ELECTRONIC WORD-OF-MOUTH: APPLICATIONS IN PRODUCT RECOMMENDATION AND CRISIS INFORMATION DISSEMINATION

NARGIS PERVIN

(M.Tech, I.S.I. Kolkata, M.Sc. I.I.T. Roorkee)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF INFORMATION SYSTEMS

NATIONAL UNIVERSITY OF SINGAPORE

2014

I would like to dedicate this thesis to my loving parents who taught me that even the largest task can be accomplished if it is executed one step at a time.

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis. This thesis has also not been submitted for any degree in any university previously.

(NARGIS PERVIN)

Acknowledgements

Productive research and educational achievement require the collaboration and support of many people. A Ph.D. project is no exception and in fact, its building blocks are laid over the years with the contribution of numerous persons. As I complete this thesis, bringing to a close another chapter in my life, I wish to take this opportunity to write a few lines to express my appreciation to the many persons who have assisted and encouraged me in this long journey.

First and foremost, I would like to express my deep and earnest gratitude to my supervisor, Professor Anindya Datta for the opportunity to work with his esteemed research group, especially for allowing me a great degree of independence and creative freedom to explore myself.

I am grateful to Professors Kaushik Dutta, Professor Tulika Mitra, and Professor Tuan Quang Phan, who commented on my research and reviewed the thesis. My special thanks to Professor Narayan Ramasubbu, Professor Debra Vandermeer for their encouragement, guidance, and helpful suggestions in different stages of my PhD journey.

My sincere thanks go to Professor Hideaki Takeda, National Institute of

Informatics, Japan, for providing me the internship opportunity in his group and supporting me to work on exciting projects. I would likewise convey my deep regards to Professor Fujio Toriumi (The University of Tokyo, Japan) for permitting me to use the dataset for my research analysis.

I am grateful to all past and present members of our research group. I would take this opportunity to thank all my lab mates: Dr. Bao Yang, Dr. Fang Fang, Xiaoying Xu, Kajanan Sangaralingam for all their help in last four years. In my daily work I have been privileged with a friendly and upbeat group of fellow students : Prasanta Bhattacharya, Vivek Singh, for the stimulating discussions and exciting research ideas they shared. My special thanks go to Satish Krishnan, Rohit Nishant, Supunmali Ahangama, Nadee Goonawardene, and Upasna Bhandari. I would love to thank all my friends in Singapore for all the fun-filled moments we shared during all those years in Singapore. It was hardly possible for me to thrive in my doctoral work without the precious support of these personalities.

Finally, I am eternally indebted to my parents, my brother for supporting me spiritually throughout my life and having the perpetual belief in me. This thesis would not have been completed without the immense assistance and constant long-distance support, personal divine guidance from my beloved husband Dr. Md. Mahiuddin Baidya and my supportive parent-in-laws.

Contents

Appendix

This page is left intentionally blank.

Summary

Electronic word-of-Mouth (eWOM) can be perceived as

"Any positive or negative statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet."

- Hennig-Thurau, Qwinner, Walsh and Gremler (2004).

The eWOM plays a central role starting from product recommendations to social awareness, which is the quintessence of this thesis. It contains three essays. The first one aims to study how eWOM, in the form of user comments, is beneficial in recommendations of high-scale products. The other two essays investigate the role of eWOM in information diffusion in the context of online social networks. Prior researchers have shown that eWOM is extremely useful in case of recommendations for various items such as movies, books, etc. However, as far as the scale is concerned, domains like mobile app ecosystem are several times larger than any of these existing consumer products, both in terms of number of items and consumers. Hence, the existing recommendation techniques cannot be applied directly to mobile apps. In the first essay, we have proposed an approach to generate mobile app recommendations that combines the association rule based recommendation technique along with collaborative filtering technique. Our proposed approach recommends apps solving the monotonicity and scalability issue. To evaluate the approach, we have experimented with mobile app user data. Experimental results yield good accuracy (15% increase in precision) while maintaining diversity (91% inter-list diversity) in the recommendation list in a scalable fashion. The second essay examines information propagation using the retweet feature on Twitter where information flows in a large network through cascades of followers. In extant literature, the bias in diffusion analysis is inevitable because of the unstandardized retweet practices. Our approach combines the activity network with the follower network and introduces the concept of Information Diffusion Impact (IDI), which represents the overall impact of the user on the diffusion of information. With two event-centric Twitter datasets, we characterize important user roles in information propagation at the time of crisis and discuss the evolution of these roles over time along with other retweetablity factors. Our findings show that user roles in information propagation are very much crucial and evolves due to event. In addition, we have experimentally shown that disruptive events have a strong influence on retweetability and replicated our findings in another dataset to validate the robustness of our approach. Hashtags in microblogs provide discoverability and in turn increase the reachability of tweets. Despite its significant influence on retweetability, a little has been unravelled to understand what contributes to the popularity of a hashtag. Further, the majority of the hashtags (around 50%) in a tweet generally occurs in groups. The third study proposed an econometric model to investigate how the co-occurrence of hashtags affects its popularity, which is not addressed heretofore. Findings indicate that if a hashtag appears with other similar (dissimilar) hashtags, popularity of the focal hashtag increases (decreases). Interestingly, however, these results reverse when dissimilar hashtags appear along with a URL in the tweet. These findings can direct the practitioners to implement efficient policies for product advertisement with brand hashtags. Overall, eWOM in the field of app recommendation and information diffusion on Twitter at the time of crisis have been critically investigated, which will not only lead to deep understanding of eWOM in emerging domains, but more importantly, provides practical implications for efficient policy making in product recommendation, advertisement, and information diffusion.

This page is left intentionally blank.

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

The most well-defined and extensive definition of electronic word-of-mouth (eWOM) till date is given by Hennig-Thurau et al. (2004):

"Any positive or negative statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet."

With the emergence of Web 2.0 massive user-generated-contents are produced online in social media, product reviews, blogs, etc. The escalating use of the internet as a communication platform capacitates word-of-mouth as a powerful and useful resource for consumers as well as merchandisers (Peres et al., 2011; Chevalier and Mayzlin, 2006;

Okada and Yamamoto, 2011). In fact, social media turns out as a relatively inexpensive platform to implement marketing campaigns for organizations. This overwhelming information on web 2.0 also concurrently offers consumers the direct access to the digital word of mouth (eWOM) before making a purchase decision (Hennig-Thurau et al., 2004). In addition, through this one way communication medium the consumers can express their views of satisfaction or dissatisfaction by writing an online review after experiencing a product. While positive WOM results in a good brand experience and are spread by satisfied customers or 'brand ambassadors', negative messages are spread by unsatisfied customers or 'detractors' (Charlett et al., 1995; Chatterjee, 2001). Earlier researches (Okada and Yamamoto, 2011; Chatterjee, 2001) have investigated the influence of electronic word-of-mouth on customers' purchase intention and also explored the varying effects of positive and negative word-of-mouth.

Similar to online product reviews, eWOM has also been adapted in social networking sites or blogs in a multifaceted manner where users can engage themselves not just in one way conversation but also in bi-directional communication. Particularly, in Twitter, followers can comment on posts or retweet to agree with and/or to promote it. By the act of retweeting the same message is visible to a larger audience, enhancing the popularity of the message and thus, social networks act as a medium of transmission of electronic word-of-mouth. Contrary to face-to-face conversation, in digital communication messages travel over long distances very quickly. If everyone passes a message only to two people in their friends circle, the message can reach to an exponential number of people. However, in practice the behavior of users is not so predictable. Hence, the transmission of a message through the social network tools turn out to be fairly an intricate process to model. Overall, word-of-mouth plays a central role starting from product recommendations to social awareness, which is the quintessence of this dissertation. The thesis contains three separate essays dealing with electronic word-of-mouth.

The first essay uses word-of-mouth in the form of user comments for generating recommendations of high-scale products. Here, by high-scale products we mean the products with rapid growth rate, e.g., mobile applications (mobile apps). The mobile apps are different from other digital products. While 100 books and 250 music get released weekly, there are 15000 mobile apps that release world-wide on a weekly basis (Datta et al., 2011) as per 2011 statistics, which has increased up to 32,5000 for mobile apps only in the iTunes app store (Costello, 2014). Here, we ask ourselves the question, "do the traditional algorithms used for books and music recommendations can be applied for mobile apps?" We anticipate that the existing mechanisms seem to be not applicable as they take a longer time to run and by the time new products are factored in, the recommended products would have grown older. In addition, a large volume of apps makes the discovery of a particular app more challenging. In order to generate recommendations for a mobile app user, it is necessary to know the apps which are available in the user's mobile device. However, gaining the access to this information is not straightforward and raises privacy concerns. These limitations could be mitigated by using the user's app reviews in the corresponding app store. The fact that app users can write app reviews, if and only if the user has installed the app on his smart device, makes app reviews as the best representative of app usage. Therefore, in this research, mobile app reviews have been used to recommend mobile apps to smartphone users. A scalable recommendation algorithm has been built for mobile applications and it has been experimented against the baseline algorithms to show its applicability in a practical scenario.

Currently, Twitter is one of the most popular social media for communication (Krishnamurthy et al., 2008; Kwak et al., 2010). In Twitter, information diffuses very rapidly through reposting of someone else's tweet. The repost of a tweet is commonly called as a retweet, which is another form of eWOM. Billions of dollars are spent for advertising products, political campaigning, and marketing in these social media. Particularly, in product advertising and campaigning through social media, brands or companies seek attention from a large audience very rapidly. This demands recognition of the potential and influential target audience in the Twitter network, who in turn can promote the product by tweeting/retweeting the product related information to his or her friends and followers. Therefore, it is very important to identify the communicators in the diffusion process and investigate their roles in diffusion mechanism. In addition, it is also essential to understand the factors affecting retweetability (probability of a tweet getting retweeted) in the first place. This motivates us to examine information propagation using the retweet feature in Twitter, which is the focus of our second study. Here, we classify the user roles in information propagation and systematically investigate the impact of these user roles on retweetability along with other factors.

Twitter (and other social media) does not only diffuse the information rapidly, but also remains active during natural calamities when traditional communication systems like television, radio, telephones, newspaper, etc. are not at all useful, mostly because of power outage. In emergency situations, it is of utmost importance to broadcast event-related information to a large audience, especially to the needy users very quickly. This is why in this study, we have also examined whether event (e.g., earthquake) has any effect on the retweetability factors and how the effects of these factors change due to emergency situations. The third essay entitled "Hashtag Popularity on Twitter: Analyzing Co-occurrence of Multiple Hashtags" uses the Twitter dataset of the Great Eastern Japan earthquake and investigates the factors affecting the popularity of hashtags. Hashtags are used to bookmark topics of interest by adding a "#" before keywords or phrase which facilitates users to categorize and track interesting events or topics. The concept was first introduced by the Twitter users and recently gained popularity in other social media like Facebook, Instagram etc. On Twitter, one can note that hashtags appear in groups, i.e., a hashtag usually comes with other hashtags. Sometimes these co-appearing hashtags are similar, one is a variant of another and often they are totally dissimilar. This spawns the question whether this similarity/dissimilarity is random or carry certain patterns. Herein, we investigate the characteristics of the hashtags that co-appear. Literature on metacognition states that when there is unfamiliarity towards an information, metacognition difficulty to process and recall the information increases (Pocheptsova et al., 2010). With the increase of difficulty level, popularity of the hashtag decreases. In such a circumstance, introduction of extra information will improve its popularity. It will be interesting to examine the effect of adding URL in the tweet when the hashtags are dissimilar. Moreover, we will check whether an external event has any impact on the process.

