that we examined this with rounded estimated rates). These figures show the changes in the rates are normally smooth (and not a big jump). Generally, these figures convey that it is hard for users to change their opinions to other ones which are very far from theirs.

Figures 4.7 and 4.10 show the spread of estimated rates of users in comparison to actual ones. Note that the red circles are the average rate of each estimated column. These figures show that our model significantly changed the user's opinion.

4.4.2.2 In the level of Users

In order to count the neighbors of a specific user, we considered the links from the current user to her neighbors (the user out-going links). In case of the Epinions dataset, if the link is negative, we consider the neighbor as a negative neighbor and otherwise positive neighbor.

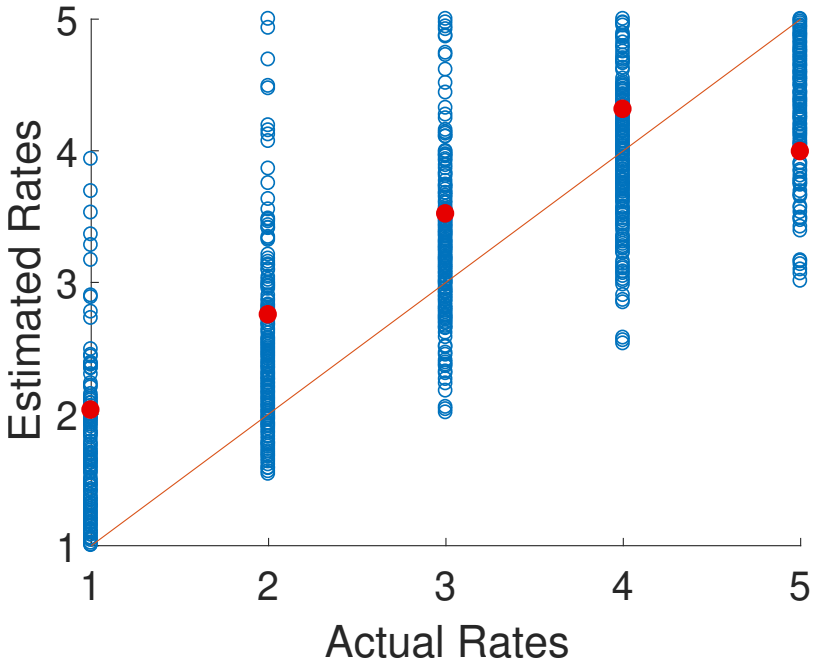


Figure 4.7 – Epinions dataset - The spread of different rates of users. The red circle is the average of estimated rates

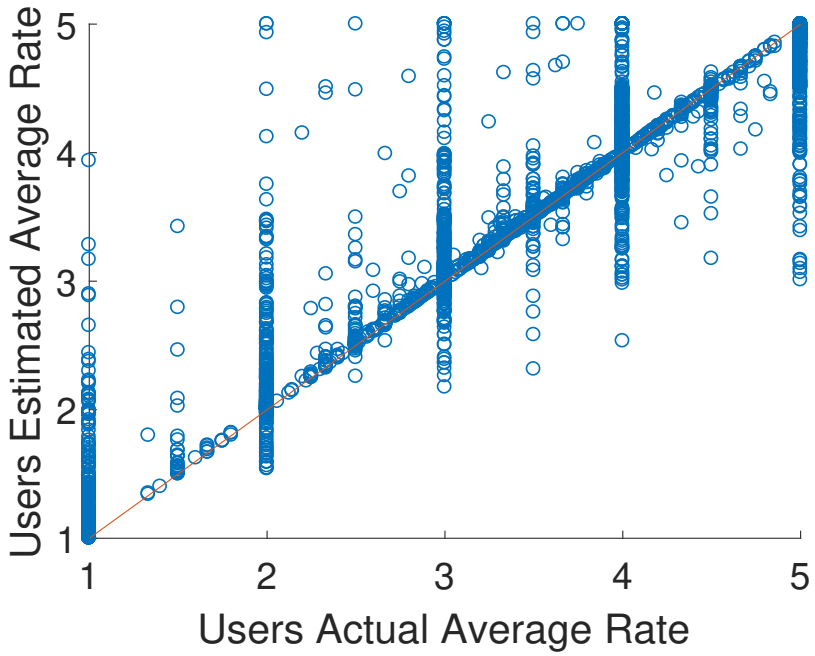


Figure 4.8 – Epinions dataset - The spread of different average rates of each user.

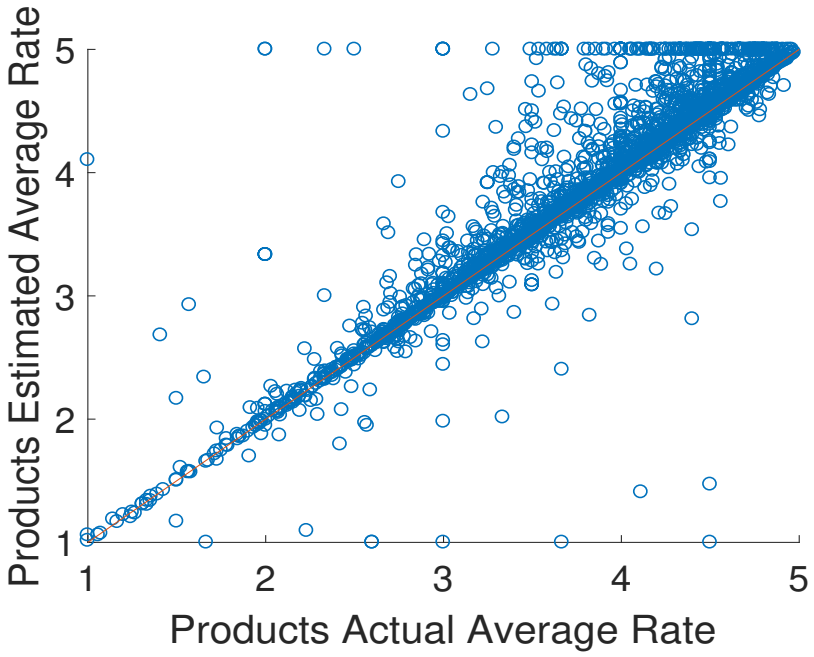


Figure 4.9 – Epinions dataset - The spread of different average rates of products.

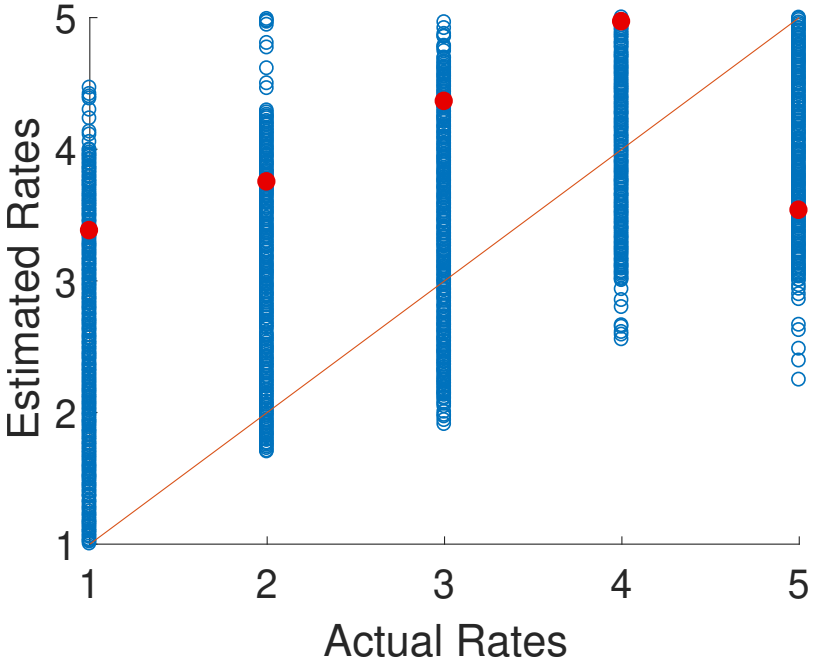


Figure 4.10 – Etsy dataset - The spread of different rates of users. The red circle is the average of estimated rates

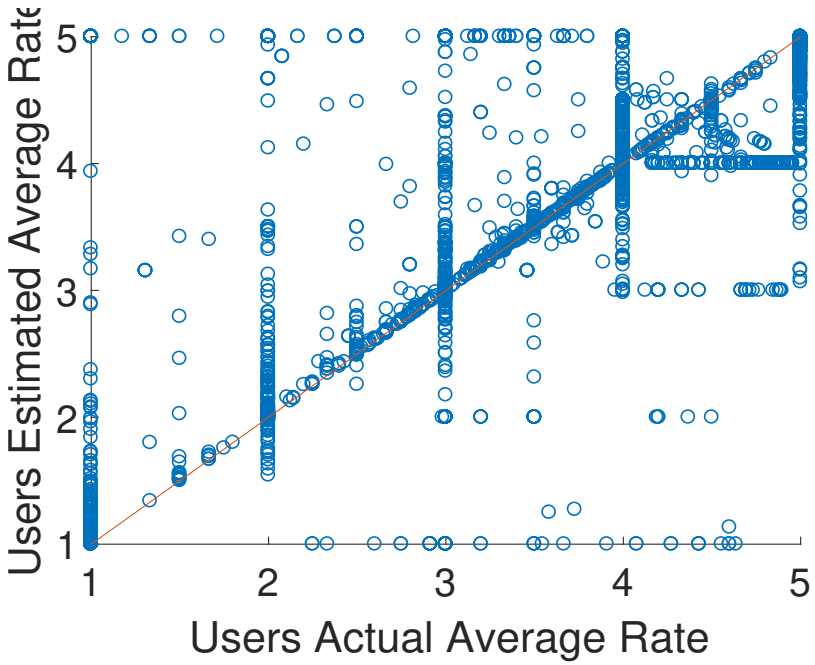


Figure 4.11 – Etsy dataset - The spread of different average rates of each user.

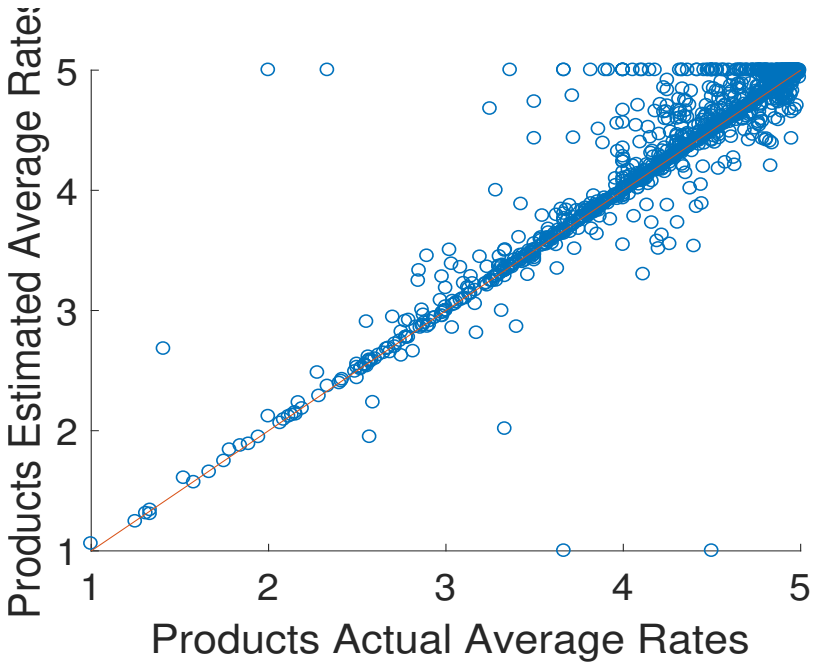


Figure 4.12 – Etsy dataset - The spread of different average rates of products.

The average number of neighbors for each user is 10.29 (with 8.81 positive neighbors and

1.48 negative neighbors) in Epinions and 27.87 for Etsy. Figures 4.8 and 4.11 show the spread of average rates of users in their networks. These figures indicate how the average rates of users changed. As we can see, most of the changes are near to the actual rates which means that the average rates of user change around the actual rates.

4.4.2.3 In the level of Products

Figures 4.9 and 4.12 show the spread of average rates of products in the network and indicate how average rates of products changed. The changes are mostly near to the actual rates with a small fraction of users who made big changes. One interesting observation is analyzing the products which their average opinions changed significantly. Among the Epinions products, we found 22.68% of them have big jumps (significant difference between estimated and actual average opinions). This percentage is 26.59% for Etsy. Moreover, we investigate the products which have the most significant changes (top 2% of them) in order to evaluate their opinions and the users who rated them. We observed that the users rated these products have some neighbors who have high scores in the network which we tagged them as influential users in section 4.3.4 (the scores of the neighbors are in the range of top 5% ranks of users). This observation implies to the fact that influential users are involved in changing the average opinion of the products and indicates the impact of them in propagating the opinions.