1.2 Contribution

Our studies aim to investigate the role of word of mouth (WOM) in the context of web 2.0. Precisely, the contribution of each study is discussed below:

- In study 1, we have investigated how word of mouth plays a role in the context of recommending products. Prior researchers have shown that word of mouth is very useful in the case of recommending movies, books, etc. However, as discussed earlier, products like mobile applications are very different compared to digital goods like movies as per the scale of the products. Therefore, generating recommendations for the mobile apps is very challenging from the perspective of scalability while maintaining accuracy. Further, a good recommender system should offer a diverse choice of relevant items, allowing users to select from a broad range of options related to their taste. It is important to mention that generating diverse recommendations is not simply a matter of selecting a set of highly dissimilar items - one still has to give importance to relevance. Overall, generating accurate and diverse recommendations in a scalable fashion is highly demanding, but most of the prior studies primarily focus on improving the accuracy of the recommendation results and neglect the diversity and scalability issues. In fact, traditional recommendation techniques (collaborative filtering techniques, content-based techniques) suffer from well known scalability and monotonicity issues. In this work, we have proposed an elegant approach to generate recommendations diversified by different categories, using the association rule mining based CF approach. Work has been done in the area of ARM based CF technique, but the rules are generated on items, which turns out to be inefficient when the product space is growing rapidly. Therefore, instead of generating associations among the items, which are highly dynamic in nature, we have generated associations among the categories and these rules are later used to extend the user preference vector for the categories. To evaluate this method, we have experimented with a real world data (mobile application user data from the iTunes app store). Experimental results yield good accuracy (15% increase in precision) while pertaining diversity (91% inter-list diversity) in the recommendation list in a scalable fashion (quasi-linear increase of response time with an increase of user-base).

- In study 2, we have investigated the word-of-mouth in the context of social networks like Twitter. On Twitter, while most of the tweets go into oblivion, only a few of them get massive user attention and are retweeted extensively. Here lies the evident question, "what makes a tweet retweeted widely". Prior researches have been conducted to unfold the factors affecting retweetability using content features (hashtags, URLs, etc.) of tweets along with indegree (number of followers) of a user. However, indegree of a user does not reflect the real contribution of the user in the information dissemination process. This prompts us to characterize user roles based on their impact on information diffusion and investigate the significance of user roles in the retweet phenomenon. To study information propagation through retweets, one needs to build a retweet network[1], which captures interaction among the users through retweeting. Earlier investigations have constructed retweet network using only the tweet content (i.e., observing the citations in the tweet), which suffers from several biases due to unstandardized retweet practices. Users can retweet using the official retweet button or they can simply copy and paste the original tweet and post. Users tend to keep only the original author of the tweet, and not intermediates, in particular to meet the 140 character limit of Twitter. Even when using the official retweet function of Twitter, only the initial poster is kept. As information flows on Twitter through the cascades of followers, bias in the constructed retweet network from citation information in a tweet can be avoided by imposing the follower network[2] information.

We have combined both activity and follower networks and introduced the concept of *Information Diffusion Impact (IDI)* of users on network to characterize important user roles in information propagation to investigate their importance in the retweet phenomena. Further, we have studied whether an emergency event has any significant impact on these factors. With a Twitter dataset during the Great Eastern Japan Earthquake (11^{th} March, 2011), we first classified users using *IDI* into three important roles, namely, idea-starter, amplifier, and transmitter.

[1] Retweet network is an *interaction graph* which captures who is retweeting whom on Twitter

[2] Follower network is the directed graph where each node represents a user and links between them represent relationships. This allows users to follow people of their interests without requiring them to reciprocate. However, this network cannot capture the social interaction among the users.

Next, retweet model has been studied to understand the importance of these roles in retweetability. Further, the effect of the earthquake on the factors affecting retweetability has been investigated. Results indicate that *amplifiers* and *information-starters* affect retweetability significantly and due to an event these effects change substantially. We have also replicated the investigation in another dataset of the Boston marathon bomb blast of 15th April, 2013. The results obtained from the Boston marathon bomb-blast data reestablish our findings from the Japan earthquake data.

- In study 3, we investigate the evolution of hashtags. On Twitter, certain hashtags gain a lot of popularity while most of the hashtags are used by only a few people. On a close observation on hashtags appearing in tweets, one can note that hashtags usually appear in groups. The reason users use more than one hashtag in a tweet might be manifold; however, the outcome of such practice increases the discoverability of the tweet (in Twitter search results all the hashtags in the tweet will contribute to the discoverability of the tweet) as well as the popularity of all the hashtags. While earlier researches have already focused on popularity prediction using hashtag contents and the graph structure of the network, co-appearance of hashtags are not taken care of. This study investigates the effect of co-appearing hashtags on hashtag's popularity.

  Prior literature suggests that preference of particular information depends on the ease of recalling and processing the information. For instance, a word that is hard to pronounce is perceived as risky (Song and Schwarz, 2008). Information that is unfamiliar or dissimilar increases the metacognitive difficulty in processing. Our findings support this in the context of the hashtag, which implies that when a hashtag appears with dissimilar hashtags, popularity decreases. Nevertheless, when dissimilar hashtags appear with URL, interestingly, its popularity increases. This phenomenon can be explained by the fact that the introduction of additional information spurs uniqueness and surprisingness of the hashtag, resulting in an increase of its popularity. Moreover, the investigation of the model in three different time-windows centering around an event reveals that at the time of the event, the effect of the similarity of hashtags is much stronger compared to pre- and post-event time windows. Interestingly, interaction plots show that the presence of URLs with similar hashtags does not have significant impact. It will facilitate in the policy making for the brand-advertisers while launching a new product in the market. The practical contribution of the study lies in strategic decision making for using hashtags for branding or advertising. Dissimilar hashtags with extra information like URL can enhance the attractiveness and uniqueness of a tweet, which is the key to getting it retweeted to a broad audience. In addition, the event-centric analysis of the hashtag popularity model suggests that this property of hashtags is much important in the time of the event, which can assist the government agencies to create emergency hashtags in tweets in a more receptive way.

1.3 Overview

The remainder of the thesis is structured as follows:

In chapter 2, we have investigated the role of electronic word-of-mouth in the context of recommending products. We have proposed an elegant approach to generate recommendations diversified by different categories using the association rule mining based CF approach. Foremost, we have presented a brief introduction to the problem followed by related literature in product recommendations. Next, we discussed the proposed model for recommending mobile applications. After that, we have presented the analytical overview tackling the computational complexity of our algorithm and discussed the experimental results. Lastly, we summarized our findings.

In chapter 3, we have classified user roles in the context of information diffusion and investigated the change in user roles in the time of crisis (earthquake in this case). First, we have briefly introduced the problem in the light of prior research. Subsequently, we have classified the user roles followed by the dataset description. Following that, we have analyzed the dataset to investigate the evolving user roles at the time of crisis. Further, we have investigated the factors affecting retweetability and have analyzed the correlation of factors with that of the popularity of a tweet. First, we have described the problem in a nutshell, followed by the discussion of related literature. Next, we proposed our model. Further, we have investigated the effect of an earthquake in this regard and provided a brief summary of our findings at the end.

In chapter 4, we have investigated the factors impacting the popularity of hashtag.

First, we reviewed the related literature. Followed by that, we have described the dataset used in the study. After that, an overview of the solution details has been given and the probable factors affecting the popularity of hashtag are discussed. In the subsequent section, we have described the experimental details and the model proposed for measuring hashtag popularity. Finally, we summarized our findings.

Finally, in chapter 5, we have summarized the findings of these three studies and provided conclusion and future direction.

Chapter 2

Towards Generating Diverse Recommendation on Large Dynamically Growing Domain

2.1 Introduction

Recommendation technology has been around for a long time and is quite well understood. A review of the recommendation literature demonstrates its use in certain classes of products such as books (Linden et al., 2003), movies (Lekakos and Caravelas, 2008), music (Davidson et al., 2010), etc. Here arises the decisive question, would these traditional recommendation algorithms be applied to a new class of products - mobile apps, a domain of digital goods? The injection volume of this new class of products is in orders of magnitude higher than products like movies, books, etc. The domain of mobile applications has enormous growth of its number of apps (Tweney, 2013; Adam Lella, 2014; Perez, 2014). While on an average over 15,000 new apps are launched weekly, only 100 new movies and 250 new books are released worldwide (Datta et al., 2011) as per 2011 statistics, which has increased up to 32,5000 for mobile apps only in the iTunes app store (Costello, 2014). In fact, currently there are over 3 million apps on the Apple (1.2 million), Android (1.3 million), Blackberry, and Microsoft native app markets (Statistica, 2014). In addition, in these cases the number of app users also concomitantly grows in massive numbers (mobiForge, 2014). So the scale problem arises both from the volume of apps as well as app users. In the iTunes app store, a popular mobile app domain, it is possible to navigate the popular apps, so called 'hot apps', but it is still hard for the mobile app users to find their preferred apps manually from the extensive list of apps. For mobile app domain, existing recommendation mechanisms will take very long time to run and most likely to return the similar apps as being used by users. However, for mobile apps recommending exactly similar apps has less of a value. It is preferable to recommend apps that are similar but has different functionalities. For example, if a user already has a map app, it is not valuable to recommend another map app, rather other travel apps such as gas station finder or traffic prediction will be more useful. This study proposes a recommendation system that does exactly the same and is suitable for large item and user space like mobile apps. It addresses the issue of scalability and recommend a diverse set of apps without sacrificing other performance parameters such as precision and recall.

Among various existing approaches collaborative filtering technique (CF) continues to be most favoured, where items have been recommended considering either similar items rated by other users or items from users sharing similar rating pattern for different items. The main stream researches for generating "good recommendation" have been engaged to improve the accuracy of exact item prediction by reducing the Root Mean Square Error (RMSE) or the Mean Absolute Error (MAE). Recently, methods for non-monotonous predictions have also been addressed (Ziegler et al., 2004; Zhang and Hurley, 2008, 2009; Vargas and Castells, 2011). However, the issues of scalability, data sparseness (Sarwar et al., 2000), and association problems (Kim and Yum, 2011) remain vastly underdeveloped and are challenging till date. In fact, these general recommendation methods (e.g., user based CF, item based CF, and content- based technique) are quite computationally intensive and when new products or reviews come in, the system has to be re-run to factor in their effects.

Attempts have also been made to generate recommendations in the area of Association Rule (ARM) based CF techniques. Similar to traditional CF methods, application of ARM based CF techniques also turned out inefficient for the rapidly growing app space. We reasoned the failure of this approach arises due to generation of rules on items (mobile app) which are highly dynamic in nature. We anticipated that a promising solution of these issues could be a diminution in the cardinality of the large user-item rating matrix. Thus, instead of generating associations among the items (app), generation of associations among the categories, which is quasi-static in practice, could be a convenient route.

Our study tackles with the scalability issue of the recommendation algorithm of mobile apps while introducing diversity and maintaining an acceptable degree of accuracy. To address the problem of scalability, sparse user-item rating matrix[1] has been converted to denser user-category rating matrix[2]. The proposed framework for recommendation uses the co-liked categories by several users derived from user-category rating matrix, which inherently introduces diversity in the recommendation lists. To show the utility of our approach in practical scenario, we have implemented as well as experimented the algorithm using real world mobile application user data from Mobilewalla (Mobilewalla is a venture capital backed company which accumulates data for mobile applications from four major platforms Apple, Android, Windows, and Blackberry).

We have used user-based (*UCF*) and item-based (*ICF*) collaborative filtering technique and content-based recommendation technique (*CR*) as the baseline algorithms. The experimental results demonstrate the superiority of our approach over traditional CF techniques on most of the performance parameters (recall, diversity, and entropy) while not degrading the others (precision). Experimental results achieve good accuracy (15% increase in precision) while maintaining diversity (91% inter-list diversity) in the recommendation list in a scalable fashion (a quasi-linear increase in response time with a linear increase in user-base).

The rest of the chapter is organized as follows: next section discusses the brief overview of the related literature followed by the problem formulation and our pro-

[1] In user-item rating matrix, for each user-item pair, a value represents the degree of preference of that user for that item.

[2] In user-category rating matrix, for each user-category pair, a value represents the degree of preference of that user for items in that category.

posed approach. After presenting our empirical results, we summarized our findings.

2.2 Literature Review

An overwhelming increase in the amount of information over internet raise a requirement of personalized recommendation system for filtering the abundant information. The traditional recommender system predicts a list of recommendations based on two well-studied approaches, collaborative filtering and content-based techniques (Goldberg et al., 1992; Herlocker et al., 2004; Miller et al., 1997). 'Collaborative filtering' (CF) concept was pioneered by Goldberg et al. (1992) that uses the historical records of users' behaviour, either the items previously purchased or the numerical ratings provided by them. Similar users are mined and their known preferences are used to make recommendations or predictions of the unknown preferences for other users (Miller et al., 1997). There are several CF techniques known in literature which can be broadly classified into user based and item based CF technique (Herlocker et al., 2004). Though traditional CF techniques are adapted by many e-commerce portal, Amazon (Linden et al., 2003), YouTube (Davidson et al., 2010), and Netflix (Bennett et al., 2007), it has few fundamental drawbacks pointed out earlier and the most important one is scalability issue. For instance, Netflix was founded in 1997 and there are 50 million subscribers, 100,000 titles on DVD globally by 2014 (Wikipedia, 2014b). On the other hand, in the mobile app domain, iTunes app store was launched on 2008 and by 2014 there are 1 million apps, 150 million users who have provided reviews for apps (mobiForge, 2014). On average, every user has reviewed 3-4 app reviews. So we can enumerate the growth of the mobile app store compared to traditional items, which gives rise to the scalability issue.

CF technique is very much compute-intensive and the computational cost grows polynomially with the number of users and items in a system leaving the system ineffective in practice. Recently, attempts have been made by several research groups to improve the efficiency of collaborative filtering techniques in different domains. A detailed survey of recommendation algorithms can be found in Schubert et al. (2006). Takács et al. (2009) have employed Matrix Factorization method on Netflix dataset and showed that their method is scalable for large datasets. The efficiency of the method was also verified on MovieLens and Jester dataset. Koren (2010) introduced a new neighbourhood model with an improved accuracy on par with recent latent factor models, and it is more scalable than previous methods without compromising its accuracy. Several incremental CF algorithms are designed (Papagelis et al., 2005; Khoshneshin and Street, 2010; Yang et al., 2012b) to handle the scalability issue. Papagelis et al. (2005) proposed an incremental CF method which updates the user-to-user similarities incrementally and hence suitable for online application. Khoshneshin and Street (2010) proposed an evolutionary co-clustering technique that improves predictive performance while maintaining the scalability of co-clustering in the online phase. Yang et al. (2012b) have also proposed incremental item based CF technique for continuously changing data and insufficient neighbourhood problem is handled based on a graph-based representation of item similarity. However, the app growth is enormous and new apps and new users enter the app market very rapidly compared to other digital goods. Moreover, the existing approaches do not take care of diversity issue of recommendation. This is why the existing approaches cannot be applied to the app world directly. Moreover, unlike other digital commodities where recommender systems are available (e.g., Netflix and Amazon), the absence of any existing mobile app recommender system motivates us to delve into the platform.

Another drawback of CF technique is the data sparsity problem. Because of the fact that in practical scenario, most of the users rate only a few numbers of items, a very sparse user-item rating matrix is generated and the sparsity increases with the growth of item space resulting low accuracy of the system. Cross-domain mediation can be used to address the sparsity problem as well as to widen and diversify the recommendation list. In Li et al. (2009), sparsity problem is addressed by transferring a dense user-item rating matrix to target domain. The basic assumption here was that related domains (e.g., books and movies share similar genres) share similar rating patterns and hence can be transferred from one domain to target domain. Ziegler et al. (2004) have proposed a hybrid approach that exploits taxonomic information designed for exact product classification to address the product classification problem. They have constructed user profiles with a hierarchical taxonomic score for super and subtopic rather than an individual item. This method attempted to overcome the sparsity problem in CF techniques and contributed toward generating novel recommendations by topic diversification. However, because one item may be present in more than one super or sub topic, the structure became more complicated.

Ziegler et al. (2004) have proposed to diversify the topic and return items to the end user by topic diversification, but these generated recommendations are still from the same domain. Overspecialization in recommendation list refers to the problem of generating similar recommendations for a user which reduces the diversity. Jiang and Sun (2012) proposed a dynamic programming algorithm to address overspecialization in recommendation list and generate diverse and relevant recommendations. The algorithm uses a nested logit model of the item pool which is not scalable for large dynamically growing domain like mobile apps. Adomavicius and Kwon (2014) proposed a greedy maximization heuristic and graph-theoretic approach to improve diversity of recommendation list and experimented using Netflix and MovieLens dataset. Graph-theoretic (Huang and Zeng, 2011) and probabilistic cut-off model (Prawesh and Padmanabhan, 2014) have been presented to improve diversity in several domains.

Association rule (Agrawal and Srikant, 1994; Agrawal et al., 1993) mining technique has also been applied to CF for mining interesting rules for recommendation generation (Kim and Yum, 2011; Sarwar et al., 2000). The top-N items are generated by simply choosing all the association rules that meet the predefined thresholds for support and confidence, and the rules having higher confidence value (sorted and top N items are chosen finally) have been selected as the recommended items. To address data sparseness and non-transitive associations Leung et al. (2006) proposed a collaborative filtering framework using fuzzy association rules and multilevel similarity.

In all these studies, the authors attempted to determine the associations among the items and the consequent items in the rules are the candidates for recommendations. In contrast, we have used the association rules to find the association pattern in the categories chosen by the users. Since the rules are generated offline on less dynamic categories, it does not add to computational complexity.

2.3 Solution Intuition

Recommendation generation is a single step process that works on item set and user set. However, since both users and items are large and dynamically growing in the mobile app domain, generating scalable, accurate, and diverse recommendation becomes challenging. In this research, we use the following process where in the first step we focus on generating association rules on categories of items rather than the item itself, which reduces the scalability issue significantly because the number of categories is far less than the number of items. We determine the category affinity vector of all users and using the association rules on categories, users' category affinity vectors are updated. We create the user profiles based on their item and category affinity vector information, which we term as item feature and category feature respectively. Next, items are recommended from similar user's item list by computing the similarity score of user pairs, that comprises of two features, category feature and item feature.

With the knowledge of item preference of the users, one can easily derive the category preference vector of those users. From a big population, we first generate the category preference vectors of the users and use them to generate association rules among the categories. For example, if 80% of the users using *'travel'* and *'vehicle'* apps also use *'maps and navigation'* apps, then a rule $travel, vehicle \rightarrow maps\ and\ navigation$ is generated with confidence value 0.8. Now suppose we have to generate recommendations for a new user who has *travel* and *vehicle* apps, our recommendation algorithm will generate recommendations from all three categories, *travel*, *vehicle*, and *maps and navigation* using the above rule. In this way diversity of recommendation is achieved inherently. The rule generation process is done offline (as the categories are quasi-static in practice) and the recommendation generation process is done online.

On the other hand, item feature has been considered to maintain the relevance of the output. Here, we consider the semantic similarity of user's items with that of the recommended items. Using an existing semantic similarity measure, the similarity of two users is calculated as the similarity of focal user's itemset with other users' itemset. This feature will take care that though the recommended items are diversified, they are also semantically similar with the user's current item list.

2.4 Dataset Description

User-based (*UCF*) and item-based (*ICF*) collaborative algorithm and content-based (*CR*) recommendation technique have been used as the baseline algorithm. The experiment has been conducted with a real world data of mobile app users' reviews as a surrogate of installed apps on user's mobile phone. A sample of user review of Apple app users and the corresponding app information has been collected from Mobilewalla[1], which contains the following dimensions depicted in Table 2.1. A total of 1744811 users' information has been collected, out of which only 22213 users who have rated more than 5 apps are considered in this study in order to sample the user-space under experiment.