4.4.3 FMO Evaluation

According to the third main part of the algorithm, we use Fuzzy Majority Opinion characterized in section 4.3.6 to evaluate our methodology. The opinion of group members will lead to the majority opinion [78]. In this study, they showed that the opinion of users will change to the opinion which is the majority opinion of the network users. We use this as a criterion for evaluating the estimated opinion propagated by OPIU model. In other words, if the estimated opinions lead to the majority opinion (getting near to it), the model is working properly. We compare the Voter model and the proposed OPIU model considering the Fuzzy Majority Opinion. To do so, we first assign the dominant opinion for each product and then compare the OPIU computed opinions of users toward products with the FMO of each of them as algorithm 1. Note that here we used two similarity relation function (Sim). The first function (Maj_Op1) is described in equation 4.10 [86] and the second one (Maj_Op2) is defined as follows [84]:

$$Sim(a_i, a_j) = \begin{cases} 1 & if |a_i - a_j| < 2 \\ \frac{1}{2}(4 - |a_i - a_j|) & if 2 < |a_i - a_j| < 4 \\ 0 & otherwise \end{cases} \tag{4.12}$$

Table 4.2 – MSE of OPIU and Voter opinions with normal and Fuzzy Majority Opinion

Dataset	Model	Maj_Op1	Maj_Op2	Mode
Epinions	OPIU	0.77	0.71	1.23
	Sznajd	2.21	2.09	2.98
	Modified Sznajd	1.97	2.07	2.65
	Deffuant	1.39	1.32	1.73
	Voter Model	1.19	1.13	1.66
Etsy	OPIU	0.89	0.86	1.19
	Sznajd	2.54	2.41	3.19
	Modified Sznajd	2.31	2.25	2.55
	Deffuant	1.48	1.53	1.69
	Voter Model	1.36	1.29	1.44

Table 4.2 shows the mean square error (MSE) of OPIU and voter model with Fuzzy Majority Opinion for both data-sets. Note that this computation is done towards the users' opinions which are changed (there are some opinions which remain unchanged after applying the method). In addition to the baseline method (Voter model), two other methods namely Sznajd and Deffuant are implemented to compare their results with the proposed method's (OPIU). In Sznajd, two users will be selected randomly, and if their opinions are same their neighbors will have that opinion. In addition to the Sznajd method, we used a modified Sznajd and put a threshold (THR = 1) to check the opinion similarity of the selected users. If the U_i has the opinion op_i and U_j has the opinion op_j then:

$$Opinion(N_i) = Opinion(N_j) = op_i \quad if(op_i == op_j) \tag{4.13}$$

where N_i contains the neighbors of U_i. and in modified Sznajd

$$Opinion(N_i) = Opinion(N_j) = max(op_i, op_j) \quad if(|op_i - op_j| <= THR) \tag{4.14}$$

Note that, in modified Sznajd, if the opinions of the selected users are not equal but their difference is less than the threshold, we choose the maximum opinion between op_i and op_j to propagate it towards their neighbors. The reason is that we observed that the users usually tend to be positive rather than negative, and hence their opinions will change to the greater opinions (we discussed it above in section 4.4.2.1). The Deffuant model is one of the most studied in socio-physics. It has been used in various communication topologies, starting from fully connected network, where any agent may communicate with any other grid models with limited range of interactions and complex social networks [87]. The bounded confidence approach forms a special class of opinion change models, characterized by continuous distribution of possible opinions within a given range and a simple, intuitive mechanism for individual opinion changes. In Deffuant, for each user a random neighbor

will be chosen and if the difference of their opinions is too big, the communication process is impossible, and there is no change in the respective opinions as the effect of the interaction. On the other hand, the model postulates that if the initial opinions are close enough, the interaction between the agents brings them even closer. In other words, if their opinions are in the bounded confidence (near enough to change each other opinions) these two will interact based on their actual opinions.

$$x(i) = x(i) + \mu[x(i) - x(j)]x(j) = x(j) + \mu[x(i) - x(j)] \tag{4.15}$$

where $x(i)$ and $x(j)$ are the opinions of U_i and U_j, respectively. The μ can be in the interval of $[0, 1]$ and determines the speed of convergence of the opinions and here we assumed it as 0.5. We can see that for Epinions and Etsy, the modified Sznajd has better MSE rather than Sznajd, which means that the opinions of the users are changed in the correct way. Considering that we computed the modified Sznajd by using the maximum opinion, this confirms that the users tend to be positive. The Deffuant method in general is performing better than the Sznajd and modified Sznajd. However, The OPIU and Voter are performing better than Deffuant. This result shows that the estimated users' opinions are leading to the Majority Opinions and the MSE of OPIU is better than other methods for both datasets. This confirms that the performance of OPIU is better than baseline propagation Voter model.

4.5 Conclusion and Future Works

Today people take lots of decisions in their lives such as shopping, where to go for a trip, renting a hotel and etc. Some of these decisions are made online and as the statistic shows, the tendency of people for online shopping is growing day by day. Users of a website usually take the decision about its products based on the current information they have, the opinion of their neighbors and influential users of that website. There are a lot of studies paid attention to the dynamics of opinion which shows the importance of the subject. In this chapter, we have proposed an opinion propagation model (OPIU), where the impact of influential users was considered as a crucial factor on propagation in both signed and unsigned networks. The OPIU is based on the link analysis approach inspired by baseline propagation method (Voter). For each user, OPIU considers her neighbors and the degree of their expertise in the network and based on that propagate their opinion toward the current user. In this case, we consider the impact of influential users of the network. Furthermore, in order to analyze the performance of the proposed model, we introduced a method namely Fuzzy Majority Opinion. We found that users usually tend to improve their opinions rather than decreasing it e.g. diminish the rates. In addition, users rarely make a

lot of changes in their opinions. Furthermore, the empirical experiments with the Epinions and Etsy datasets show that our approach outperforms baseline method significantly and the fact that identifying the expertise of neighbors (influential users) have a crucial impact on opinion propagation.

The future works regarding this chapter are: 1) Considering the influence of friends with a different opinion: In our computations, we used the impact of neighbors who are in the bounded confidence. However, there might be a close friend who is not in the bounded confidence but she can effectively change the current user's opinion. Detecting such users and taking in to account their impacts can make the opinion propagation more precise. 2) Using the profile of the users as prior knowledge for their decisions: The other future work is considering the prior knowledge of the users for propagation. If a user has her own information regarding a product or even has her own experience in using a product, that can considerably affect her opinion. This prior knowledge can be achieved by reviewing and investigating the users' profiles. A user may put a note that she used a product or a picture showing a product with the user and etc.

Chapter 5

Aggregation of the Opinions

Contents

5.1 Introduction

The following chapter provides the opinion aggregation method used for presenting the consensus opinion toward the product or problem.

Today there are numerous groups of people in different categories such as the smallest (but powerful) one like family or the biggest ones like Walmart company. We believe almost every person in the world is within one of the groups. For instance, the groups of scientists, judges, family, drivers, fighters, markets, and thousand other existing groups. The groups have a lot of responsibilities, such as making food, fixing a problem, providing information, getting a decision and so on. Getting a decision regarding a problem is one

the most important responsibilities of each group. Wrong decisions have a bad influence on the future and in some cases, they can make a disaster. In contrast, good choices lead to a better life, hence the decision is important. In of In our case, we want to know how can we make a consensus opinion for a group to help them take an appropriate decision. Consensus means a general opinion shared by all the people in a group. Unlike the precise mathematical formulas used in technical analysis, we cannot easily reduce human behavior to a mathematical equation that can be plotted on a graph as a trend-line or as a series of variables that we can examine in detail, throughout history.

That said, much of the current research in social sciences is attempting to bring psychology more in line with mathematics for the precision that it gives to experimental methods. Mathematical methods are applied to behavioral science for the purpose of observing and comparing human behavior, according to a set of strict numerical criteria, the only stable benchmarks that allow comparison of behavior from person to person and from time to time. In relation to trading and investing, we can consider two very different approaches to psychology in the markets: individual psychology and group psychology. Individual psychology obviously only looks at the behaviors of the single individual trader. Attempting to draw conclusions based on the actions of the herd, mass psychology (or group psychology or crowd behavior) examines how the behavior of all investors exerts an effect on a stock price (or option price or currency value).

In online markets, a lot of online shopping centers such as Wanelo, Etsy, Fancy, Pinterest, Fab, Shoptagr, Hotels etc. exist which provides several products to the customers. In these online shopping websites, it is needed to construct preference relations of users by comparing a finite set of opinions and aggregate them to a collective one to derive a common rate (opinion) or solution [88, 89]. However, the group of users usually have the inconsistency problems due to different backgrounds and knowledge on the decision making problem faced [90–96]. It is preferable that the users reach consensus (agreement) before applying aggregation procedure. This topic has attracted the interest of a large number of researchers in this field and group decisions [97–101]. The group interaction consensus has been proved to be an effective method to reduce or eliminate inconsistency [102–106].

5.2 Problem Definition

The online shopping centers rate their products (assign a score to the products based on their quality reported by users). This score is based on the rate of other users who have bought or used the current product. Later, the other consumers will decide and get their decisions based on these rates toward the products. We called these rates as opinions which has a crucial effect and impact on users decision. These websites should aggregate

the users' opinions to a collective one to drive a common opinion and present it to their consumers. However, the group of users usually have the inconsistency problems due to different backgrounds and knowledge on the decision making problem faced. To overcome this problem, we need to reach an agreement between the users and then aggregate their opinions. One way to reach the agreement among the users of a group is to let them negotiate with each other. In our case, we propagate the opinions in the network to reach the agreement e.g. users will talk and change their opinion considering their own opinion and others ones. Furthermore, the true agreement cannot be reached if we don't consider the impact of each user. In other words, some users have a greater impact on changing the opinions of others. We called them influential users and we taking it to the account, propagated the opinions to reach the agreement.

Therefore, in the case of online shopping websites, there are a number of products and the customers of these websites. The customers want to decide regarding the products and the task is providing an appropriate rate for each product and help the customers to have better decisions. This solution can be applied to any group who need to find a convenient decision regarding the encountered problem.

5.2.1 Different network types and robustness of the method

There are several types of social networks and in the domain of business, there are three distinct ones, and an integrated social business strategy uses all three to improve communications with different groups of people [1]:

1. Public social networks like Facebook and Twitter - good for making contact with customers and prospects.

2. Social extra nets including customer communities, for deeper communication and collaboration with customers, and private business-to-business networks for communication with partners and B2B customers.

3. Employee networks for internal company communication

The focus of this study is the social networks which are in the first category. Among these networks, there are other properties as below:

- **Bipartile networks**: a bipartite network is a network whose nodes can be divided into two disjoint and independent sets U and V such that every link connects a vertex in to one in V. Vertex sets U and V are usually called the parts of the network. As an example, we can mention the dating online websites which matches the men and women.

[1] https://www.business2community.com/social-business/three-types-social-network-0606551

- **Complete networks**: a complete network is a simple undirected network in which every pair of distinct vertices is connected by a unique link. A complete di-network is a directed network in which every pair of distinct vertices is connected by a pair of unique links (one in each direction).
- **Directed Network**: a directed network is a network that is made up of a set of vertices connected by links, where the links have a direction associated with them.
- **Hyper Network**: a hyper network is a generalization of a network in which a link can join any number of vertices. Formally, a hyper network H is a pair $H = (X, E)$ where X is a set of elements called nodes or vertices, and E is a set of non-empty subsets of X called hyper links or edges.
- **Multi Network**: a multi network (in contrast to a simple network) is a network which is permitted to have multiple edges (also called parallel edges), that is, edges that have the same end nodes. Thus, two vertices may be connected by more than one edge. There are two distinct notions of multiple edges: 1) Edges without own identity: The identity of an edge is defined solely by the two nodes it connects. In this case, the term "multiple edges" means that the same edge can occur several times between these two nodes, 2) Edges with own identity: Edges are primitive entities just like nodes. When multiple edges connect two nodes, these are different edges. A multi-network is different from a hyper network, which is a network in which an edge can connect any number of nodes, not just two.
- **Random Network**: The random network is the general term to refer to probability distributions over networks. Random networks may be described simply by a probability distribution, or by a random process which generates them. The theory of random networks lies at the intersection between graph theory and probability theory. From a mathematical perspective, random networks are used to answer questions about the properties of typical graphs. Its practical applications are found in all areas in which complex networks need to be modeled - a large number of random network models are thus known, mirroring the diverse types of complex networks encountered in different areas. In a mathematical context, random network refers almost exclusively to the Erdo-Renyi random network model. In other contexts, any network model may be referred to as a random network.
- **Weighted Networks**: A weighted network is a network where the ties among nodes have weights assigned to them. A network is a system whose elements are somehow connected. The elements of a system are represented as nodes (also known as actors or vertices) and the connections among interacting elements are known as ties,

edges, arcs, or links. The nodes might be neurons, individuals, groups, organizations, airports, or even countries, whereas ties can take the form of friendship, communication, collaboration, alliance, flow, or trade, to name a few. In a number of real-world networks, not all ties in a network have the same capacity. In fact, ties are often associated with weights that differentiate them in terms of their strength, intensity, or capacity. On the one hand, the strength of social relationships in social networks is a function of their duration, emotional intensity, intimacy, and exchange of services. On the other, for non-social networks, weights often refer to the function performed by ties, e.g., the carbon flow between species in food webs, the number of synapses and gap junctions in neural networks, or the amount of traffic flowing along connections in transportation networks. By recording the strength of ties, a weighted network can be created (also known as a valued network).

- **Signed Networks**: In addition to the weighted networks, there are some networks that the weights are limited to +1 and -1. These networks can be used to illustrate good and bad relationships between humans. A positive link between two nodes denotes a positive relationship (friendship, alliance, dating) and a negative link between two nodes denotes a negative relationship (hatred, anger).

In this manuscript, we focused on the real networks and based on that devised a solution for the explained problem. Current online shopping shops are usually non-bipartile, simple, non-hyper, undirected and unsigned networks. It is worth mentioning that randomness and completeness will not affect the proposed method as the OPIU will consider the links and the number of links has no effect on the method. Considering that the most of today real world online shopping sites (networks) fall in the group of unsigned directed ones, the experiments are implemented based on two different types namely, Epinion (signed and directed) and Etsy (unsigned and directed) networks. Table 5.1 shows the impact of these networks on the proposed model and its robustness:

5.2.2 Using the bounded confidence for propagation

In order to propagate the opinions in proposed OPIU model in previous section, for each user, we considered the neighbors who are in her bounded confidence (whose opinions are close to the current user) and according to their opinions, we update her opinion. This process will continue for all of the users, until the change in the users' opinions are not sensible (the difference of new opinions and previous ones are less than ϵ). The reason for considering these neighbors is related to the convergence issues. It is proven that using the bounded confidence will converge after a finite steps [35, 107, 108]. In fact, the proposed method runs for some steps and in each step, updates the users' opinions. The problem of

Table 5.1 – Different networks and their impact on the proposed method

Network Type	**Can impact**	**Description**
Bipartile	Yes	There is two groups of users and therefore, it needs to reconsider the propagation based on different groups
Complete	No	It is already considered and it does not influence the method
Directed	No	The directions of the links are considered in the method
Hyper	Yes	As the link can join any number of users, the computation of new opinions needs to be redefined
Multi	Yes	There are several links and thus, it needs to reconsider the connection between users
Random	No	It is already considered and it does not influence the method
Weighted	Yes	We considered the +1 and -1 weights but not the other values. However, the method can be developed to considet
Singed	No	The proposed mechanism considers the signed network so, it does not influence the method

the process is the number of these steps and the convergence of the users' opinions. Hence, it order to satisfy the convergence of the opinions in proposed method, we use the neighbors who are in the bounded confidence. In this way, the method and its updating will finish after finite steps. In addition, as a future work of this manuscript we propose finding the impact of those neighbors who are not in bounded confidence but are important to the current user. For instance, there may be a close friend who has a great influence on the current user's opinion, but his opinion is far from her. This situation cannot happen a lot as close friends usually think like each other who have same (or close) opinions. However, it is possible that for a product, they have completely different opinions. In order to find these rare neighbors, we suggest using the Jaccard distance. In other words, we consider two users as close friends when the portion of their similar neighbors is high. Later, the opinion of this close friend will be added to the bounded confidence of the current user if her opinion is far from the user. There could be other evaluation and analysis such as the threshold for begin close friend, the impact weight of close friend's opinion and etc. that can be considered for opinion propagation which we will discover in the future work of this manuscript.

5.2.3 The false positive values

We tried to find the influential users in the network based on optimism and pessimism scores. One problem regarding the found influential users is false positive values. In other words, is there any way to find the false positive influential users or not. One way to find the false positive values is using a data-set with ground truth. In real world networks, it is hard and costly to have such network as the online social networks managers may refuse to diffuse their data. In addition, the two implemented networks does not have this information. In order to solve this problem, we introduced a method (namely credibility 3.6) to measure of being the influential user. The credibility has a direct relation with being influential and hence, it can be used as our ground truth. Hence, we compute the users credibility and compare the most credible users with the found influential users by the method. To prevent the false positive values, we can define a threshold for the credibility of the influential users and put aside the found influential users whose credibility is less than this threshold. The false positive influential users will influence the proposed method as they have a direct impact on propagation. Hence, we verified the credibility of each of the found influential users by proposed OPIU. We found that the introduced influential users buy OPIU have high credibility in previous section and also the proposed method is performing better than other methods 4.4.3. However, this verification should be done right after the first phase (finding the influential users) to put aside the false positive influential user.

5.3 Consensus Formation Method

In order to solve the above problem, we defined three distinct steps. First, we find the influential users and then propagate their opinions to reach the agreement and at the end aggregate the opinions. The first two steps, i.e. finding the influential users and propagating the opinion are presented in chapters Chapters 3 and 4. Here, we apply the aggregation methods. To do so, we introduce three methods namely, Lehner-Wanger, ordered weighted averaging and Fuzzy majority. When users of a company or any group make judgments or decisions, their members interact with each other: they exchange relevant information, put forward arguments and deliberate the reasons for a particular position. Then, they will act based on the decided opinion or pass the group decision to the manager and she will act based on that. In an online shopping center, the users rated a product form a group and their opinions toward the product will be aggregated and presented as the final opinion. This opinion is the overall rate of users which helps the other consumer who has an inquiry about the current product. Thomas et al. [109] compare three different models: an independent model, where the group reliability is just the probability that each

group member has solved the problem, a rational model, where the group makes a correct judgment as soon as a single member is right, and a consensus model, which assumes the groups' inclination toward uniformity. Using an arithmetically simple, but conceptually tricky mathematical problem, the authors find that the consensus model describes the outcomes better than the other two.

We implied that to provide a consensus opinion within a group, we need the users to negotiate and update their opinion based on their relations. This update is done by the proposed opinion propagation method. The reason is that mutual respect among the group members should prompt every group member to revise her initial opinion [110]. This respect can be epistemically motivated (e.g., by realizing that the other group members are no less competent than oneself), but also reflect degrees of care or relations of social power, dependent on whether there is a matter of fact to the subject of disagreement. Conditional on such mutual respect, blending one's opinions with the opinions of the other group members seems to be a requirement of rationality. This will assure the consistency of the group. One justification for aggregation is consistency, since refusing to aggregate is equivalent to assigning everyone else a weight of zero [111]. In other words, refusing to blend one's opinions would amount to unjustified dogmatism [112].

Algorithm 2 is the pseudo code of the whole process of consensus opinion model. It has three main processes. First, we find the influential user's scores (ranks). Then, we propagate the opinions of users considering the impact and rank of users and at the end, we aggregate the opinions to provide the consensus opinion of each product or group.

Algorithm 2 Consensus Opinion Model

Finding influential users in the network
  Personality definition
  Rank the users using POPRank
  Using Credibility to evaluate the ranking
Propagating the opinion
for each user i $[U_i]$ **do**
    Consider all the $neighbors_i$ and their ranks
    Set the neighbors whose opinions are in confidence bound U_i
    Updating the U_i opinion based on these neighbors
      $p_i^{t+1} = p_i^t + \frac{1}{|N_i^t|} \sum \mu \frac{SP_j}{SP_i + SP_j} [p_j^t - p_i^t]$
end for
  Using Fuzzy Majority Opinion to analyze the opinion propagation
Aggregating the opinions of the users for each product
for each product i $[product_i]$ **do**
    Consider all the users who has opinion toward $product_i$
    Aggregate the users Opinion of current product
end for
Using influential users opinion to evaluate the provided consensus opinion

5.3.1 Lehner-Wanger Aggregating method

Among models of opinion aggregation, the Lehrer-Wagner model is most prominent [112]. The model tries to estimate the quantity x, from the individual estimates v_i of every group member i. The x is normally thought of as objective and independent of the group members' cognitive states. The quantity x could be the opinions of the users toward a product. Their central idea consists in ascribing the agents' opinions about each others expertise, or in other words, mutual assignments of respect. Then, the w_{ij} describe the proportion to which j's opinion on the subject matter in question affects i's revised opinion. These mutual respect assignments are used to revise the original estimates of the quantity in question, and codified in an NxN matrix W (where N denotes the number of agents in the group):

$$W = \begin{pmatrix} w_{11} & w_{12} & ... & w_{1N} \\ w_{21} & w_{22} & ... & w_{2N} \\ & & ... & \\ w_{N1} & w_{N2} & ... & w_{NN} \end{pmatrix} \tag{5.1}$$

The values in each row are non-negative and normalized so as to sum to 1: $\sum_{j=1}^{N} w_{ij} = 1$. Thus, the w_{ij} represent relative weights which the agents ascribe to themselves and to others when it comes to estimating the unknown value x. Then, W is multiplied with a vector v that contains the agents' individual estimates of x, obtaining a novel updated value for v:

$$W.v = \begin{pmatrix} w_{11}v_1 & w_{12}v_2 & ... & w_{1N}v_N \\ w_{21}v_1 & w_{22}v_2 & ... & w_{2N}v_N \\ & & ... & \\ w_{N1}v_1 & w_{N2}v_2 & ... & w_{NN}v_N \end{pmatrix} \tag{5.2}$$

In general, this procedure will not directly lead to consensus, since the entries of $W.v$ differ: $(Wv)_i \neq (Wv)_j$. However, later they showed that under very weak constraints, the sequence $(W^k)_k \in N$ converges to a matrix W^∞ where all rows are identical, that is, where all agents agree on their relative weights. That is, when the procedure of averaging is repeated, the agents will finally achieve a consensus and not only agree on the factual subject matter but also on the differential weight that each group member should obtain. The method is similar to ordered average weighting (see next section).

5.3.2 Ordered Weighted Average (OWA)

The OWA is introduced in [113] to aggregate the group decision. The author introduced a type of operator for aggregation called an ordered weighted aggregation (OWA) operator and investigated the properties of this operator. The OWA's performance is found to be

between those obtained using the AND operator, which requires all criteria to be satisfied, and the OR operator, which requires at least one criteria to be satisfied.

An OWA operator of dimension n is a mapping $F : R^n \to R^n$ that has an associated weighting vector W of dimension n having the properties

$$\sum_{j=1}^{n} w_j = 1, w_j \in [0, 1] \tag{5.3}$$

and such that

$$F(a_1, ..., a_n) = \sum_{j=1}^{n} w_j b_j \tag{5.4}$$

where b_j is the jth largest of the a_j. Central to this operator is the reordering of the arguments, based upon their values. That is, the weights rather than being associated with a specific argument, as in the case of the usual weighted average, are associated with a particular position in the ordering. We note this reordering introduces a non-linearity into an otherwise linear process. If B is a vector corresponding to the ordered arguments, we shall call this the ordered argument vector, and W^T is the transpose of the weighing vector then we can express the OWA aggregation in vector notation as

$$F_w(a_1, ..., a_n) = W^T B \tag{5.5}$$

The OWA operator provides a very rich family of aggregation operators parametrized by the weighting vector [114]. Among the operators included in this family are the average, max, min, and median of the variables. The operator can be defined appropriately to the situation. For instance, consider there is a war and some sources regarding the number of enemy troops approaching. The officer should know how many enemy troops are coming to stand against them, so he should combine the sources of the information. As long as underestimating the enemies could be costly, it is better to use max operator.

5.3.3 Fuzzy Majority Aggregation

On one hand, as described above, the Lehner-Wanger will not always proceed to a consensus opinion model and it needs a promising constraint. On the other hand, the aggregation performed by the OWA operator depends upon the form of the weighting vector W which can be simply the average of opinions ($w_j = (\frac{1}{n})$). The process of determining the weighting vector is important. One method for obtaining the weighting vector is to associate the OWA aggregation with a linguistic quantifier represented as a fuzzy subset Q on the unit interval. Here, we use the described fuzzy majority in 4.3.6 which tries to find the majority opinion

among the users as their aggregated one. Under this explication, the majority opinion is no longer represented as a value, but as a fuzzy subset. This will provide in addition to a value for the majority opinion a sign of the strength of that value as a delegate of the majority opinion. In this approach, the weights associated with an aggregation of degree are obtained as

$$w_j = Q(\frac{j}{n}) - Q(\frac{j-1}{n}) \qquad for \qquad j = 1, ..., n \tag{5.6}$$

Here, for each product, we consider the users who have an opinion toward it (who rated the current product). These users form a group. Then, we use the fuzzy majority opinion to present the aggregated opinion of this group. In order to evaluate the model, we compare the aggregated opinion with expert satisfaction [115, 116]. Using our method, each group generate an aggregated score toward the product. In order to assess the performance of this score, we compare the similarity between the aggregated opinion with the opinion of the most prominent expert in that group (the expert should be ranked high in POPRank). Ultimately, if the provided consensus opinion is changed in a way that gets near to the expert opinion, that means that the aggregation method satisfies the expert.

5.4 Experiments

We used two methods for aggregating namely Ordered Weighted Average and Fuzzy Majority Aggregation. In addition to the ranking in signed networks, we performed the influential user detection (POPRank) in unsigned network Etsy. Figure 5.2 illustrates the credibility of ranking method for Etsy data-set (in addition to the experiments we showed in chapter 3). We observed that for Etsy data-set the percentage of found influential users in POPRank have more credibility rather than other methods. This again implies the power of users' personality in finding the influential users. It is worth mentioning that the Spearman correlation of more credible nodes Prestige, Pagerank and POPRank are 3.68%, 10.51%, and 13.74% respectively. This shows that the users found by POPRank algorithm have more credibility in comparison with others and as a consequent confirms the performance of POPRank in finding the influential users.

Figure 5.1 shows the PMF of the Etsy data-set in addition to the presented PMF of Epinions in 3.6.1. Again, we can see that the slope of the PMF is steep at the beginning, which means that the first ranked users have a higher level of credibility and they are different from other users (this difference is meaningful). Then, the slope becomes smooth for lower ranked users, which means that their ranks are getting almost same as each other. Please note that here (in the Etsy data-set), the network is sparser in comparison with the Epinions data-set. Hence, there is more difference between the upper (first 100 users) found

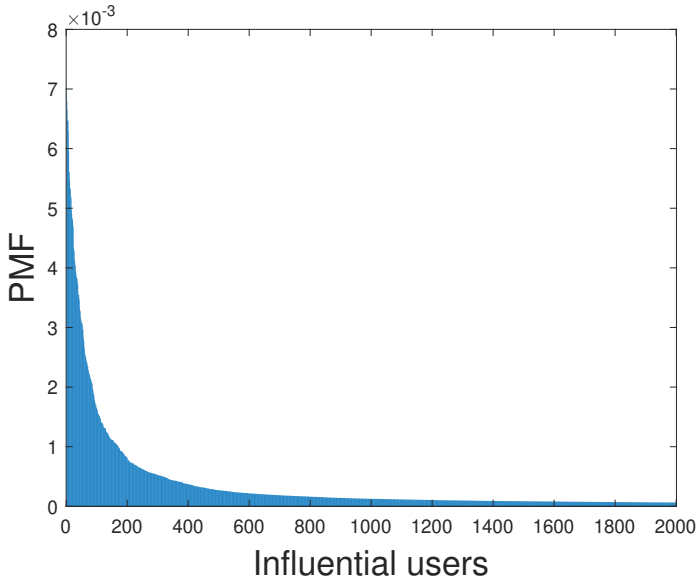


Figure 5.1 – The probability mass function of credibility in Etsy data-set

influential and the rest ones in comparison with Epinions.

As we discussed before, we used expert satisfaction to evaluate the aggregation methods. In order to verify the performance of the aggregation methods, for each product, we compared the similarity between the aggregated opinion with the opinion of the two most prominent influential users in each group. To do so, we performed two aggregations OWA and FMO in both Etsy and Epinions datasets. For both datasets, we first chose the top 10% of influential users and then determined the products that they rated. Among these products, we chose one hundred products to analyzed the aggregation performance. Each product has a number of users who have the opinion towards them. Hence, for each product, we found the two best influential users ($IU1$ and $IU2$) and compared their opinion with the aggregated one. Figures 5.3 and 5.4 show the similarity of different aggregation methods with the two chose influential users for Epinions and Etsy datasets respectively.