Table 2.1: App and User Details

App and User Details
iTunes ID of the author (unique)
Application ID
Name of the app
Description of the app
Category of the app

Table 2.2: Descriptive Statistics

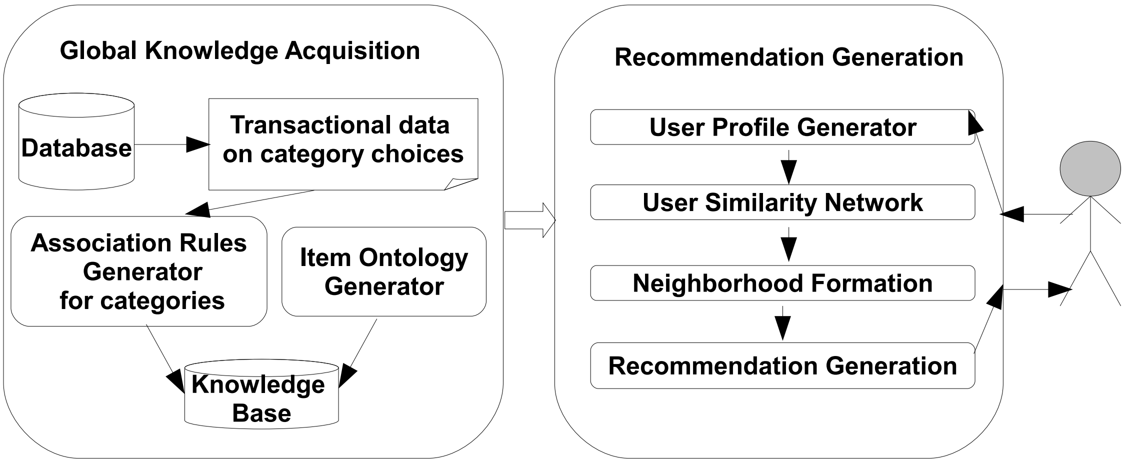
Variables	**Numbers**
Total Number of Users	22213
Total Number of Products	66137
Average Number of Products Rated per User	3
Total Number of Categories	194

2.5 Solution Details

Our proposed system has two main components: (a) Global Knowledge Acquisition Module (GKA) and (b) Recommendation Engine (RE) (See Figure 2.1). Prior one is done offline while the later one is an online process. At a high level, GKA identifies the categories the user has an interest in and also pre-computes the item-item similarity

[1] Mobilewalla is a company for mobile app search http://mobilewalla.com/, which accumulates data for mobile applications from four major platforms that includes Apple and Android

Figure 2.1: Recommendation Architecture


Global Knowledge Acquisition
Database
Transactional data on category choices
Association Rules Generator for categories
Item Ontology Generator
Knowledge Base
Recommendation Generation
User Profile Generator
User Similarity Network
Neighborhood Formation
Recommendation Generation

based on meta-data information about the items. The online component, RE operates on the output of GKA, i.e., the association rules of the categories and item-item similarity matrix to create a profile vector for each user. The generated profile vector is then used to compute the similarity across users and recommend new items accordingly. Next, we describe the details of GKA and RE. The notations used are shown in Table 2.3.

ALGORITHM 1: Build Category Interest

```
Input: Items used by user $u_i$ and a taxonomy of items for $\bar{d}$ categories
Output: Category Interest vector $C^{\bar{d}}_{u_i}$ of $u_i$
for $l = 1; l <= \bar{d}; l++$ do
    Initialize $r^{u_i}_{d_l} \leftarrow 0$
    $r^{u_i}_{d_l} \leftarrow r^{u_i}_{d_l} + \sum_{I_{ij} \in d_l} r^{u_i}_{I_{ij}}$
    $C_{u_i}[l] \leftarrow r^{u_i}_{d_l}$ //update score at $l^{th}$ position
end
return $C^{\bar{d}}_{u_i}$
```

Table 2.3: Notation Table

Notation	Meaning
$U = \{u_1, u_2, ..., u_n\}$	set of n users
$I = \{I_1, I_2, ..., I_l\}$	set of l items in the itemspace
$D = \{d_1, d_2, ..., d_m\}$	set of m categories items belong to
I_{u_i}	items perceived by user u_i, where $u_i \in U$
$I_{u_i}(d_k)$	items perceived by user u_i from category d_k
D_{u_i}	category set of items perceived by u_i
$\{C^d_{u_i}\}$	category interest vector of user u_i of dimension $d = \|D\|$
$r^{u_i}_{d_l}$	rating of u_i for category d_l
Sup	support threshold
$Conf$	confidence of an association rule
$Score_{Category}(u_i, u_j)$	category score of user u_i and u_j
$Score_{Item}(u_i, u_j)$	item similarity score of user u_i and u_j
$Sim(I_i, I_j)$	semantic similarity of item I_i and I_j
β_{th}	similarity threshold of binary precision and recall
ν_{th}	similarity threshold of fuzzy precision and recall

2.5.1 Global Knowledge Acquisition Module (GKA)

Input to GKA is the meta-data (name, description, categories) of the existing items that users have used. Symbolically, if I_{u_i} is the set of items user u_i has used such that $I_{u_i} \subset I$, the input to GKA are $I_{u_i} \forall u_i \in U$. The GKA consists of two main sub-components, category association rule generator and item-item similarity generator.

Generating Transactional Data on Category Choices: The goal of this task is to transfer this sparse user-item matrix to denser user-category matrix. Thus, each record in the new matrix corresponds to the transactional information on category for a user (See Algorithm 1). Consider, for each of n number of users, we have the set of items I_{u_i} used by u_i and a taxonomy of the $\bar{d}$ ($\bar{d} = |D|$) categories as input to the Algorithm 1. Initially, the category score for $d_l = 0$ (Algorithm 1, line 2). For each category d_l in category space

Table 2.4: Abbreviation Table

Abbreviation	**Meaning**
Sup	Support threshold
Conf	Confidence threshold for association rules
ARM	Association Rule Mining
GKA	Global Knowledge Acquisition
CF	Collaborative Filtering
UCF	User based collaborative filtering
ICF	Item based collaborative Filtering
CR CPU	Content based recommendation Central Processing Unit
RAM	Random Access Memory

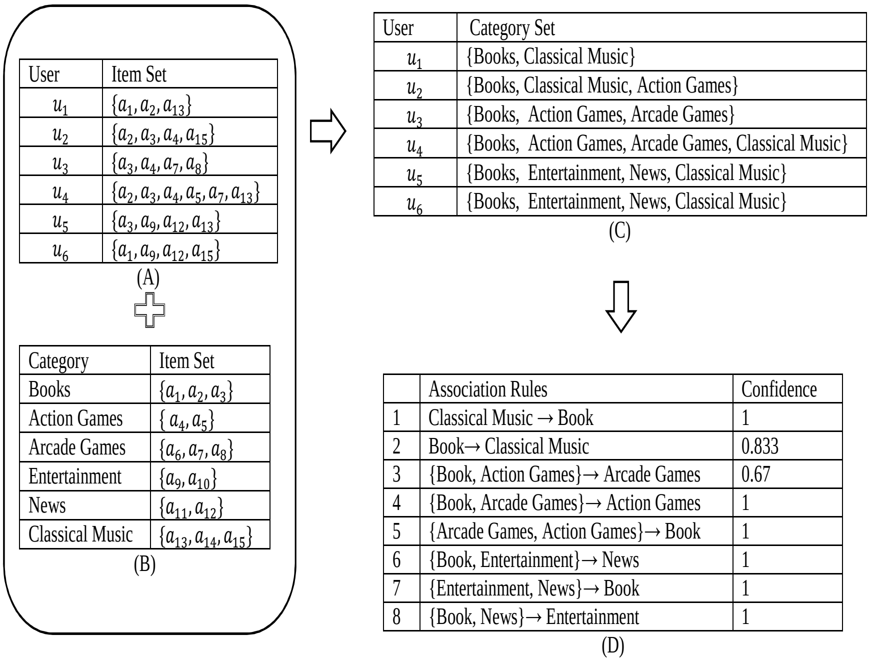
D, we sum up the rating of u_i for category d_l (Algorithm 1, line 3). Using this algorithm we derive a set of categories $C_{u_i}^{\bar{d}}$ used by user u_i of dimension $\bar{d}$ using the item-category mapping from I_{u_i}.

Association Rule Generator for Categories: In this work we have employed association rule mining (ARM) on the transactional data on categories to find the associations of different categories. Analogous to 'Market Basket Analysis', we identify the usage pattern in various categories simply by finding the 'togetherness' of the categories in the data with a support and confidence value chosen experimentally. The calculated confidence for each rule is used as the *score of closeness* of the categories. To illustrate, if a user likes a *Travel* application, he might be interested in *Restaurant* applications in that area.

It is worth to emphasize that in a practical scenario mobile app-space is more dynamic in nature compared to category taxonomy. As a result, frequent re-evaluation of the rule set in ARM based CF on items is inevitable and retards the system proficiency.

Item-Item Similarity Generator: In item-based CF techniques, the similarities among the items are computed by exploiting the similar rating pattern by the users. In contrast, in this work semantic similarity has been pre-computed for an item pair using item-information meta-data, *"Info = Description, Name"*, i.e., the description of the item and the name of the item. Apache Lucene (Apache, 2001) is used to first index the items and then compute the item-item similarity score based on Cosine similarity. It is independent of the previous module and hence can be performed in parallel.

Figure 2.2: Association Rule Generation Process



User	Item Set
u_1	$\{a_1, a_2, a_{13}\}$
u_2	$\{a_2, a_3, a_4, a_{15}\}$
u_3	$\{a_3, a_4, a_7, a_8\}$
u_4	$\{a_2, a_3, a_4, a_5, a_7, a_{13}\}$
u_5	$\{a_3, a_9, a_{12}, a_{13}\}$
u_6	$\{a_1, a_9, a_{12}, a_{15}\}$

(A)

Category	Item Set
Books	$\{a_1, a_2, a_3\}$
Action Games	$\{ a_4, a_5\}$
Arcade Games	$\{a_6, a_7, a_8\}$
Entertainment	$\{a_9, a_{10}\}$
News	$\{a_{11}, a_{12}\}$
Classical Music	$\{a_{13}, a_{14}, a_{15}\}$

(B)

User	Category Set
u_1	{Books, Classical Music}
u_2	{Books, Classical Music, Action Games}
u_3	{Books, Action Games, Arcade Games}
u_4	{Books, Action Games, Arcade Games, Classical Music}
u_5	{Books, Entertainment, News, Classical Music}
u_6	{Books, Entertainment, News, Classical Music}

(C)

	Association Rules	Confidence
1	Classical Music → Book	1
2	Book→ Classical Music	0.833
3	{Book, Action Games}→ Arcade Games	0.67
4	{Book, Arcade Games}→ Action Games	1
5	{Arcade Games, Action Games}→ Book	1
6	{Book, Entertainment}→ News	1
7	{Entertainment, News}→ Book	1
8	{Book, News}→ Entertainment	1

(D)

Let us assume a dummy example shown in Figure 2.2. Say, there are 15 mobile apps, $\{a_1, a_2, ..., a_{15}\}$ are available in the iTunes store from 6 different categories, namely 'Book', 'Action Games', 'Arcade Games', 'Entertainment', 'News', and 'Classical Music' and 6 users $\{u_1, u_2, ..., u_6\}$ have used those 15 apps as shown in Figure. 2(A). Say, $\{a_1, a_2, a_3\} \in$ *Books*, $\{a_4, a_5\} \in$ *Action Games*, $\{a_9, a_{10}\} \in$ *Entertainment*, $\{a_{11}, a_{12}\} \in$ *News*, and $\{a_{13}, a_{14}, a_{15}\} \in$ *Classical Music* (See Figure 2.2(B)). Figure 2.2(C) shows the category mapping of the users from the item-space. From the dataset total of 8 association rules are mined (Figure 2.2(D)) using minimum support = 0.2 and minimum confidence = 0.65. Each of these rules is of the form $[rule.antecedent \rightarrow rule.conseqeunt, confidence]$.

Additionally, to generate item-item similarity descriptions and names of these 15 mobile apps are crawled and indexed using the Lucene indexer. With these two information meta-data, Lucene similarity score has been calculated among these apps. So at most 210 app-pairs will have similarity scores, which is then stored in a knowledge base. In practice, very few app-pairs will have non-zero similarity scores.

Once the offline processes of generating association rules and item-item similarity computation are done, they are fed to the recommendation system through the central knowledge base which comprises of the association rules on categories and item-item similarity score. This information is accessed each time recommendations are generated for a user.

2.5.2 Recommendation Generation Module

The recommendation generation module which is core to generate online recommendations consists of 4 sub-modules. The first step is to generate the profile for the users using pre-computed category association rules and item-item similarity matrix. Afterwards the generated user profiles are updated in the knowledge database. Next, the

ALGORITHM 2: Inject Association Rules

```
Input: Category Interest vector $C_{u_i}^{\bar{d}}$ of $u_i$ and set of Association rules $R$
Output: Updated Category Interest vector $C_{u_i}^{\bar{d}}$ of $u_i$
for $rule \in R$ do
    Initialize $r \leftarrow 0$
    Initialize $A \leftarrow rule.antecedent$ //get the antecedent part of the rule
    Initialize $flag \leftarrow true$
    forall the $d_l \in A$ do
        //Check if $u_i$ owns at least one item from each categories in antecedent part
        if $C_{u_i}[l] == 0$ then
            // $u_i$ does not own any item from category $d_l$
            $flag \leftarrow false$
            break
        end
        else
            $r \leftarrow r + C_{u_i}[l]$ //fraction of items for the categories
        end
    end
    if $flag == true$ then
        Initialize $P \leftarrow rule.precedents$
        Initialize $Conf \leftarrow rule.confidence$
        forall the $d_k \in P$ do
            $C_{u_i}[k] \leftarrow Conf \times \frac{r}{|A|}$
        end
    end
end
return $C_{u_i}^{\bar{d}}$ //return updated category interest
```

neighbourhood of the active user u_i is formed. Finally, the recommendations are generated from the u_i's top N-similar users' item list. Next follows the detailed discussion of these four steps.

User Profile Generator

User profile comprises of two features, namely category affinity vector and item feature. Below, we define the category score and the category affinity vector for a user.

Definition: (Category Score)

For a user u_j category score for category d_k is the fraction of items u_j own from category d_k.

$Score_{u_j}^{d_k} = \frac{|I_{u_j}(d_k)|}{|I_{u_j}|}$

If the user does not have any item from category d_k, then $Score_{u_j}^{d_k} = 0$.

Definition: (Category Affinity Vector)

For n categories $\{d_1, d_2, ..., d_n\}$ for user u_j, category affinity vector is an n dimensional vector with each entity being the category score defined as above, i.e.,

$C_{u_j}^d = \{Score_{u_j}^{d_1}, Score_{u_j}^{d_2}, ..., Score_{u_j}^{d_n}\}$

For a user, the category affinity vector defines the preference of the user over different categories in the application domain.

Say, user u_2 has 4 mobile apps $\{a_2, a_3, a_4, a_{15}\}$ installed in his cell phone (Figure 2.2A), 2 from 'Books' category and 1 from 'Action Games' category, then initial $C_{u_2}^d = [0.5, 0.25, 0, 0, 0, 0]$ (Table 2.6).

Updating category affinity vector: Once initial $C_{u_j}^d$ is calculated for each user, generated association rules are injected to update $C_{u_j}^d$. If the user has expressed interest in categories in the antecedent part of the rule, then the categories in the consequent part of the rule are updated with an average score of the antecedent categories weighted by the confidence of the rule (See Algorithm 2, lines 16-21).

From the dataset, 8 association rules are mined as mentioned earlier (Figure 2.2(D)). Extending the previous example, 'Arcade Games' is added to $C_{u_2}^d$ with score (0.5 +

0.25)/2×0.67 = 0.25125 (using association rule 'Book', 'Action Games' → 'Arcade Games', 0.67) respectively. Thus, the category affinity vector reduces to $[0.5, 0.25, 0.25125, 0, 0, 0]$. Next, the vector is normalized to unity. Note that, the other rules were skipped as u_j does not own 'Entertainment' or 'News' apps resulting category affinity vector as $[0.1998, 0.1998, 0.3996, 0, 0, 0]$. Similarly, the category vectors are updated for other users (See Table 2.5).

Item Feature: Once the category affinity vector is calculated for the users, items set I_{u_j} for a user is added to the profile to find the semantic similarity of items in I_{u_j} with $I_{u_k} \forall I_{u_k} \notin I_{u_j}$ in later phase. Continuing the same example, for user u_2, references of 4 apps $\{a_2, a_3, a_4, a_{15}\}$ installed in his cell phone are added to his profile.

For the existing users, profile generation can be done offline and stored in the database, whereas for the new users, the generated profile can be updated for future reference regularly. It is worthy to note that association rules do not require to be generated regularly as association rules are based on categories, those are less dynamic in nature.

Table 2.5: User Profile Generation

	Initial Category Affinity Vector						User Profile						
							Updated Category Affinity Vector (After Normalization)						
	Books	Action Games	Arcade Games	Entertainment	News	Classical Music	Books	Action Games	Arcade Games	Entertainment	News	Classical Music	Item Feature
u_1	0.667	0	0	0	0	0.333	0.667	0	0	0	0	0.333	$\{a_1, a_2, a_{13}\}$
u_2	0.5	0.25	0	0	0	0.25	0.399	0.199	0.201	0	0	0.199	$\{a_2, a_3, a_4, a_{15}\}$
u_3	0.25	0.25	0.5	0	0	0	0.207	0.207	0.414	0	0	0.208	$\{a_3, a_4, a_7, a_8\}$
u_4	0.333	0.333	0.167	0	0	0.167	0.333	0.333	0.167	0	0	0.167	$\{a_2, a_3, a_4, a_5, a_7, a_{13}\}$
u_5	0.25	0.25	0	0	0.25	0.25	0.25	0	0	0.25	0.25	0.25	$\{a_3, a_9, a_{12}, a_{13}\}$
u_6	0.25	0	0	0.25	0.25	0.25	0.25	0	0	0.25	0.25	0.25	$\{a_1, a_9, a_{12}, a_{15}\}$

Neighbourhood Formation

For an active user u_i, we need to find the similar peers using the well-known proximity measure described below:

Proximity Measurement For similarity computation there are several measures exist (e.g. Cosine similarity, Pearson correlation, Spearman correlation etc.) in the literature, however Pearson correlation has been used widely. While Cosine similarity can be interpreted as the cosine of the angle between two vectors, Pearson correlation can be interpreted as the demeaned Cosine similarity. However, Pearson correlation is invariant to shift of the vector element, which means if x is shifted to $x + 1$ the Pearson correlation will not change. Spearman correlation is used when we want to measure similarity between ranked vectors. However, in our study we need to find similarity among non-ranked user profile vectors, where Pearson correlation is used extensively. Pearson correlation can be defined as follows:

$$r(u_i, u_j) = \frac{\sum_{k=1}^{n}(v_{ik} - \bar{v}_i)(v_{jk} - \bar{v}_j)}{\sqrt{\sum_{k=1}^{n}(v_{ik} - \bar{v}_i)^2 \cdot (v_{jk} - \bar{v}_j)^2}}, \tag{2.5.1}$$

For two different features the similarity score is calculated.

- Category Feature: For user pair (u_i, u_j), Pearson correlation has been computed between $C_{u_i}^d$ and $C_{u_j}^d$ using equation Eq.2.5.1 and is denoted by $Score_{Category}$. In Table 2.6, for each pair of users, the category score has been calculated using Eq.2.5.1.

Table 2.6: Calculation of Category Score

$Score_{Category}(u_i, u_j)$	u_1	u_2	u_3	u_4	u_5	u_6
u_1	1	0.00029	0.133	0.365	0.463	0.463
u_2		1	0.9096	0.6653	-0.7906	-0.7906
u_3			1	0.6652	-0.7905	-0.7905
u_4				1	-0.5259	-0.5259
u_5					1	1
u_6						1

- Item Feature: For u_i and u_j, item similarity score ($Score_{Item}$) has been computed as the average semantic similarity of the pair of item set I_{u_i} and I_{u_j} normalized by $|I_{u_i}| \cdot |I_{u_j}|$. Here, by semantic similarity of two sets we mean semantic similarity of the elements of the two sets.

$$Score_{Item}(u_i, u_j) = \frac{1}{|I_{u_i}| \cdot |I_{u_j}|} \sum_{p \in I_{u_i}, q \in I_{u_j} p \neq q} Sim(p, q), \tag{2.5.2}$$

With the running example $Score_{Item}$ of u_2 with all the other users are calculated using Eq.2.5.2 as shown in Table 2.7.

Further, a weighted score for these two features has been calculated and the final score is computed as $Score_{u_i,u_j} = w_1 \cdot Score_{Category}(u_i, u_j) + w_2 \cdot Score_{Item}(u_i, u_j)$,

w_1 addresses that two users have similar categories, while w_2 addresses that two users have similar items. w_1 introduces diversity in recommended items. We gave higher weight to w_1 compared to w_2 to get more diverse recommendations (w_1 and w_2 are decided experimentally). In our study, we have used $w_1 = 0.6$ and $w_2 = 0.4$.

With u_2's category affinity vector, Pearson correlation has been computed with all the other users' $(u_1, u_3, u_4, u_5, u_6)$ category affinity vectors which comes out to be $0.00029, 0.9096, 0.6653, -0.7906, -0.7906$ respectively. Assume $w_1 = 0.6$ and $w_2 = 0.4$ and calculate the score for each pair of users.

Score Vector: Score vector of user u_i is defined as the vector containing all score values with other users, $Score(u_i) = [Score_{u_i,u_j}]_{\forall j, j \neq i}$.

Table 2.7: Calculation of Item Score

$Score$	u_1	u_3	u_4	u_5	u_6
$Score_{Category}(u_i, u_2)$	0.00029	0.9096	0.6653	-0.7906	-0.7906
$Score_{Item}(u_i, u_2)$	0.7	0.7	0.9	0.8	0.8
$Score(u_i, u_2)$	0.280174	0.82576	0.75918	-0.15436	-0.15436

Neighbour Selection To find the top neighbours, two approaches can be used: by either setting a threshold value above which the peers are considered as similar users, or selecting top-N users similar to active user u_2. Setting a threshold value for the similarity score will give neighbors based on the value chosen. In this case we might end up getting no neighbors for some users. However, selecting top-N users does not depend on the value of similarity threshold, rather it depends on N. This process will result in a non-empty neighborhood set in maximum cases. Therefore, in this work, we have chosen the top-N users as neighbours.