These figures show that for most of the products, the similarity value of aggregated opinion and the opinion of the influential user who rated the current product is high. In other words, the aggregated opinion is near to the opinion of the influential users. This confirms the performance of aggregated methods. However, the performance of OWA and FMO are different. Table 5.2 illustrates the MSE of FMO and OWA methods with two influential users ($IU1$ and $IU2$) for both data-sets.

We observed that in general, the performance of FMO is better than OWA however, it does not mean that FMO always makes better aggregation value. For example, the MSE of the OWA method for $IU1$ in Etsy is slightly better than FMO which makes the OWA better, however, from the results we can say FMO has better performance rather

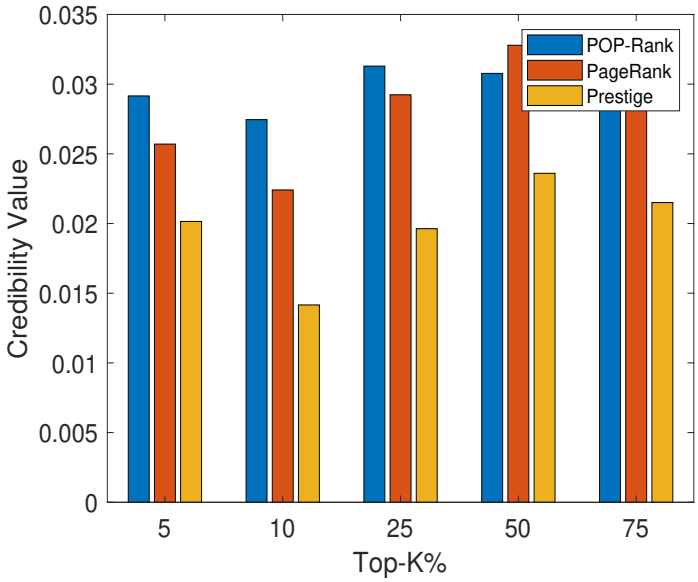
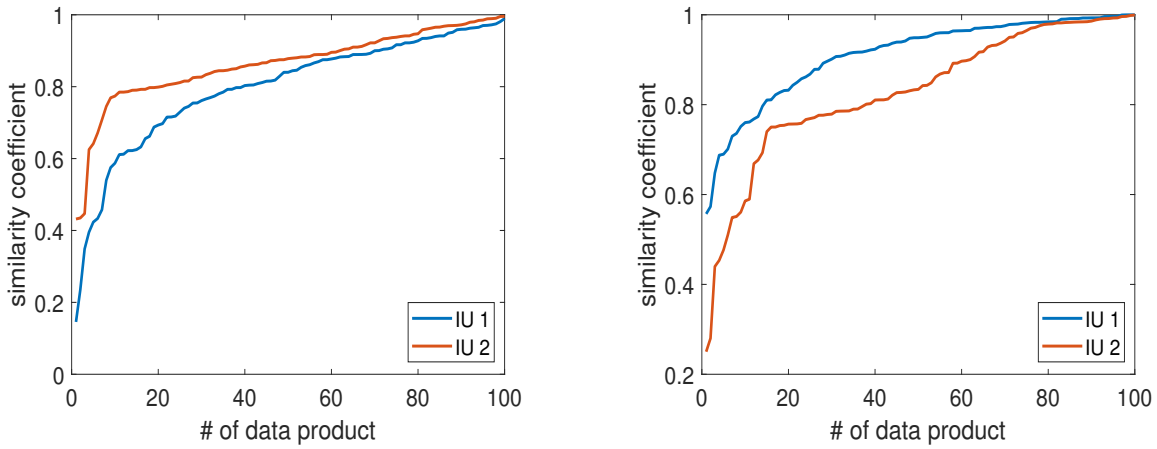


Figure 5.2 – Normalized Credibility of each ranking algorithm regarding different percentages of found influential users in Etsy. X axis represents different percentages of top found users and Y axis shows the normalized credibility value of found users



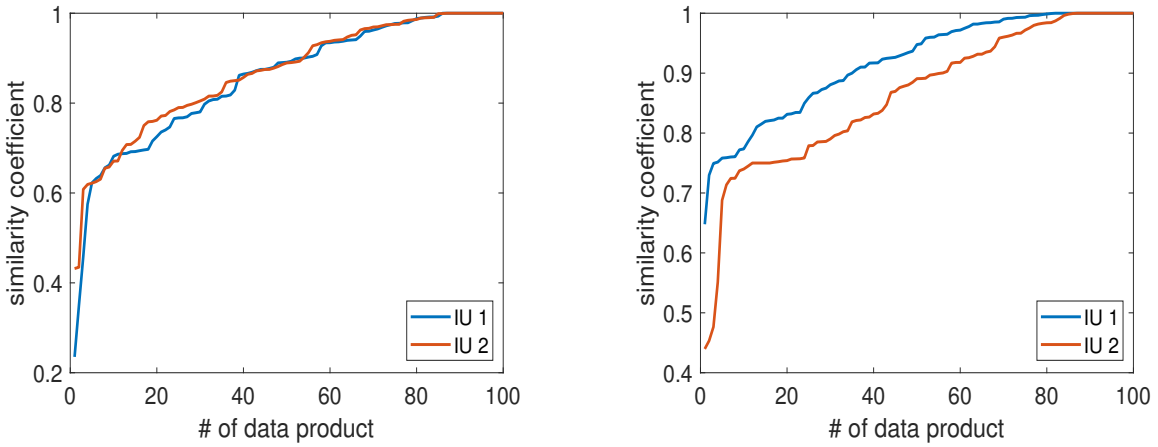
(a) OWA aggregation

(b) FMO Aggregation

Figure 5.3 – Aggregation in Epinions dataset

Table 5.2 – MSE of different Aggregation methods

Dataset	**OWA**		**FMO**	
	IU1	IU2	IU1	IU2
Epinions	0.89	1.09	0.79	0.64
Etsy	0.96	0.62	0.99	0.59



(a) OWA aggregation (b) FMO aggregation

Figure 5.4 – Aggregation in Etsy dataset

than OWA. As long as we used expert satisfaction to evaluate the aggregation, it is worth mentioning that this model is not able to detect when the influential user is wrong or inconsistent. This explains the low similarity of influential users and aggregated methods in some products.

5.5 Conclusion and Future Work

We proposed a novel consensus opinion approach that has been specially designed to model opinion formation in social groups such as online shopping centers. Assuming the social power and different impact of each individual user in the network, we proposed a consensus model. To do so, we first identified the influential users using personality of them in the network by POPRank methodology. Then, in order to maintain the consistency, the users need to negotiate their opinion in which we proposed a new method of opinion propagation (OPIU) considering the impact of influential users. Ultimately, we aggregated the opinions by OWA and Fuzzy methods to present the consensus opinion. We performed several experiments in signed and unsigned networks namely Epinions and Etsy datasets. It is worth mentioning that the proposed POPRank algorithm demonstrated that the personality features can effectively improve the ranking and consequently, finding the influential users in different social networks. Furthermore, in OPIU we observed that most of the users like to increase their opinions rather than decreasing them during the propagation period. Finally, in general, the fuzzy aggregation method performed better than OWA which proves that FMO suits best to our context. On the other hand, we should point out that we can not evaluate the performance of aggregation methods if the opinion of influential users is wrong. As a potential future work, we investigate the incorporation of a mechanism to evaluate

the performance of aggregation regardless of the opinion of the expert satisfaction.

Chapter 6

Conclusion and Future Work

Contents

Today the online shopping websites got a lot of attention from the people. The tendency of customers to use these websites for their shopping is increasing day by day due to their reliable possibilities such as time-saving, better prices, fast comparison and etc. On the other hand, one of the most important issues regarding the products of these websites is that how can the online shopping portals provide a valuable information for their customers and help them decide about their shopping. In other words, can they present an appropriate rate for their products so that the customers obtain a proper knowledge regarding the quality of them? This is very important as it helps the customers to find the right product they are searching for which ends with their satisfaction and more shopping from the current website (double satisfaction). The online shopping websites usually ask their customers to rate the product after they buy and use it. Then, these opinions will be aggregated and presented as the final rate of the products for other customers. However, a suitable aggregation cannot be achieved by simply averaging the opinions. There are inconsistency and incomplete problems that need to reach an agreement (consensus opinion) before aggregating the opinions. To do so, the opinions are propagating through the users of the network. In addition, the propagation methods consider the neighbors of the user to update her opinion. However, some of these neighbors (such as the influential users) have more impact on propagation. Hence, identifying the influential users is also needed before propagating the opinions. The current thesis provides a comprehensive solution to the problem of opinion

consensus in online shopping groups.

6.1 Summary of the Contributions

We divided the above problem into three sub-problems. The aggregation which needs the propagation and the propagation which needs to detect the experts in the network. As the solution and our contributions, the devised model for the opinion consensus is as follows (three phases):

1) **Finding the influential users in the network**:
One way to find the trustful information is relying on the information provided by experts or influential users. In a network, the influential users are those who have more effect on others and their information can be trusted more than usual ones. These special users can be found by analyzing their profile, background, and links they have in the network. In real networks, the most available information regarding the users is their links and the connection to each other. Therefore, in the current thesis, we used link analysis to find them. To do so, we used the users' personality to rank them in a given network and consequently detect the influential users. The users' personality consists of two features namely Optimism and Pessimism. Finally, using these features, the POPRank algorithm was proposed to rank the users and find the influential ones.

2) **Propagating the opinion in the network**:
People take a lot of decisions in their lives. The customers of an online shopping website usually take the decision about its products based on the current information they have, the opinion of their friends (neighbors in the network) and the influential users of that website. Therefore, in this way, the users' opinions will propagate in the network. This propagation will help the network managers who want to aggregate their users' opinions as it maintains the consistency. In addition, the managers can predict and find the opinions of their new customers based on the links they have. There are a lot of studies paid attention to the dynamics of opinion and in this thesis, we proposed an opinion propagation model namely OPIU. The OPIU considers the user's neighbors and the degree of their expertise in the network and based on that propagate their opinion toward the current user. As a result, the OPIU presented a more precise opinion propagation.

3) **Aggregating the opinions and providing the consensus opinion**:
The online shopping websites, as well as the social groups, are increasing continuously. Usually, the customers of online shopping websites who bought a product, put their opinions as the rate of that product on these websites. Then, using these opinions, the websites present an aggregated score of that product and present it to the other customers in order to help them find the best product they are looking. This presents the consensus opinion for

the products of the online shopping portals. In order to provide a precise consensus opinion regarding a product, the first and second phases were employed to find the influential users and propagate the opinion in the network. Ultimately, the Fuzzy majority opinion and OWA aggregation methods were performed to reach the aim. The Fuzzy aggregation method outperformed the OWA in our context.

For each phase, several experiments on two real datasets (Epinions and Etsy) were performed and finally, we were able to provide the consensus opinion model. This solution is essential for online shopping managers and helping the customers of them in shopping which can be used in any group such as companies who needs to reach an agreement or consensus opinion among their employees and users, however, in our case, we aimed the online shopping websites.

6.2 Conclusion

In this study, we proposed a novel consensus opinion approach that has been specially designed to model opinion formation in social groups. The aim of this study is presenting a mechanism for online shopping centers (and any other social groups) to provide a comprehensive information for their users regarding each product of their websites. Assuming the social power and different impact of each individual user in the network, we proposed a consensus model. To do so, we divided the problem into three phases. First, we identified the influential users using link analysis and the personality of the users of the network. In other words, we used two features from social science namely Optimism and Pessimism to add the personality of each user in the ranking algorithms. We applied the user's personality to PageRank algorithm and created a new ranking POPRank for signed and unsigned networks. Second, in order to maintain the consistency, the users need to negotiate their opinion in which we proposed a new method of propagation OPIU considering the impact of influential users as a crucial factor on propagation in different networks. The OPIU is based on the link analysis approach inspired by baseline Voter propagation method. For each user, OPIU considers her neighbors and the degree of their expertise in the network and based on that propagate their opinion toward the current user. In this case, we consider the impact of influential users of the network. In the end, we aggregated the opinion by OWA and fuzzy methods to present the consensus opinion. We performed several experiments in signed (Epinions) and unsigned (Etsy) networks. The performance of each phase is verified with different methods. We used the concept of credibility to verify the performance of our POPrank ranking. Furthermore, in order to analyze the performance of the proposed OPIU model, we introduced a method namely Fuzzy Majority Opinion. Ultimately, the expert satisfaction was utilized to confirm the aggregation performance. As

some results of our experiments, we can mention that the proposed POPRank algorithm demonstrated that the personality features can effectively improve the ranking and consequently, finding the influential users in different social networks. Furthermore, in OPIU we observed that most of the users like to increase their opinions rather than decreasing them during the propagation period. In addition, users rarely make a lot of changes in their opinions and most of the opinion changes are smooth. Finally, in general, the fuzzy method performed better than OWA in the aggregation phase which proves that FMO suit best to our context. We studied the need of customers in online shopping websites which are getting a lot of attention during this era. Eventually, we presented a model to help both customers and the owners and managers of these online social groups. In other words, with the presented study, the online shopping websites can have more benefit by providing a better information regarding their products and help their customers to decide more precisely and choose the right product they are searching for.

6.3 Future Research

We conclude our thesis by mentioning some of the future research directions:

Using profile information We identified the influential users using link analysis. As a future guideline, we plan to use the profile of the users in addition to their connections as the users' prior knowledge. In other words, the link analysis can identify the expert users but we plan to investigate the effect of each user profile on the accuracy of expert detection.

Communities role in influential users detection Different communities have different strategies. For instance, a user who is a member in accommodation community has more weight for being an influential user in hotel.com. Taking the users communities into the account can add an efficient value in finding the influential users.