For user u_2 if we take top-1 neighbour, then u_3 becomes the selected one (similarity score = 0.82576).

Recommendation Generation

Once the neighbourhood is generated from the set of users, recommendations are generated from the items of the top-N users' list (See Algorithm 3). The input to the algorithm is the set of top N users $(u_1, u_2, ..u_N)$ with score vector $[s_1, s_2, ..s_N]$. For each of u_j, set of item = I_{u_j}. Say, for a user u_i we want to generate recommendation R_i. From the top-N neighbours, we find the items those are not in u_2's item list (Algorithm 3, line 5). These items are assigned the similar users' similarity value as the score (Algorithm 3, lines 7-10). If one item is recommended from many users in top-N user list, then the user's score is added up to assign a higher score to that item (Algorithm 3, lines 11-13). Finally the top-k items have been recommended to the active user (Algorithm 3, lines 17-22). Recommendations are generated from u_3's item list where only two apps (a_7, a_8) can be recommended to user u_2 as he owns all the other apps from u_3's list. Since $Score(u_2, u_3) = 0.82576$ (Table 2.7), both the items a_7 and a_8 are assigned score as 0.82576 (Algorithm 3, lines 4-15).

2.6 Analytical Overview

In this section we calculate the computational complexity of our algorithm.

The goal of this study is to produce diverse recommendations for a dynamically growing dataset. For large datasets the main concern lies in the scalability of the system. We measure the scalability of our algorithm by computing the computational complexity of our algorithm.

ALGORITHM 3: Recommendation Generation

```
Input: top N similar users $U_N$ with score vector $[s_1, s_2, ..s_N]$, item set $\{I_{u_1}, I_{u_2}, ..I_{u_N}\}$
        respectively for each of $u_k \in U_N$
Output: top-K Recommendations list $R_i$ for $u_i$
Initialize $R_i \leftarrow \emptyset$
Initialize $List.Item \leftarrow \emptyset, List.Score \leftarrow \emptyset$
Initialize $p \leftarrow 1$
for $j = 1; j <= N; j++$ do
    Initialize $TemList \leftarrow I_{u_j} \cap I_{u_i}$
    for $I \in TemList$ do
        if $I \notin List.Item$ then
            $List.Item[p] \leftarrow I$
            $List.Score[p] \leftarrow s_j$
        end
        else
            $List.Score[p] \leftarrow List.Score[p] + s_j$
        end
        $p \leftarrow p + 1$
    end
end
if $List.Item \neq \emptyset$ then
    Sort $List.Item$
    for $p = 1; p <= K; p++$ do
        $R_i \leftarrow R_i \cup List.Item[p]$
    end
end
return $R_i$ //return generated recommended list for user $u_i$
```

Computational Complexity Association rule mining for categories and item-item similarity is done offline and ahead of recommendation generation and hence does not add to the computational cost of generating recommendations for users. We will discuss the complexity of generating recommendation for each user in the online process.

From the whole dataset, say, R number of rules are generated, wherein the antecedent part of the rule, p number of categories are there ($Max \;\; p \leq 10$, in practice). Assume, user u has n items from d categories. We will calculate the average case complexity for each of the steps in the recommendation module.

Step 1: Building Category Affinity: First the item-category mapping is loaded from the database for focal user, which is $O(n)$ database fetch. Initial category affinity vector calculation takes $O(d \cdot (n/d)) + O(n) \approx O(n)$.

Step 2: Updating Category Affinity Vector: For each rule $\{d_1, d_2, ..., d_p\} \rightarrow d_m$, u's category list C is checked with maximum of the p number of categories. In the worst case, if only the last category in the rule $d_p \notin C$, then p comparisons are made. If the first category in the rule $d_1 \notin C$, then there is no need to check for the other categories. Thus, on average, there are $0.5p$ number of categories needed to be checked before moving for the next rule. On the other hand, if the rule is satisfied, then exactly p comparisons are made. Assume for a user, m% of R rules satisfying all the p categories in the antecedent part of the rule, are present in u's category list C.

Overall, for the rules where user's profile is not updated, the number of comparisons $= (1 - m) \times R \times 0.5 \cdot p$. For the rules where u's profile is updated, also the consequent part of the rule is verified in $O(1)$ time. Thus, overall time in step 2 $= O((1 - m) \times R \times$

$0.5 \cdot p + m \times R \times p) + O(1)) \approx O(0.5 \cdot R \times p \times (1 + m))$.

Step 3: Neighbor Formation: For each user, we need to find the neighbor users. Say, there is a total of U number of users in the application domain and total number of categories is D. Thus, for each user the category affinity vector is of dimension D. To compute the Pearson correlation for each user it takes $O(c \cdot D)$ steps, as sample Pearson correlation computation takes one pass algorithm. Ideally, for each user we need to compare with remaining $(U - 1)$ users. To reduce this number, the user profiles are indexed with the categories they prefer. To get the neighbor users for user u, only the user profiles with category list C is searched to prune the user space. Since the choice of categories for a user is independent of choice of categories for a different user, at most a user has to be compared with all other users. However, in practice one user does not have all the categories and two users having a set of same categories (or even a high percentage (say, 80%) similar) is very low. While computing the association rules say support (Sup) was 2% and confidence ($Min\ Conf$) threshold was 70%. If we consider these thresholds, then on average for each user we need to compare with at most $Sup \times U$ users, which reduces the overall number of users drastically. Let u_s be the number of users to compare. Then for u_s two vector comparison takes $c \cdot u_s \times D$ steps.

Also to compute the semantic similarity of the n items (say, on average every user has n items) complexity is $O(u_s \cdot n^2)$ (for u_s users). After computing the similarity score of a user with its u_s similar users, similarity score is sorted and top N users are selected. Overall, this step takes $O(u_s \cdot n^2) + O(c \cdot u_s \times D) + O(u_s \cdot log u_s)$.

Step 4: Recommendation Generation: To find the top K items from top N users, the similarity score of users is given to each item. As each user has on average n items, total nN items from N users are sorted in $O(nNlognN)$ and top K items are chosen.

So total complexity reduces to $O(n) + O(0.5 \cdot R \times p \times m) + O(u_s \cdot n^2) + O(c \cdot u_s \times D) + O(u_s \cdot logu_s) + O(nNlognN)$.

The average number of categories in antecedent part of a rule p, and percentage of rules satisfying the p categories and m $(0 < m < 1)$ are all constant. Similarly, number of top users to select, N and number of domains D, are small numbers.

Hence the complexity reduces to $O(n) + O(c.R) + O(c \cdot u_s) + O(u_s \cdot n^2) + O(u_s \cdot logu_s) \approx O(u_s \cdot n^2) + O(c.R) + O(u_s \cdot logu_s)$, i.e., the computational cost is quadratic in terms of the number of items a user owns, linear in terms of rules to be checked and quasi-linear in terms of the number of similar users. In the worst case $u_s = Sup \cdot U$, i.e., a fraction of total number of users in the domain. We list the complexity computation as follows:

Deduction 1: The computational complexity of online recommendation is quadratic in term of number of items a user owns.

Deduction 2: The computation complexity of online recommendation is linear in terms of number of association rules.

Deduction 3: The computation complexity of online recommendation is quasi-linear in terms of the number of similar users.

2.7 Experimental Results

In this section we will discuss the experimental results evaluating the accuracy, scalability, diversity, and entropy of the recommendation system. Experimental settings are described first, followed by the findings of the proposed algorithm.

2.7.1 Experimental Settings

All modules are implemented in Java 8 and MySQL v5.1 was employed as a database back-end. All modules and the database reside on the same computer (a server equipped with a 2.33 GHz quad-core CPU and 8 GB RAM, and running on Linux operating system).

Baseline Algorithms Our proposed algorithm has been compared with three traditional collaborative filtering techniques Item-based CF (*ICF*), User-based CF (*UCF*) and Content-based technique (*CR*).

2.7.2 Data Acquisition

Acquiring the real mobile app usage data is hard in a practical scenario. Thus, to evaluate the effectiveness of our method, review data from the mobile app users have been used as surrogate for usage data. In Table 2.2 descriptive statistics of the data are shown; on average there are 3 reviews per app. Each mobile application belongs to one or more categories. Now we will discuss the generation of association rules among these categories. With a support threshold of 0.1 and a confidence threshold of 0.7, total 72407 frequent item sets are mined generating 977678 association rules.

To find the similar apps first they are indexed using Lucene and later the similarity score (in [0,1]) has been calculated for each pair of apps[1]. Training (A_{tr}) (50%), and test dataset are chosen randomly (A_{ts}) (50%) where both the datasets contain mutually exclusive items as well as categories at user level. To illustrate, say a user has four apps $\{a_1, a_2, a_3, a_4\}$ from two categories, Games and Entertainment, where $\{a_1, a_2\} \in$ 'Games' and $\{a_3, a_4\} \in$ 'Entertainment'. Then we keep $\{a_1, a_2\}$ in A_{tr} and $\{a_3, a_4\}$ in A_{ts}. The experiment is conducted for 5000 users. Both collaborative filtering technique (user-based and item-based CF method) and content based method are used to compare the results with the proposed one. While recommendations are generated using the training dataset, the test dataset is used for evaluation.

2.7.3 Evaluation Metrics

In traditional recommendation systems, performance (precision and recall) is measured by calculating the exact match of the items in the test set with that of generated recommendations. While the exact match would be preferable, the similar predictions should not be overlooked. Thus, instead of evaluating the predictions against the exact item set, we have examined the closeness of user's actual taste and generated recommendations. To evaluate recommendation performance, we have each algorithm generate a ranked list of recommended items for each user and then the recommended items are compared with the actual transactions in test data. Measures used for evaluation are discussed in turns.

[1] Apache, Lucene, http://lucene.apache.org/core/

- **Binary Precision and Binary Recall:** In binary precision and recall, we assume if two items are semantically similar with respect to a predefined threshold (β_{th}), then the two items are same. "Binary Precision" and "Binary Recall" are formulated similar to the standard precision and recall. The only difference in our metrics with that of the standard ones is that two items are considered same when they are similar with a threshold β_{th}. We define "Binary Precision" and "Binary Recall" as follows:

  $Binary\ Precision(\beta_{th}) = \frac{|A_{ts} \cap A_o|}{|A_o|}$

  $Binary\ Recall(\beta_{th}) = \frac{|A_{ts} \cap A_o|}{|A_{ts}|}$,

  $I_i = I_j$ if $Sim(I_i, I_j) > \beta_{th}$, where $I_i \in A_o, \quad I_j \in A_{ts}$

  $Binary\ Precision(\beta_{th})$ depicts the fraction of the items in the recommendation list similar to the expected ones with a similarity threshold β_{th}. On the other hand, $Binary\ Recall(\beta_{th})$ explains the fraction of the items in the expected list similar to the recommenced ones for a similarity threshold β_{th}.

  If $\beta_{th} = 1$ binary precision and recall boils down to traditional precision and recall.

- **Fuzzy Precision and Fuzzy Recall:** Fuzzy precision and fuzzy recall (Bartosz Ziolko and Wilson, 2007) is defined by a membership function of an element (I_i) in a set A_k by the maximum similarity score of I_i with all the remaining elements in A_k. Similar to binary precision and recall, two items are considered same when they are similar with a threshold ν_{th}.

  $Fuzzy\ Precision(\nu_{th}) = \frac{\sum_{I_i \in A_o} f_{A_{ts}}^{\nu_{th}}(I_i)}{|A_o|}$

$Fuzzy\ Recall(\nu_{th}) = \frac{\sum_{I_i \in A_{ts}} f_{A_o}^{\nu_{th}}(I_i)}{|A_{ts}|}$,

where membership function

$$\begin{aligned} f_{A_k}^{\nu_{th}}(I_i) &= Max_{\forall I_j \in A_k} Sim(I_i, I_j),\ \ if\ \ Sim(I_i, I_j) > \nu_{th},\ \ where\ \ I_i \in A_o,\ \ I_j \in A_{ts} \\ &= 0\ \ \ Otherwise \end{aligned}$$

i.e., $I_i = I_j$ if $Sim(I_i, I_j) > \nu_{th}$, where $I_i \in A_o,\ \ I_j \in A_{ts}$

It is worthy to mention that binary precision and recall are special cases of fuzzy precision and recall with a membership value of 1. More precisely, if $f_{A_k}^{\nu_{th}}(I_i) = 1,\ \ where\ \ Sim(I_i, I_j) > \nu_{th} \forall k$ fuzzy precision and recall are equivalent to binary precision and recall.

- **Intra-list Diversity:** Intra-list item diversity measures how different items are recommended to users. We borrow the measure of diversity from Zhang and Hurley (2008) where diversity of any set A_o is defined as

  $Intra - diversity(A_o) = 2/p(p-1) \sum_{I_i \in A_o} \sum_{I_j \neq I_i \in A_o} d(I_i, I_j), p = |A_o|,$

  $d(I_i, I_j) = 1 - Sim(I_i, I_j)$

- **Inter-list Diversity:** Inter-list item diversity measures how different are the recommended items from users' current list of items. We have measured inter-list diversity of any recommended set A_o relative to training set A_{tr} as

  $Inter - diversity(A_o) = 1/p \sum_{(I_i \in A_o)} d(I_i, A_{tr}), p = |A_o|$

  and $d(I_i, A_{tr}) = Min_{(I_j \in A_{tr})}(1 - Sim(I_i, I_j))$

- **Entropy of Recommendation List:** The entropy (Shannon, 2001) of a recommendation list is defined (Pavlov and Pennock, 2002) as

  $H = -\sum_{i=1}^{i=n} p(i) log(p(i))$

  where $p(i)$ is the probability of occurrence of item i in the recommendation list and is calculated based on a popularity fraction of that item.

  $p(i) = number\ of\ users\ commented\ for\ item\ i/total\ number\ of\ users\ in\ the\ system$

  Higher entropy denotes that the distribution is less biased to only popular items.

2.7.4 Experimental Findings

The proposed method *Accurate Diverse Recommendation* (*ADR*) has been compared with three traditional collaborative filtering techniques, Item-based CF (*ICF*), User-based CF (*UCF*)) and Content-based technique (*CR*). In *CR* the content similarity of two items is computed based on the tags extracted from item descriptions. The comparative results of three aforementioned techniques with the proposed one are discussed in terms of accuracy, scalability, diversity, and entropy of recommendation in the following section.

The benchmark values of the experimental parameters are listed in Table 2.8.

Table 2.8: Benchmark Values of Parameters

Parameter	Benchmark Value	Range
β_{th}	0.6	0.2 - 0.8
ν_{th}	0.6	0.2 - 0.8
#users	5000	1000-8000
#items	66,137	N/A
Recommended set size	40	10-50

Accuracy

The accuracy of recommender system is of utmost importance, which determines whether the recommended items are correct output. We have measured the accuracy of our proposed algorithm *ADR* with that of the three baseline methods specified earlier. Traditionally, accuracy of recommender system has been measured by precision and recall, however, we have defined a new set of precision and recall measures (extension of traditional measures) as defined in the previous section. In case of binary precision and binary recall, similarity threshold (β_{th}) has been varied from 0.2 to 0.8, as shown in Figure 2.3 and Figure 2.4. Similarly, for fuzzy precision and fuzzy recall membership threshold value (similarly denoted as ν_{th}) has been varied from 0.2 to 0.8 and results are plotted in Figure 2.5 and Figure 2.6. The values of the other parameters have been kept in the benchmark values as in Table 2.8.

From Figure 2.4, it is clear that our algorithm (*ADR*) has very high recall value for small β_{th} and it reduces with increase of β_{th}. For $\beta_{th} = 0.8$ the recall value coincides with other three methods. On the other hand, the precision value for *ADR* (Figure 2.3) is higher than that of *ICF* and *UCF* but lower than CR for all values of β_{th}. Fuzzy precision and recall have also been compared where the fuzziness gives the true average closeness of the test set and the recommended set (Figure 2.5 and Figure 2.6). Similar to binary recall, *ADR* outperforms all three methods in terms of fuzzy recall. However, for fuzzy precision *ADR* is comparable to *UCF* but lower than *CR*. As the *CR* is based on content similarity and precision measures the similarity of recommended items with the actual items, the precision value of *CR* is expected to be higher than the other methods.

Figure 2.3: Binary Precision

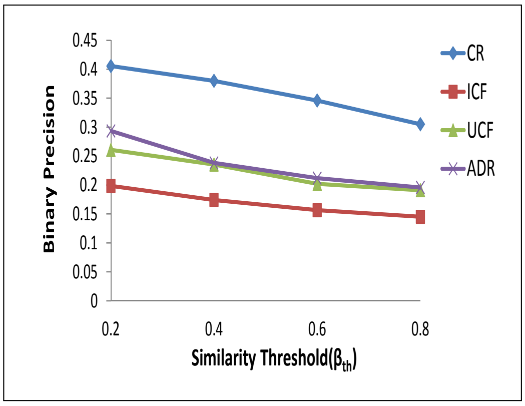


Figure 2.4: Binary Recall

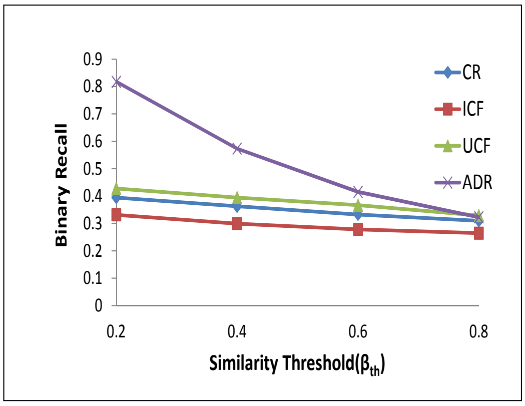


Figure 2.5: Fuzzy Precision

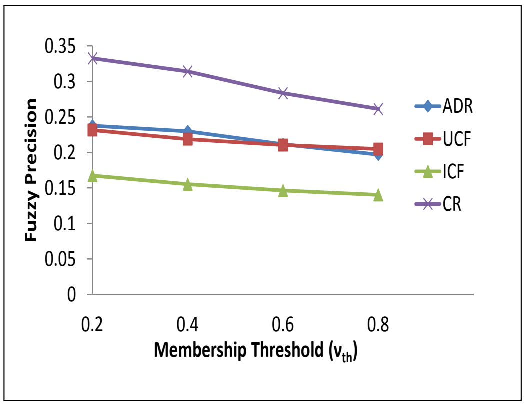
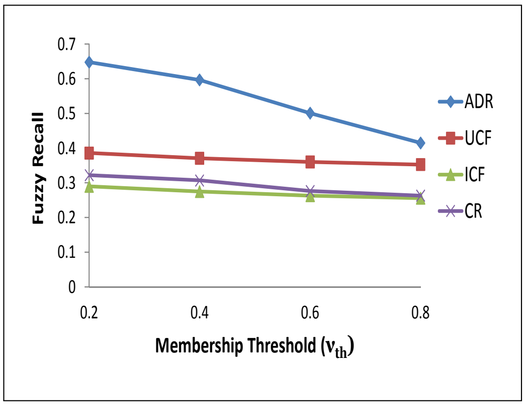


Figure 2.6: Fuzzy Recall



Diversity

Typically, research on recommender systems is concerned about finding the most accurate recommendation algorithms, however, the quality of the recommended items depend on many other factors, such as diversity of recommendation list. Our algorithm, *ADR*, promises to give diverse recommendations as inherently the algorithm chooses items from diverse categories. Diversity among the recommended itemset (*Intra-list Diversity*) as well as recommended itemset and user's own itemset (*Inter-list Diversity*) have been computed and compared against the baseline algorithms. Both diversity measures together determine the quality of the recommendation list in terms of novel diverse recommendations.

Figure 2.7: Intra-list Diversity

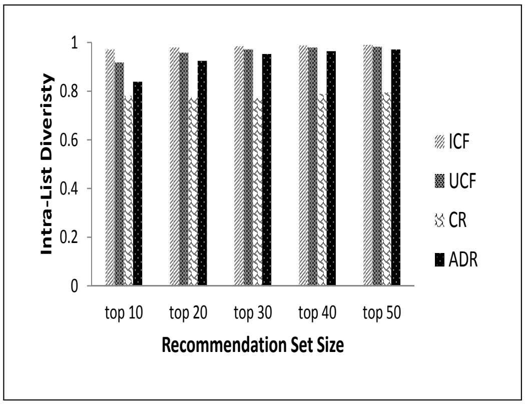
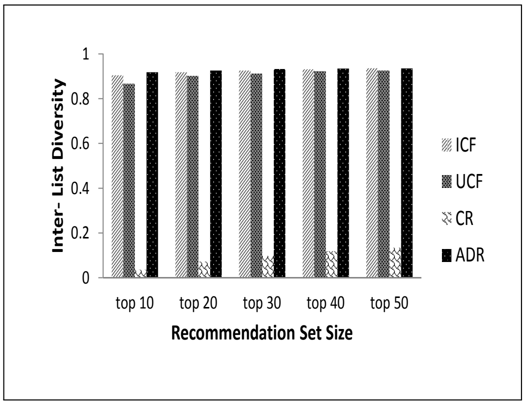


Figure 2.8: Inter-list Diversity



In order to measure diversity, top-k recommendations are generated for different values of k, $k = 10, 20, 30, 40$, and 50 for all four algorithms keeping all the other parameters at benchmark values as in Table 2.8. For each method both intra- and inter-list diversity have been plotted separately against recommendation set size k (Figure 2.7 and Figure 2.8). It is noted that *ADR* has similar diversity value as existing approaches in both cases. For inter-list diversity CR performs very poorly. This is obvious because in CR recommendations are generated based on similar content of the items. For the same reason *CR* gives high precision compared to the baseline algorithms.