Friends with opposite opinion in propagation In the propagation process, we updated the opinion of users based on the opinion of their neighbors who are in the bounded confidence. The close friends have a crucial impact on opinion propagation however, they are hard to detect in link analysis. Finding a mechanism (such as the number of mutual friends between two users) to find these friends and considering their influence in propagation is another future work.

Evaluating the aggregation In our experiments, we used the expert satisfaction concept to evaluate the aggregation methods. However, if the expert's opinions are wrong the evaluation can be ended in an incorrect realization. Hence, one of the future guidelines of this study is to investigate a mechanism to evaluate the performance of the aggregation regardless of the opinion of the experts (such as having the ground truth).

Chapter 7

Other research as a part of my studies: A Framework to Detect Users' Life Events in Online Social Networks

Contents

7.1 Introduction

Online Social Networks (OSNs) plays an essential role in today's human life that almost all people use them as a daily basis. The OSNs are made of users, groups and their connections and these individual users create tremendous amount of data in the form of texts, audio,

videos and photos. Access and investigate these data reveal unique information regarding individual users and their preferences and define paradigms for social research analysis which result in many useful applications e.g. business and marketing, healthcare, education, social and community engagement application and life stories. Meanwhile, mining and analyzing such contents could help identifying one's life significant times and events such as Facebook Look Back and Google Awesome generates short video clips for users to summarize and visualize their time-lines. Furthermore, detecting such moments could be used for recommendation systems to suggest new items based on the new status of the user. To this end, events and detection of them got a lot of attention from social networks.

Shared information by social media users could be used for different purposes. One of the popular research subject regarding social media analysis is event detection. Event detection is related with finding special events and incidents from streams of social media updates. There is a growing body of literature that recognizes the importance of identifying such events based on the social media [117–119]. In human life, there are two categories of events: *1) personal life events 2) non-personal events*. Personal life events are the incidents such as birthdays, work and education changes, relationships and etc. which are directly related to the life of the users. As talking and sharing about personal life events is a frequent activity in social media, major social networks even provide specific attributes in users' profile for this sort of activity such as the snapshot of Figure 7.1 showing the mentioned attributes in Facebook. On the other hand, non-personal ones are those on a large and global scale which have international significance and take place in society e.g. earthquake, traffic, attack and etc. Apart from the clearly mentioned events by users in their profile, there are also bunch of events that is expressed in different type of activities which identifying those events can be very interesting challenge. In this domain, most of the researchers paid attention to non-personal events detection. Therefore, there are a few studies focused on personal life events. For instance B. Di Eugenio et al. in [120] tried to identify two personal life events from Tweeter text stream. In addition, most of the previous studies used only the text mining techniques to find the events. According to this, S. Choudhury et al. verified the text flew of users in order to find the personal events [121], J. Li et al. did similar process in Tweeter [122] and P. Cavalin et al. analyzed the Tweeter stream to explore the personal events [123]. Their methods can find events if the users post a text related to her new life incident. Nevertheless, they still have lack in detecting the events which are not mentioned explicitly in the user's profile.

This chapter proposes a new hybrid framework to detect a majority of the possible personal life events based on available social information. In order to enhance prior detection methods, not only the text streams but also profile attributes and activities are considered. To this end, we proposed a method consisting of three layers, Input, Data processing and

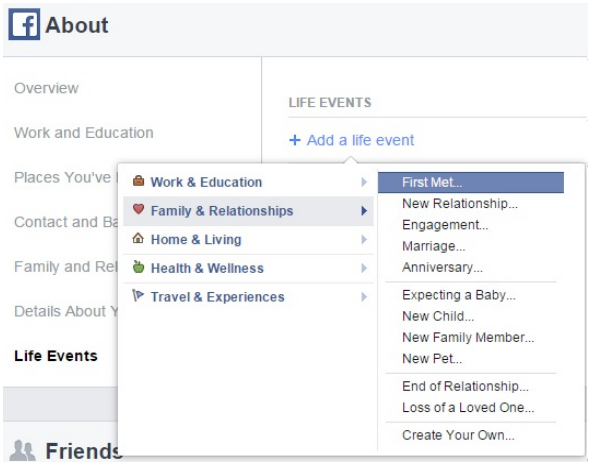


Figure 7.1 – A snapshot of available profile attributes for life events in Facebook.

Visualization (output) layers as shown briefly in Figure 7.2. The first layer, i.e. input, is connected to users' source of data which includes the profile and activity information. The framework first gather activities (text and image posts), reactions (comments, likes, share) as well as the profile attributes (e.g. age, gender, friend list, city and job etc.). Then different modules inside the data processing layer will identify the events (see section 7.3). At the end, the detected events will be presented in the last layer of the framework and will be stored as well in the corresponding elements to be used in the event prediction phase.

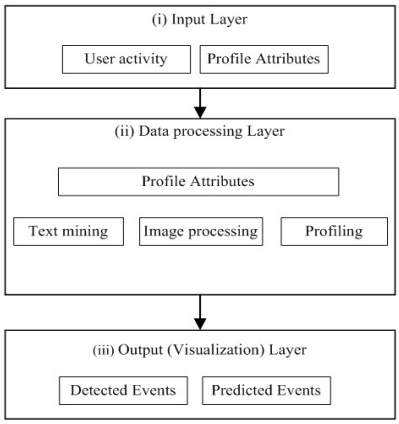


Figure 7.2 – The overall framework of Event detection method including modules

It is worth nothing that ML methods is well-known to produce solutions to deal with ambiguous and noisy texts. However, text mining and ML classifier is used to find the events by text stream extracted from the user. Moreover, image processing is used to find events from the images the user posts. In addition, profiling will find events from the user

profile attributes. This can be done by analyzing the recent added friends to the user's friend list and verifying these friends similarities. This section discusses event detection and the previous approaches used to detect them from social media streams. The organization of this chapter proceeds as follows: Section 7.2 summarizes the current methods used for event detection. The proposed hybrid methodology is presented in section 7.3. Section 7.4 discuss about challenges of predicting events. Section 7.5 consludes this study and finally section 7.6 present some ideas as future directions of this chapter.

7.2 Background and State-of-the-Art

This section attempts to provides a comprehensive report of the current methods for event detection based on the different techniques including text mining, image processing and other available methods. To start, table 7.1 provides a comprehensive summary of the relevant techniques on event detection from the literature. Next we present details about the main mentioned items in the table.

7.2.1 Text mining studies

The main and popular method of event detection is in text mining domain where there are several approaches to analyze the text to identify an event. As the most important ones we can indicate LDA (Latent Dirichlet allocation), TDT (Topic Detection and Tracking) and NER (Named-Entity Recognition). LDA [124] is a generative topic model that finds the similarity of data by sets of observations. It assumes that each document is a mixture of topics and each word could be assigned to one of the document's topics. According to [125], LDA models can detect the topic of happened stories. However, their performance will decrease when the task is identifying in noisy events data.

NER is an information extraction method which classify the texts elements into the pre-defined categories such as the names of persons, organizations, locations and etc. The task consists of two main parts i.e. breaking the text to names and classifying them to the defined groups.

TDT [126] is an automatic technique which discover the topical structure of streaming text by breaking down it to smaller pieces. The event detection is a sub-task of TDT [126].

7.2.2 Image processing studies

Image processing methods can be used to recognize the image objects and relate them to events. Image mining uses three distinguishable types of feature vectors for images description in order to reach high accuracy. These feature vectors are Histogram of Oriented Gradients (HOG) descriptors for object detection, Grey-Level Co-occurrence Matrix

Table 7.1 – Summery of the current techniques on Event Detection methods

Methods and techniques		References	Event category	Details
Text Mining	LDA	J. Li et al (2014) [122]	Personal events (job, wedding, birthday, etc.)	Their method used LDA-Clustering and Human-identification. They first filter the noisy tweets and then cluster them based on LDA.
		Z. Tan et al (2014) [127]	Non-personal event (global, local)	They designed a multilayer LDA to cluster the events. The definition of global and local events is described by means of topic community matrix. Meanwhile, the interests of different communities were obtained by a multilayer event detection method. Then the events categories were formulated and in the last step they presented the identified events.
		K. Sathiyamurthy et al (2014) [128]	Non-personal events	They used an unsupervised learning, LDA clustering and summarization. Their method has two parts: detection and summarization. First they used clustering to detect the events then summarized the data that describe the event properly.
	NER	P. Khare et al (2014) [129]	Non-personal events (breaking news)	In their method they first gathered the text data from tweeter. Then using NER they have clustered them to detect the events.
	TDT	CC. Aggarwal et al (2012) [130]	Non-personal event (global)	Their method clusters and detects the events in social streams. In order to have a high accuracy in event detection text mining task, they used supervised, unsupervised and clustering techniques.
	Bngram	S. Choudhury et al (2014) [121]	Personal events (Marriage, graduation, new job, new born, surgery)	Their model is based on Unigram method. They compared four models (four modified Unigram) and they have shown that a hybrid model based on Unigram outperformed among others.
	Hybrid	B. Di Eugenio et al (2013) [120]	Personal events (marriage, birth of a child, graduation, losing or getting a job)	They used and compared different text mining methods to detect and classify the events (Unigram, Naive Bayes, Decision table, SVM, CNB (Complement Naive Bayes)). At the end they have shown that the hybrid method using SVM and CNB has the best performance in their application.
Image processing	HOG, GLCM	S M. Alqhtani et al (2015) [131]	Non-personal event	In order to improve the mining performance of their method they used fusing, visual and textual information. For the text mining part they used TF-IDF and for the image processing the HOG and GLCM is used. At the end they used SVM for classifying the detected events.
	Meta Data	M. Zaharieva et al (2013) [132]	Non-personal event	The only data that is used in their method is the meta data of images. Later, this data were verified by text filtering and clustering methods for the task.
		G. Petkos et al (2014) [133]	Non-personal events (e.g. concert, sport, celebration, protests)	They used supervised clustering technique to retrieve events from meta data of images. Their method first cluster the social media data and then recover those events that meet some criteria (location, type, entities).
Other	Wavelet signals	J. Weng et al (2011) [134]	Non-personal event (protest)	For each word a signal is created using wavelet analysis and frequency of the words. Then, they filtered away the obvious used words and as the final task they clustered the remaining words to form events with the modularity based graph partitioning procedure.
	Segmentation	Q. Zhao et al (2007) [135]	Non-personal events (any communicating topic between actors group)	They exploited clustering and temporal segmentation method. First the words are compared with a database. Then similarity is calculated to specify the class of new text.
		Q. Zhao et al (2007) [136]	Non-personal events (e.g. hurricane)	The events are detected by combining text based clustering. The temporal segmentation and information flow based graph cuts is used for detection task.
	Graph	H. Sayyadi et al (2009) [137]	Non-personal events (e.g. news, election, movies)	First a key graph and community detection is used to extract event keywords and then they clustered them.
	Hierarchical	G. Ifrim et al (2014) [138]	Non-personal events (e.g. news, election, movies)	Their method filters the links and unrelated words form the text to remove the noisy tweets and then hierarchical clustering is used to cluster the remaining tweets to the relevant events.
	Similarity	H. Becker et al (2010) [139]	Non-personal event (e.g. festival, concert, street art)	They proposed a method based on clustering and similarity metric learning approaches for identifying the events in social media document.
	Other clustering	H. Becker et al (2009) [140]	Non-personal event (concert, party)	They used weighted cluster ensemble algorithm in which applies the multiple features (title, description, tags, location and time). In general, they used clustering technique which has validated weights.
		R. Li et al (2012) [141]	Non-personal event (e.g. crimes, earthquake, accident)	They developed an efficient CDE (crime and disaster related events) focused crawler, and explored valuable features from Twitter to classify and rate tweets.
		T. sakaki et al (2010) [142]	Non-personal real time Events (earthquake)	The semantic analysis is used to increase the filtering accuracy. Also Kalman and particle filtering is employed to estimate the locations of events.
		P. Cavalin et al (2014) [123]	Personal events (Marriage, graduation, travel, birthday, birth, death)	They first applied a method with two layers of filtering to find the related words. Then they used ML classification for detection.
		M. Walther et al (2013) [143]	Non-personal events (house fires, on-going baseball games, bomb threats, traffic jams, broadway premiers, conferences in an area)	They employed ML clustering and Baysian model to identify geo-spatial and real world events from tweets. They explored the clusters of tweets that are temporally and spatially close to each other to attain if it can describe a real world event.
		L. Jalali et al (2014) [144]	Personal events (e.g. dinning, shopping, meeting)	Their method has three parts: observation, event stream and situation. First a simple analysis performed on data stream and then they classified the analyzed data using Naive Bayes and random forest classifier.

(GLCM) for texture description, and color histogram [145, 146]. HOG is an object detection method base on the occurrences of gradient orientation of the images. GLCM is based on image distribution and is used to measure the texture of surfaces. Color histogram is defined as the representation of distribution of colors of an image. In addition, the data models such as support vector machine (SVM) could be used. SVM is a prominent classification method which analyzes input data and finally recognize patterns that are used in classification. Other aspect is analyzing the text and tags of the images (not the image itself). A SVM model is the representation of the example elements as points in a defined space layout that is mapped so that the examples of the separate categories are actually divided by a distinctive gap which is wide enough.

7.3 Proposed Methodology to Identify the Events

Personal life events are categorized based on users life incidents ranging from desirable to undesirable events, such as births or promotions and deaths or accidents. Based on our studies from Facebook, Google+, Twitter, and Stanford studies, personal events is mostly categorized to five different groups i.e.: work (job, school and university), family (engagement, having child and divorce), home (changing city), health (disease and wellness issues) and travel (as shown in Figure 7.1). Specific features should be considered to analyze the users in a social media. These features can be classified into two categories: 1) activity 2) profile attributes. Activity features are based on user's activity (posting text, image, following, share, or react to a post, etc.) while profile attributes includes general information on subscribers' profile (city, job, friend list, interests list, etc.). In our proposed method we use both activity and profile attributes to find the events. Moreover, for detection it is needed to consider a specific period of time such as day, week, month or etc. Using above features we will detect the users events in the chosen period. Users events can be revealed in three type of users behavior as follow:
i) changing their profile attributes e.g. hometown when they change the city
ii) posting e.g. text or image about new incident
iii) a group of new behavioral changes, e.g. adding group of new friends with similar profile characteristic (from a specific city or job).

According to this, we consider three phases in data processing layer. In this section, we explain each module in the proposed method in detail. Figure 7.3 illustrates our overall platform to find personal events.

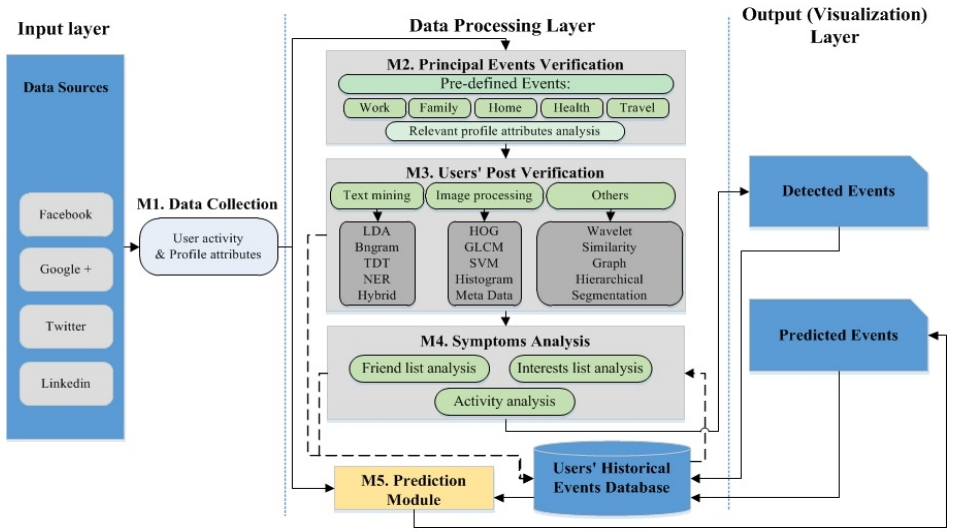


Figure 7.3 – The block diagram of the event detection framework including the three layers and different modules.

7.3.1 Input layer

Data sources: This layer includes different sources of data, e.g. social media, which the information of targeted users will be gathered for the data collection module. Also, data pre-processing such as filtering unrelated information will be done here. Overall, in this module the text, images and the log of each user will be prepared from the social media.

M1. Data collection: This module is connected to the source of information and extract the profile and activity data of users.

7.3.2 Data processing layer

All the existing events will be detected in three different phases namely principal events verification, activity verification and symptoms analysis.

M2. Principal events verification: As the first step in the methodology, In this module the framework verifies whether the main events (work, family, home, health and travel) are changed or filled in profile of the user. If so, obviously that event has occurred. After confirming the principal events, we verify the activity of the user.

M3. Users' post verification: For each new published posts of the user (text and image) the mentioned methods in section 7.2 are used. Then we verify the posts of the user to detect them. For text and image posts of her, common methods mentioned in 7.2 are used. Moreover, the tag of the images (meta data) is analyzed to get better accuracy.

M4. Symptoms analysis: The aim of this module is to find those events which are not explicitly mentioned in the profile of users. To this end we use three features namely friend

list, interests and activity rate (number of posts, shares and reactions) in the given time period as indirect indicators of an occurred event. We first analyze above features to find out if an event took place and then try to identify its category by similarity detection. We say an event took place if number of added friends is more than average rate of adding friends (T_i in algorithm 3) and if the rate of adding interests is more than usual (T_{i1} in algorithm 4) for a user. The other possibility is when user's actual activity rate is more than her average rate. In order to find the event occurrence, for each feature we compare user's previous average rate of changes from user history (Fig. 7.3, module 6) with current status and if the difference is more than a threshold we assume it as an event. These activities are meaningful, means that something happened to the user life. In order to detect such activities, one can analyze the previous activity time series from user history and then compare it with current status. If a user's average adding friend rate in a period is X and now she has added more than X friends on that period it could be meaningful.

Then we do the similarity detection in which we are looking for the common attributes among the features that recently has been changed. In case of friend list, we seek the similarity between new added friends and in case of interests we explore the similarity among members of added group or page. Apart from these two features, it is almost impossible to find event yet we can find the occurrence of it from activity if user actual activity rate is more than her average one (T_{i2} in algorithm 4). The identified similarity (working in same company, living in same city and etc.) will be labeled as the detected event for that user. We try to cover all the possible situations to extract the event category. For instance, consider user A adds X = 3 friends in a week (assume her average rate of adding friend is 2), which are working in a same company. First we conclude that an event has been occurred (X>2) and eventually by analyzing the similarity of new added friends we conclude that she changed her job. Furthermore, there is a possible situation that user A may adds new friends from a meeting whom works on a same place. Considering available information it is tremendously hard to distinguish this situation with changing job event. However, using time series of user profile, if the user keep adding new activities related to new added friends common feature, we can conclude job change, else just a meeting.

Another example is when user B adds interests about a city in her profile in which the number of added interests are more than her average rate of adding interests. We can conclude that she may travel there, or she changed her home. However, the similarity between changing home and travel are tremendous, because in both situations, the user may adds some photos of a city, or add new friends from that city. In order to distinguish these two events the time series in user history records is used. In other words, if user keep adding new friends or photos from a new city we can infer she changed her city, else she just traveled there. Algorithms 3 and 4 demonstrate the overall process for the symptoms

Algorithm 3 Friend list analysis in symptoms module

```
1: for each user i [U_i] do
2:     T_i = Avg #(added friends) in period for U_i
3:     for each period j [P_j] do
4:         if #(added friends) by U_i in P_j > T_i
5:           if exists similarity between new friends of U_i
6:             an event has occurred [E_i]
7:             assign E_i for U_i in the data base
8:     end for
9: end for
```

event detection.

In algorithm 3 we detect the event from the friend list changes. First, using database in the history (Fig. 7.3 Update Historical Users Database) for each user we define a threshold based on average number of added friends in period in line 2. Note that period is week, month, year or etc. Then in line 4 we compare it with current number of added friends and if it satisfied the condition we will explore the similarity between new friends. If the similarity is found, we assign the found similarity as the occurred event E_i for user U_i and save it in data base in line 7.

In other algorithm we want to identify the event from interest and activity rate. Based on Facebook, Twitter and Google+, the interests are: Video, Places, Sports, Music, Movies, Shows, Books, Applications, Likes, Notes and Groups. A user may follow (like) them, or get a membership on each of them. In lines 2 and 3 we define threshold based on the rate of adding interests and activities for each user. If the number of added interests is more than average rate we seek the possible similarity among members of added interest(s) in line 6. If so, in line 8 we assign found event E_i in data base for the current user U_i. Then we verify her activity rate in lines 9 to 10. If the current activity rate is more than average rate, we conclude an event occurrence.

Update historical users databases: The history data base records every change and the order of them for each user. Given a user and period, the data base has her number of added friends, number of activities and happened event(s). As mentioned above, this history records is used in symptom analysis. Also, for each user all activities (adding friends, text and image posts and interests) before occurrence of an event is stored to predict the coming events (see section 7.4).

M5. Prediction Module: This module aims to predict future events which is discussed in section 7.4.

7.3.3 Visualization layer

Detected events: Whether the event(s) is found or not, the process of detecting events will be ended in this section. No events or the aggregated detected events of modules 2, 3 and 4 will be outputted as the system's outcome.

Predicted events: The last section will present the predicted events for each user.

Algorithm 4 Interests and activities analysis in symptoms module

```
for each user i [U_i] do
    T_i1 = Avg(rate of interests) added in period for U_i
    T_i2 = Avg(rate of activities) added in period for U_i
    for each period j [P_j] do
        if #interests added by U_i in P_j > T_i1
          if exist similarity between new interests
            an event has occurred [E_i]
            assign E_i for U_i in the data base
        if #activities added in P_j > T_i2
          an event has occurred
    end for
end for
```

7.4 Prediction Method of Events

The events with different probabilities are predicted unfavorably [147]. Generally, the studies on prediction in social network are about marketing, movie box-office, information dissemination, elections, macroeconomic, and miscellanea [148].

Lincoln is one of the laboratories working on the prediction of users and groups lives. Accordingly, their researchers discovered the pattern of individual life in social network using reality mining data-set analysis [149]. Based on [148] the major metrics in social media used for prediction are: Message characteristics (Sentiment metrics, Time series metrics) and social network characteristics (Terminology, Degree, Density, Centrality, Structural hole).

Table 7.2 – corresponding features to the events

Event	**Features**
Work	• New added friends which work on a same job • Adding interests about a job or company • Text or image posts related to it
Family	• Text or image posts related to family (love, baby, etc) • Adding new interests about this topic
Home	• New added friends who live in a same city • Adding interests about a city and it's places • Pictures from a same region • Text post about the city
Health	• This is not an easy task to predict health related events, but reaction of friends and asking about the person in their comments can be used as a hint.
Travel	• Text or image posts related to travel • Adding interests about a city or specific monument

The aim of this section is to propose a simple method to predict the coming activities and events of users based on historical activities and detected events of them. In previous section, we have presented a framework to identify events based on users social information. The outcome of the framework which are detected events and their corresponding features,

are stored in the framework database (Fig. 7.3, Update Historical Users Database). Based on the identified features for each detected event from previous historical activities of users, the prediction module aims to predict their future events by comparing ongoing behavioral pattern of users with their historical records. In general, the occurrence of events has a process consisting of user's activities. We stored them in the history database for each users during occurrence of events. By considering the prior activities of users before the occurrence of an event, we learn the process of an event occurrence and its behavioral pattern. In this way, if a user follows a similar pattern, the method will be able to predict her possible coming events. To this end, we match the current activities with the process of each event. For example, by analyzing the database information for different type of events, we get know the possible process for changing work which is adding new friends who works on a same job or adding new interests about that job. If a user follow this pattern, we can guess that she is going to change her job in near future consequently. Table 7.2 presents the features of possible process for each event that can lead us to predict the events.

Despite the similarity between travel and home (changing city) features, its a challenge to distinguish these two events from each other. However, there are some differences which can help to separate them. The one who is going to a trip will look for a hotel or a place for a short time. While most of people who change their city, have a job or looked for a job there.

7.5 Conclusion

Social media provides a golden opportunity for people to express themselves in their everyday life by posting about their feelings and their previous and future life events (e.g. changing a city, job, etc.) as well as non personal events that they care or affect them. In this era, users have the possibility to post their life events by means of text, image or video on different networks such as Facebook, Google+ and Tweeter. There has been an increasing amount of literature which aim to detect events from available users data. In this chapter, we proposed a hybrid framework for personal life event detection. Based on our knowledge, there are just few previous works targeting this goal, without providing a comprehensive approach that covers different types of activities due to the fact that most of these events are not defined in known categories. Furthermore, the existing detection methods mainly focused on text mining. Our proposed framework is a hybrid solution consisting of principal events, user post verification and a novel symptoms analysis. In these different steps, we manipulated text mining, image processing, activity and profiling methods to find and cover all possible events including events which are not explicitly mentioned by users but can be discovered from a set of evidences from the profile of users. Based on the

proposed detection framework and historical identified events, we also proposed a method to predict the potential coming events of users by comparing ongoing behavioral pattern of a user with historical records for different already identified events.

7.6 Potential future work directions

This study proposed a general hybrid framework on event detection and prediction and highlighted the key available methods in this topic of research. As the next step in follow up to this study, several ideas of research can be taken into account as follows:

Implementation and evaluation: In this chapter we introduced a comprehensive framework for detecting the events. Our method added the symptoms analysis in order to improve the accuracy of detection by identifying the events which are mentioned in users profile. According to this, one of the potential future work is evaluating this framework and produce the results accordingly.

Image object recognition: So far the tags and information (temporal, geographical and etc.) stuck to the images are used for event detection but not the actual image. It is a challenge yet actual images can be used for event detection e.g. discovering health event by analyzing an image post showing a bandaged hand. An object detection method is proposed by Felzenszwalb et al. [150] based on the mixtures of multi-scale models using Latent SVM [150]. Furthermore, Tankoyeu et al. [151] used standard clustering and ML techniques to classify the images posted by users. Hence, another future work direction is using image processing and object detection to identify events from image posts.

Anomaly/fraud detection: Since the users create massive amount of data in social medias, processing them is challenging. Therefore, one approach is improve the data processing by reducing the amount of data that should be analyzed with omitting the irrelevant text posts. Akuglu et al. [152] proposed an effective framework for discovering fraudsters and fake reviews in online review data sets which can be used to elide the fake text posts.

Text sentiment extraction: Sentiment analysis can improve text mining and consequently event detection task yet researchers pay much attention to other aspects such as event terms, textual features, temporal and person reference terms. One approach is amending the text mining by using sentiment analysis to distinguish the false negative sentences e.g. *I want hamburger so bad*. Unigram is one of the models related to this approach which can significantly detect such sentences.

Detecting the branches of main events: In addition to the main events, there are secondary events which are specific event inside the mentioned categories of event. These events are branches of the main ones such as broken bone event as a subset of health (based on Facebook event classification shown in figure 7.1). As long as current studies paid

attention to main events, an interesting direction for future work is finding the secondary ones.

Appendix A

Thesis Publications

Journal

- A. Mohammadinejad, R. Farahbakhsh, N. Crespi, *Consensus Opinion Model in Online Social Networks based on Influential Users*, Submitted to IEEE Access, 2018.

International Conferences

- A. Mohammadinejad, R. Farahbakhsh, N. Crespi, *OPIU: Opinion Propagation in Online Social Networks Using Influential Users Impact*, International Conference on Communications, IEEE ICC, Kansas City, USA, May 2018.
- A. Mohammadinejad, R. Farahbakhsh, N. Crespi, *Employing Personality Feature to Rank the Influential Users in Signed Networks*, The 9th IEEE International Conference on Social Computing and Networking, SocialCom, Atlanta, USA, October 2016.
- I. Javed, K. Toumi, N. Crespi, A. Mohammadinejad, *Br2Br: A Vector-based Trust Framework for WebRTC Calling Services*, 18th IEEE International Conference on High Performance Computing and Communications, HPCC, Sydney , Australia, December 2016.

Under Preparation Papers

- , *Detection and Prediction of Users' Life Event based on their activities on Online Social Networks*, International Conference on Communications, IEEE ICC, Shanghai, China, May 2019.

References

[1] D. Acemoglu and A. Ozdaglar, "Opinion dynamics and learning in social networks," *Dynamic Games and Applications*, vol. 1, no. 1, pp. 3–49, 2011.

[2] C. Altafini, "Dynamics of opinion forming in structurally balanced social networks," *PloS one*, vol. 7, no. 6, p. e38135, 2012.

[3] T. V. Martins, M. Pineda, and R. Toral, "Mass media and repulsive interactions in continuous-opinion dynamics," *EPL (Europhysics Letters)*, vol. 91, no. 4, p. 48003, 2010.

[4] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion word expansion and target extraction through double propagation," *Computational linguistics*, vol. 37, no. 1, pp. 9–27, 2011.

[5] C.-L. Hsu, M.-C. Chen, and V. Kumar, "How social shopping retains customers? capturing the essence of website quality and relationship quality," *Total quality management & business excellence*, vol. 29, no. 1-2, pp. 161–184, 2018.

[6] A. T. Stephen and O. Toubia, "Deriving value from social commerce networks," *Journal of marketing research*, vol. 47, no. 2, pp. 215–228, 2010.

[7] S. Chen and K. He, "Influence maximization on signed social networks with integrated pagerank," in *Smart City/SocialCom/SustainCom (SmartCity), 2015 IEEE International Conference on.* IEEE, 2015, pp. 289–292.

[8] X. Kong, J. Zhou, J. Zhang, W. Wang, and F. Xia, "Taprank: A time-aware author ranking method in heterogeneous networks," in *Smart City/SocialCom/SustainCom (SmartCity), 2015 IEEE International Conference on.* IEEE, 2015, pp. 242–246.

[9] H. Zhu, E. Chen, H. Xiong, H. Cao, and J. Tian, "Ranking user authority with relevant knowledge categories for expert finding," *World Wide Web*, vol. 17, no. 5, pp. 1081–1107, 2014.

[10] A. El-Korany, "Integrated expert recommendation model for online communities," *arXiv preprint arXiv:1311.3394*, 2013.

[11] J. Liu, Y.-I. Song, and C.-Y. Lin, "Competition-based user expertise score estimation," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval.* ACM, 2011, pp. 425–434.

[12] M. Bouguessa, B. Dumoulin, and S. Wang, "Identifying authoritative actors in question-answering forums: the case of yahoo! answers," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 866–874.

[13] P. Jurczyk and E. Agichtein, "Discovering authorities in question answer communities by using link analysis," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, 2007, pp. 919–922.

[14] J. Zhang, M. S. Ackerman, and L. Adamic, "Expertise networks in online communities: structure and algorithms," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 221–230.

[15] Z. Zhao, L. Zhang, X. He, and W. Ng, "Expert finding for question answering via graph regularized matrix completion," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 4, pp. 993–1004, 2015.

[16] M. Shahriari, S. Parekodi, and R. Klamma, "Community-aware ranking algorithms for expert identification in question-answer forums," in *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business*. ACM, 2015, p. 8.

[17] J. Zhang, J. Tang, and J. Li, "Expert finding in a social network," in *International Conference on Database Systems for Advanced Applications*. Springer, 2007, pp. 1066–1069.

[18] H. Deng, I. King, and M. R. Lyu, "Formal models for expert finding on dblp bibliography data," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008, pp. 163–172.

[19] B.-C. Chen, J. Guo, B. Tseng, and J. Yang, "User reputation in a comment rating environment," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 159–167.

[20] J. Guo, S. Xu, S. Bao, and Y. Yu, "Tapping on the potential of q&a community by recommending answer providers," in *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 2008, pp. 921–930.

[21] D. Mimno and A. McCallum, "Expertise modeling for matching papers with reviewers," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007, pp. 500–509.

[22] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 261–270.

[23] A. Kardan, A. Omidvar, and F. Farahmandnia, "Expert finding on social network with link analysis approach," in *Electrical Engineering (ICEE), 2011 19th Iranian Conference on*. IEEE, 2011, pp. 1–6.

[24] M. Rafiei and A. A. Kardan, "A novel method for expert finding in online communities based on concept map and pagerank," *Human-centric computing and information sciences*, vol. 5, no. 1, p. 10, 2015.

[25] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[26] Z. Liu and Y. Zhang, "Structures or texts? a dynamic gating method for expert finding in cqa services," in *International Conference on Database Systems for Advanced Applications*. Springer, 2018, pp. 201–208.

[27] K. Balog, T. Bogers, L. Azzopardi, M. De Rijke, and A. Van Den Bosch, "Broad expertise retrieval in sparse data environments," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 551–558.

[28] Y. Lu, X. Quan, J. Lei, X. Ni, W. Liu, and Y. Xu, "Semantic link analysis for finding answer experts," *Journal of Information Science and Engineering*, vol. 28, no. 1, pp. 51–65, 2012.

[29] H.-C. Yang, D. B. Lange, and X. Zhang, "Identifying influential users of a social networking service," Apr. 3 2018, uS Patent 9,934,512.

[30] A. Zareie, A. Sheikhahmadi, and M. Jalili, "Influential node ranking in social networks based on neighborhood diversity," *Future Generation Computer Systems*, vol. 94, pp. 120 – 129, 2019.

[31] Z. Wei, Y. Jing, D. Xiao-yu, Z. Xiao-mei, H. Hong-yu, and Z. Qing-chao, "Groups make nodes powerful: Identifying influential nodes in social networks based on social conformity theory and community features," *Expert Systems with Applications*, 2019.

[32] A. Sheikhahmadi, M. A. Nematbakhsh, and A. Zareie, "Identification of influential users by neighbors in online social networks," *Physica A: Statistical Mechanics and its Applications*, vol. 486, pp. 517–534, 2017.

[33] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge press, 1994.

[34] G. Kou, Y. Zhao, Y. Peng, and Y. Shi, "Multi-level opinion dynamics under bounded confidence," *PloS one*, vol. 7, no. 9, p. e43507, 2012.

[35] H. Liang, Y. Yang, and X. Wang, "Opinion dynamics in networks with heterogeneous confidence and influence," *Physica A: Statistical Mechanics and its Applications*, vol. 392, no. 9, pp. 2248–2256, 2013.

[36] L. Zhang, B. Liu, S. H. Lim, and E. O'Brien-Strain, "Extracting and ranking product features in opinion documents," in *Proceedings of the 23rd international conference on computational linguistics: Posters*, 2010, pp. 1462–1470.

[37] J. Yang and J. Leskovec, "Modeling information diffusion in implicit networks," in *ICDM*. IEEE, 2010, pp. 599–608.

[38] X. Wang, Y. Liu, G. Zhang, Y. Zhang, H. Chen, and J. Lu, "Mixed similarity diffusion for recommendation on bipartite networks," *IEEE Access*, vol. 5, pp. 21 029–21 038, 2017.

[39] M. Cha, A. Mislove, and K. P. Gummadi, "A measurement-driven analysis of information propagation in the flickr social network," in *ACM WWW*, 2009, pp. 721–730.

[40] Y. Shang, "Deffuant model of opinion formation in one-dimensional multiplex networks," *Journal of Physics A: Mathematical and Theoretical*, vol. 48, no. 39, p. 395101, 2015.

[41] R. Durrett, J. P. Gleeson, A. L. Lloyd, P. J. Mucha, F. Shi, D. Sivakoff, J. E. Socolar, and C. Varghese, "Graph fission in an evolving voter model," *Proceedings of the National Academy of Sciences*, vol. 109, no. 10, pp. 3682–3687, 2012.

[42] J. Wu, F. Chiclana, H. Fujita, and E. Herrera-Viedma, "A visual interaction consensus model for social network group decision making with trust propagation," *Knowledge-Based Systems*, vol. 122, pp. 39–50, 2017.

[43] P. Berkhim, Z. Xu, J. Mao, D. E. Rose, A. Taha, and F. Maghoul, "Trust propagation through both explicit and implicit social networks," Feb. 21 2017, uS Patent 9,576,029.

[44] S. Deng, L. Huang, G. Xu, X. Wu, and Z. Wu, "On deep learning for trust-aware recommendations in social networks," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 5, pp. 1164–1177, 2017.

[45] Y. Dong, Z. Ding, L. Martínez, and F. Herrera, "Managing consensus based on leadership in opinion dynamics," *Information Sciences*, vol. 397, pp. 187–205, 2017.

[46] M. Ghavipour and M. R. Meybodi, "Trust propagation algorithm based on learning automata for inferring local trust in online social networks," *Knowledge-Based Systems*, vol. 143, pp. 307–316, 2018.

[47] Q. Du, H. Hong, G. A. Wang, P. Wang, and W. Fan, "Crowdiq: A new opinion aggregation model," 2017.

[48] S. Chatterjee, A. Mukhopadhyay, and M. Bhattacharyya, "Quality enhancement by weighted rank aggregation of crowd opinion," *arXiv preprint arXiv:1708.09662*, 2017.

[49] Y. Liu, C. Liang, F. Chiclana, and J. Wu, "A trust induced recommendation mechanism for reaching consensus in group decision making," *Knowledge-Based Systems*, vol. 119, pp. 221–231, 2017.

[50] J. Wu, F. Chiclana, and E. Herrera-Viedma, "Trust based consensus model for social network in an incomplete linguistic information context," *Applied Soft Computing*, vol. 35, pp. 827–839, 2015.

[51] N. Capuano, F. Chiclana, H. Fujita, E. Herrera-Viedma, and V. Loia, "Fuzzy group decision making with incomplete information guided by social influence," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 3, pp. 1704–1718, 2018.

[52] H. Zhang, Y. Dong, and E. Herrera-Viedma, "Consensus building for the heterogeneous large-scale gdm with the individual concerns and satisfactions," *IEEE Transactions on Fuzzy Systems*, vol. 26, no. 2, pp. 884–898, 2018.

[53] M. J. del Moral, F. Chiclana, J. M. Tapia, and E. Herrera-Viedma, "A comparative study on consensus measures in group decision making," *International Journal of Intelligent Systems*, 2018.

[54] M. J. Del Moral, J. M. Tapia, F. Chiclana, A. Al-Hmouz, and E. Herrera-Viedma, "An analysis of consensus approaches based on different concepts of coincidence," *Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 4, pp. 2247–2259, 2018.

[55] Y. Dong, Q. Zha, H. Zhang, G. Kou, H. Fujita, F. Chiclana, and E. Herrera-Viedma, "Consensus reaching in social network group decision making: Research paradigms and challenges," *Knowledge-Based Systems*, 2018.

[56] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas, "Link analysis ranking: algorithms, theory, and experiments," *ACM Transactions on Internet Technology (TOIT)*, vol. 5, no. 1, pp. 231–297, 2005.

[57] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[58] A. Y. Ng, A. X. Zheng, and M. I. Jordan, "Stable algorithms for link analysis," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2001, pp. 258–266.

[59] R. Lempel and S. Moran, "The stochastic approach for link-structure analysis (salsa) and the tkc effect1," *Computer Networks*, vol. 33, no. 1-6, pp. 387–401, 2000.

[60] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of personality*, vol. 60, no. 2, pp. 175–215, 1992.

[61] M. F. Scheier, C. S. Carver, and M. W. Bridges, "Optimism, pessimism, and psychological well-being," *Optimism and pessimism: Implications for theory, research, and practice*, vol. 1, pp. 189–216, 2001.

[62] W. N. Dember, S. H. Martin, M. K. Hummer, S. R. Howe, and R. S. Melton, "The measurement of optimism and pessimism," *Current Psychology*, vol. 8, no. 2, pp. 102–119, 1989.

[63] B. Liu, *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media, 2007.

[64] V. A. Traag, Y. E. Nesterov, and P. Van Dooren, "Exponential ranking: taking into account negative links," in *International Conference on Social Informatics*. Springer, 2010, pp. 192–202.

[65] A. Mishra and A. Bhattacharya, "Finding the bias and prestige of nodes in networks based on trust scores," in *Proceedings of the 20th international conference on World wide web*. ACM, 2011, pp. 567–576.

[66] C. De Kerchove and P. Van Dooren, "The pagetrust algorithm: How to rank web pages when negative links are allowed?" in *SDM*. SIAM, 2008, pp. 346–352.

[67] A. M. Z. Bidoki and N. Yazdani, "Distancerank: An intelligent ranking algorithm for web pages," *Information Processing & Management*, vol. 44, no. 2, pp. 877–892, 2008.

[68] A. S. Butt, A. Haller, and L. Xie, "Ontology search: An empirical evaluation," in *International Semantic Web Conference*. Springer, 2014, pp. 130–147.

[69] G. Beigi, J. Tang, and H. Liu, "Signed link analysis in social media networks." in *ICWSM*, 2016, pp. 539–542.

[70] H. Mei, Y. Zhang, and X. Meng, "Csa: A credibility search algorithm based on different query in unstructured peer-to-peer networks," *Mathematical Problems in Engineering*, vol. 2014, 2014.

[71] J. M. Kouzes and B. Z. Posner, *The Jossey-Bass academic administrator's guide to exemplary leadership*. John Wiley & Sons, 2003, vol. 131.

[72] Hu, Weishu and Gong, Zhiguo, "Multi-relational reinforcement for computing credibility of nodes," *World Wide Web*, pp. 1–22, 2015.

[73] W. Hu and Z. Gong, "Assessing the credibility of nodes on multiple-relational social networks," in *International Conference on Web Information Systems Engineering*. Springer, 2014, pp. 62–77.

[74] G. Jeh and J. Widom, "Simrank: a measure of structural-context similarity," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 538–543.

[75] P. Clifford and A. Sudbury, "A model for spatial conflict," *Biometrika*, vol. 60, no. 3, pp. 581–588, 1973.

[76] V. Sood, T. Antal, and S. Redner, "Voter models on heterogeneous networks," *Physical Review E*, vol. 77, no. 4, 2008.

[77] F. Riahi, Z. Zolaktaf, M. Shafiei, and E. Milios, "Finding expert users in community question answering," in *Proceedings of the 21st International Conference on World Wide Web*. ACM, 2012, pp. 791–798.

[78] M. Moussaïd, J. E. Kämmer, P. P. Analytis, and H. Neth, "Social influence and the collective dynamics of opinion formation," *PloS one*, vol. 8, no. 11, p. e78433, 2013.

[79] B. Latane *et al.*, "The psychology of social impact," *American psychologist*, vol. 36, no. 4, pp. 343–356, 1981.

[80] D. G. Myers and G. D. Bishop, "Discussion effects on racial attitudes," *Science*, vol. 169, no. 3947, pp. 778–779, 1970.

[81] A. Mohammadinejad, R. Farahbakhsh, and N. Crespi, "Employing personality feature to rank the influential users in signed networks," in *IEEE SocialCom*, 2016, pp. 346–353.

[82] B. Golub and M. O. Jackson, "Naive learning in social networks and the wisdom of crowds," *American Economic Journal: Microeconomics*, vol. 2, no. 1, pp. 112–149, 2010.

[83] L. E. Blume, "The statistical mechanics of strategic interaction," *Games and economic behavior*, vol. 5, no. 3, 1993.

[84] G. Pasi and R. R. Yager, "Modeling the concept of majority opinion in group decision making," *Information Sciences*, vol. 176, no. 4, pp. 390–414, 2006.

[85] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Signed networks in social media," in *ACM SIGCHI*, 2010, pp. 1361–1370.

[86] H. Aghaie, S. Shafieezadeh, and B. Moshiri, "A new modified fuzzy topsis for group decision making using fuzzy majority opinion based aggregation," in *ICEE*. IEEE, 2011, pp. 1–6.

[87] X.-M. Si, W.-D. Wang, and Y. Ma, "Role of propagation thresholds in sentiment-based model of opinion evolution with information diffusion," *Physica A: Statistical Mechanics and its Applications*, vol. 451, pp. 549–559, 2016.

[88] S. Alonso, E. Herrera-Viedma, F. Chiclana, and F. Herrera, "A web based consensus support system for group decision making problems and incomplete preferences," *Information Sciences*, vol. 180, no. 23, pp. 4477–4495, 2010.

[89] E. Herrera-Viedma, F. Chiclana, F. Herrera, and S. Alonso, "Group decision-making model with incomplete fuzzy preference relations based on additive consistency," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 1, pp. 176–189, 2007.

[90] F. Chiclana, J. T. GarcíA, M. J. del Moral, and E. Herrera-Viedma, "A statistical comparative study of different similarity measures of consensus in group decision making," *Information Sciences*, vol. 221, pp. 110–123, 2013.

[91] Y. Dong, H. Zhang, and E. Herrera-Viedma, "Integrating experts' weights generated dynamically into the consensus reaching process and its applications in managing non-cooperative behaviors," *Decision Support Systems*, vol. 84, pp. 1–15, 2016.

[92] Y. Dong, X. Chen, and F. Herrera, "Minimizing adjusted simple terms in the consensus reaching process with hesitant linguistic assessments in group decision making," *Information Sciences*, vol. 297, pp. 95–117, 2015.

[93] T. González-Arteaga, R. de Andrés Calle, and F. Chiclana, "A new measure of consensus with reciprocal preference relations: The correlation consensus degree," *Knowledge-Based Systems*, vol. 107, pp. 104–116, 2016.

[94] F.-Y. Meng, Q.-X. An, C.-Q. Tan, and X.-H. Chen, "An approach for group decision making with interval fuzzy preference relations based on additive consistency and consensus analysis," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 8, pp. 2069–2082, 2017.

[95] Z.-J. Wang and J. Lin, "Ratio-based similarity analysis and consensus building for group decision making with interval reciprocal preference relations," *Applied Soft Computing*, vol. 42, pp. 260–275, 2016.

[96] Z.-J. Wang and X. Tong, "Consistency analysis and group decision making based on triangular fuzzy additive reciprocal preference relations," *Information Sciences*, vol. 361, pp. 29–47, 2016.

[97] Y. Dong, N. Luo, and H. Liang, "Consensus building in multiperson decision making with heterogeneous preference representation structures: A perspective based on prospect theory," *Applied Soft Computing*, vol. 35, pp. 898–910, 2015.

[98] E. Herrera-Viedma, S. Alonso, F. Chiclana, and F. Herrera, "A consensus model for group decision making with incomplete fuzzy preference relations," *IEEE Transactions on fuzzy Systems*, vol. 15, no. 5, pp. 863–877, 2007.

[99] I. Palomares, F. J. Estrella, L. Martínez, and F. Herrera, "Consensus under a fuzzy context: Taxonomy, analysis framework afryca and experimental case of study," *Information Fusion*, vol. 20, pp. 252–271, 2014.

[100] J. Wu and F. Chiclana, "Visual information feedback mechanism and attitudinal prioritisation method for group decision making with triangular fuzzy complementary preference relations," *Information Sciences*, vol. 279, pp. 716–734, 2014.

[101] ——, "Multiplicative consistency of intuitionistic reciprocal preference relations and its application to missing values estimation and consensus building," *Knowledge-Based Systems*, vol. 71, pp. 187–200, 2014.

[102] Z. Gong, X. Xu, H. Zhang, U. A. Ozturk, E. Herrera-Viedma, and C. Xu, "The consensus models with interval preference opinions and their economic interpretation," *Omega*, vol. 55, pp. 81–90, 2015.

[103] J. Kacprzyk, S. ZADROŻNY, and Z. W. RAŚ, "How to support consensus reaching using action rules: a novel approach," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 18, no. 04, pp. 451–470, 2010.

[104] Z. Wu and J. Xu, "A consistency and consensus based decision support model for group decision making with multiplicative preference relations," *Decision Support Systems*, vol. 52, no. 3, pp. 757–767, 2012.

[105] ——, "Managing consistency and consensus in group decision making with hesitant fuzzy linguistic preference relations," *Omega*, vol. 65, pp. 28–40, 2016.

[106] R. R. Yager and N. Alajlan, "An intelligent interactive approach to group aggregation of subjective probabilities," *Knowledge-Based Systems*, vol. 83, pp. 170–175, 2015.

[107] R. Hegselmann, U. Krause, *et al.*, "Opinion dynamics and bounded confidence models, analysis, and simulation," *Journal of artificial societies and social simulation*, vol. 5, no. 3, 2002.

[108] U. Krause, "A discrete nonlinear and non-autonomous model of consensus formation," *Communications in difference equations*, vol. 2000, pp. 227–236, 2000.

[109] E. J. Thomas and C. F. Fink, "Models of group problem solving." *The Journal of Abnormal and Social Psychology*, vol. 63, no. 1, p. 53, 1961.

[110] C. Martini and J. Sprenger, "Opinion aggregation and individual expertise," 2015.

[111] K. Lehrer and C. Wagner, *Rational consensus in science and society: A philosophical and mathematical study*. Springer Science & Business Media, 2012, vol. 24.

[112] K. Lehrer, "When rational disagreement is impossible," *Noûs*, pp. 327–332, 1976.

[113] R. R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decisionmaking," *IEEE Transactions on systems, Man, and Cybernetics*, vol. 18, no. 1, pp. 183–190, 1988.

[114] ——, "Families of owa operators," *Fuzzy sets and systems*, vol. 59, no. 2, pp. 125–148, 1993.

[115] S. Kubler, W. Derigent, A. Voisin, K. Främling, and A. Thomas, "Methods of aggregation of expert opinions in the framework of intelligent products," in *11th IFAC Workshop on Intelligent Manufacturing Systems, IMS'2013*, 2013, pp. 163–168.

[116] I. J. Pérez, F. J. Cabrerizo, S. Alonso, and E. Herrera-Viedma, "A new consensus model for group decision making problems with non-homogeneous experts," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 4, pp. 494–498, 2014.

[117] M. B. Habib and M. van Keulen, "Information extraction for social media," in *Proceedings of the Third Workshop on Semantic Web and Information Extraction (SWAIE 2014), Dublin, Ireland*, vol. W14-62. Dublin: Association for Computational Linguistics, August 2014.

[118] A. Ritter, O. Etzioni, S. Clark, *et al.*, "Open domain event extraction from twitter," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2012.

[119] W. X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E.-P. Lim, and X. Li, "Topical keyphrase extraction from twitter," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 379–388.

[120] B. Di Eugenio, N. Green, and R. Subba, "Detecting life events in feeds from twitter," in *2013 IEEE Seventh International Conference on Semantic Computing*. Ieee, 2013, pp. 274–277.

[121] S. Choudhury and H. Alani, "Personal life event detection from social media," in *Hypertext and Social Media Conference (Hypertext 2014)*. ACM, 2014.

[122] J. Li, A. Ritter, C. Cardie, and E. Hovy, "Major life event extraction from twitter based on congratulations/condolences speech acts," in *Proceedings of Empirical Methods in Natural Language Processing*, 2014.

[123] P. Cavalin, M. Gatti, and C. Pinhanez, "Towards personalized offers by means of life event detection on social media and entity matching."

[124] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[125] L. M. Aiello, G. Petkos, C. Martin, D. Corney, S. Papadopoulos, R. Skraba, A. Goker, I. Kompatsiaris, and A. Jaimes, "Sensing trending topics in twitter," *Multimedia, IEEE Transactions on*, vol. 15, no. 6, pp. 1268–1282, 2013.

[126] J. Allan, "Introduction to topic detection and tracking," in *Topic detection and tracking*. Springer, 2002, pp. 1–16.

[127] Z. Tan, P. Zhang, J. Tan, and L. Guo, "A multi-layer event detection algorithm for detecting global and local hot events in social networks," *Procedia Computer Science*, vol. 29, pp. 2080–2089, 2014.

[128] K. Sathiyamurthy, G. Shanmugavalli, and N. Udayalakshmi, "Event detection and summarization based on social networks and semantic query expansion," *IJNLC*, vol. 3, no. 5/6, 2014.

[129] P. Khare and B. R. Heravi, "Towards social event detection and contextualisation for journalists," in *In the proceedings of the AHA! Workshop on Information Discovery in Text, at the 25th International Conference on Computational Linguistics, August 2014*, 2014.

[130] C. C. Aggarwal and K. Subbian, "Event detection in social streams." in *SDM*, vol. 12. SIAM, 2012, pp. 624–635.

[131] S. M. Alqhtani, S. Luo, and B. Regan, "Fusing text and image for event detection in twitter," *arXiv preprint arXiv:1503.03920*, 2015.

[132] M. Zaharieva, M. Zeppelzauer, and C. Breiteneder, "Automated social event detection in large photo collections," in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*. ACM, 2013, pp. 167–174.

[133] G. Petkos, S. Papadopoulos, V. Mezaris, and Y. Kompatsiaris, "Social event detection at mediaeval 2014: Challenges, datasets, and evaluation," in *MediaEval 2014 Workshop, Barcelona, Spain*, 2014.

[134] J. Weng and B.-S. Lee, "Event detection in twitter." *ICWSM*, vol. 11, pp. 401–408, 2011.

[135] Q. Zhao and P. Mitra, "Event detection and visualization for social text streams," in *ICWSM*, 2007.

[136] Q. Zhao, P. Mitra, and B. Chen, "Temporal and information flow based event detection from social text streams," in *AAAI*, vol. 7, 2007.

[137] H. Sayyadi, M. Hurst, and A. Maykov, "Event detection and tracking in social streams." in *ICWSM*, 2009.

[138] G. Ifrim, B. Shi, and I. Brigadir, "Event detection in twitter using aggressive filtering and hierarchical tweet clustering." in *SNOW-DC@ WWW*, 2014, pp. 33–40.

[139] H. Becker, M. Naaman, and L. Gravano, "Learning similarity metrics for event identification in social media," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010.

[140] H. Becker and M. e. a. Naaman, "Event identification in social media," in *WebDB*, 2009.

[141] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang, "Tedas: A twitter-based event detection and analysis system," in *28th international conference on Data engineering (icde)*. IEEE, 2012.

[142] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 851–860.

[143] M. Walther and M. Kaisser, "Geo-spatial event detection in the twitter stream," in *Advances in Information Retrieval*. Springer, 2013.

[144] L. Jalali, D. Huo, H. Oh, M. Tang, S. Pongpaichet, and R. Jain, "Personicle: Personal chronicle of life events," in *Workshop on Personal Data Analytics in the Internet of Things (PDA@ IOT) at the 40th International Conference on Very Large Databases (VLDB), Hangzhou, China*, 2014.

[145] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005.

[146] L. S. Davis, S. A. Johns, and J. Aggarwal, "Texture analysis using generalized co-occurrence matrices," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, no. 3, pp. 251–259, 1979.

[147] J. Wolfers and E. Zitzewitz, "Prediction markets," National Bureau of Economic Research, Tech. Rep., 2004.

[148] S. Yu and S. Kak, "A survey of prediction using social media," *arXiv preprint arXiv:1203.1647*, 2012.

[149] C. K. Dagli and W. M. Campbell, "Individual and group dynamics in the reality mining corpus," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 2012, pp. 61–70.

[150] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, 2010.

[151] I. Tankoyeu, J. Paniagua, J. Stöttinger, and F. Giunchiglia, "Event detection and scene attraction by very simple contextual cues," in *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*. ACM, 2011, pp. 1–6.

[152] L. Akoglu, R. Chandy, and C. Faloutsos, "Opinion fraud detection in online reviews by network effects." *ICWSM*, vol. 13, pp. 2–11, 2013.

List of figures

List of tables