Diversity vs Accuracy To understand the overall performance in terms of both accuracy and diversity of *ADR* vs the other baseline methods, accuracy vs diversity comparison have been plotted in Figure 2.9 and Figure 2.10 at the benchmark values listed in Table 2.8. We have compared inter-list diversity against fuzzy precision and fuzzy recall of all four methods. Figure 2.9 and Figure 2.10 indicate that overall *ADR* outperforms in both the cases, i.e., *ADR* achieves high diversity while maintaining good precision and recall.

Figure 2.9: Diversity Vs. Recall

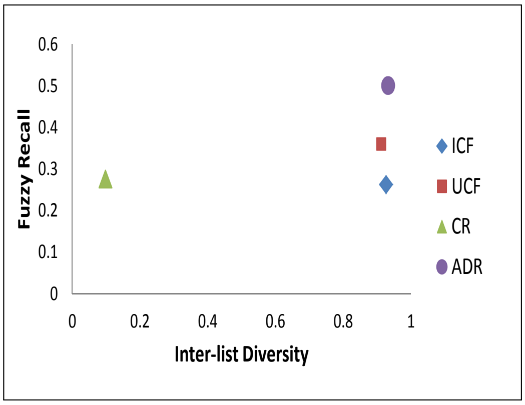
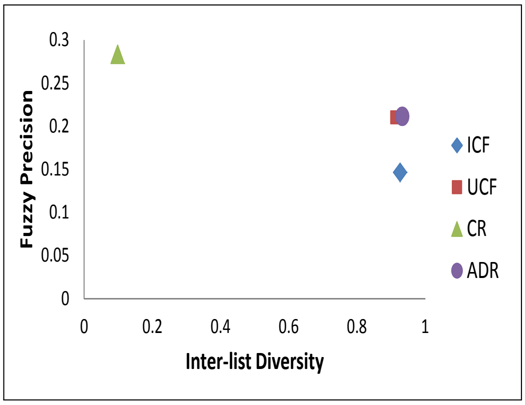


Figure 2.10: Diversity Vs. Precision



Scalability

For dynamically growing domain like mobile app domain, item-base increases rapidly which demands a scalable system that can process the massive number of items efficiently and provide real-time recommendations to users. To measure the scalability of the system, time spent in an offline and online recommendation generation processes have been plotted for all the algorithms (Figure 2.11, Figure 2.12) where user-base has been increased (1000-8000) keeping the number of items fixed (66,137). All the benchmark values have been listed in Table 2.8. For *ADR* offline process measures the time for generating association rules and the user-profile generation for the existing user-base which varies from 1000 to 8000. On the other hand, online process measures the average time for generating recommendation for each user when the user-base varies from 1000 to 8000.

From Figure 2.11, it is clear that the scalability in the offline process for *ICF*, *UCF*, and *ADR* are comparable and superior to *CR*. Figure 2.12 shows that in online process time spent in *ICF* and *UCF* increases rapidly with increasing user-base. In comparison, under the same conditions, time spent in *ADR* is maximum 400 milliseconds (0.4 Sec). In the context of massive scale problem online responsiveness is critical and recent literatures (Rui and Whinston, 2011) show that online system should respond within 2 seconds. Hence, performance of *ADR* is acceptable. *CR* consumes minimum time because the item-item similarity has already been calculated during the offline process and stored in the database for online recommendation.

Figure 2.11: Offline Time Spent

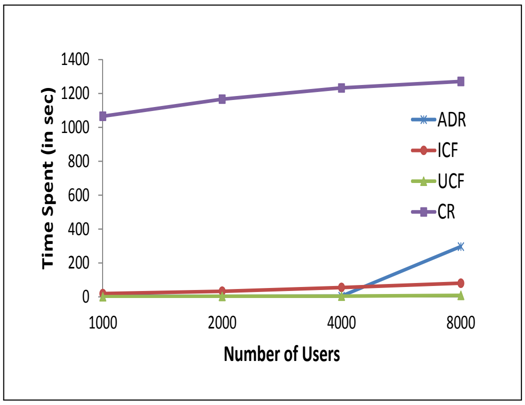
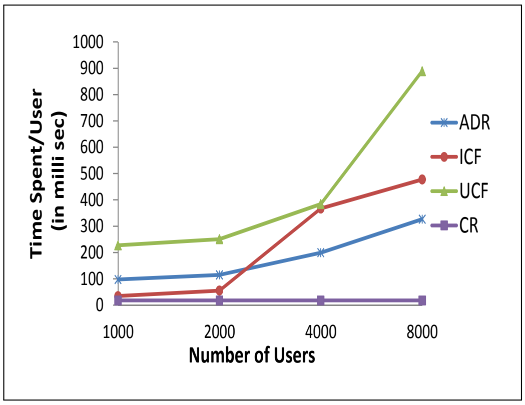


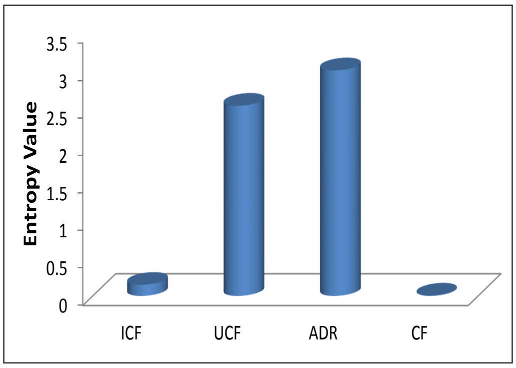
Figure 2.12: Online Time Spent



Entropy

The quality of the recommendation lists also depends on the fact that they should not be biased to popular items and to examine that we have computed the entropy of recommendation list. For each user, top 40 recommendations are generated and entropy for the recommendation list is calculated using entropy formula discussed earlier. Final entropy value is the average entropy values calculated for all the users. The highest value of entropy for *ADR* indicates that the generated recommendations are not biased to only popular items (Figure 2.13). *ADR* generates recommendations diversified by different categories which leads to a high entropy value.

Figure 2.13: Entropy in Recommended Items



Summary of Findings

Overall, *ADR* performs better compared to the baseline methods in terms of accuracy, diversity, and scalability. For membership threshold of 0.6, *ADR* improves fuzzy recall by 15% compared to *ICF*, with a similar fuzzy recall value (except *CR* method). On the other hand, both intra- and inter-list diversity (91% for *ADR*) of the recommendation lists was better than CR method (78% intra-list and 8% inter-list diversity) and comparable to the baseline methods. To get an overall comparative picture of accuracy with diversity, inter-list diversity and fuzzy precision and fuzzy recall have been plotted in Figure 2.5 and Figure 2.6, which reveal that *ADR* performs the best amongst all. Besides, time taken

Table 2.9: Comparison of Algorithms

Algorithm	Binary Precision	Binary Recall	Fuzzy Precision	Fuzzy Recall	Interlist Diversity	Intralist Diversity	Online	Offline	Entropy
ADR	0.29	1	0.47	1	1	0.85	0.50	0.999	1
ICF	0	0	0	0	0.99	1	0.05	0.96	0.05
UCF	0.24	0.65	0.47	0.41	0.975	0.94	0	1	0.84
CR	1	0.398	1	0.056	0	0	1	0	0

in offline and online method for *ADR* method scales well with the increase of the user base. Moreover, a high entropy value of *ADR* (3.5) confirms that the recommendations are not biased to popular items. For a better understanding, comparative values for each measure have been summarized in Table 2.9 for all baseline algorithms and *ADR* at the benchmark values. Each cell value in the table defines the standardized value std_{dim}^{Al} for a specific measure *dim* and algorithm *Al*. Hence, a higher value std_{dim}^{Al} for a measure *dim* determines superiority of the algorithm *Al* in that dimension. For the scalability measures for offline and online time, we have taken $1 - std_{dim}^{Al}$, so that the definition of a cell value remains same for measures. Cells in the table are colored in the gray-scale where gray-level is determined by the value of std_{dim}^{Al}. It is clear from Table 2.9, overall *ADR* performs better in all dimensions, whereas *ICF* performs well for diversity metric, but does not do well for accuracy metrics. On the other hand, *UCF* is good at binary recall and diversity metric and performs bad for online time spent. Finally, *CR* performs poorly at diversity metrics (inter-list and intra-list diversity). Overall, *ADR* achieves a balanced accuracy, diversity, scalability, and entropy measurement for the mobile app recommendation.

2.8 Summary

In this study, we have described a novel approach to generate mobile app recommendations for users in a scalable fashion. Association rule mining approach is used to generate rules for interrelated categories of users' transactions and following which user's profile is updated using pre-computed rules to redefine his category interest. In distinct contrast to traditional approaches where association rules are structured among the items (dynamic in nature), we have generated these rules among the categories (quasi-static in practice). Our findings show that this method can be used to predict mobile app with a legitimate recall value and comparable precision value. System scalability has been verified by measuring both offline and online time spent in the system. A comparison with the baseline algorithms demonstrates the superiority of our approach *ADR* in all the dimensions. Two measures have been proposed to find the modified precision and recall value when recommendation evaluation is an issue.

Moreover, this work is one of the first work to develop a recommendation system for the mobile app users. The system has been experimented with real-world mobile app users' data, which shows its applicability in online systems. Besides online responsiveness, the diversity of the recommended items enhances the quality of the recommendations. Therefore, it will be of much use to the mobile app marketers to target a wide range of audiences.

The dataset used for experiment contains information of apps installed on real mobile app users. Accuracy and diversity measures on the training and test set gives a good empirical evidence of the efficacy of our approach. However, human evaluation of the algorithm is important for recommendation systems. The best way to carry out the evaluation process is experimenting on human subjects based on how do they find the recommendation meaningful. This can further justify the quality of the recommendations. It would be meaningful to extend our evaluation approach on human subjects.

For future work, it would be interesting to investigate how social information can be integrated with the user profiles to understand their product preference. This would lead us to find the users in the community who share similar taste with the active user, for which there are now limited methods available, but will be very important in the future.

This page is left intentionally blank.

Chapter 3

Factors Affecting Retweetability: An Event-Centric Analysis on Twitter

3.1 Introduction

Twitter has progressed to be the most popular microblogging service by far which can disseminate up-to-the-minute information rapidly. It endows users to share information in real time beyond geographic constraints and has gained increasing attention for political campaigning (Abel et al., 2011), news media, crime information (Chu et al., 2010), and disaster communication (Hughes and Palen, 2009; Mendoza et al., 2010). Research has been conducted in the line of diffusion of information on Twitter, particularly, in the context of adoption cascades[1] (Gruhl et al., 2004) and trending topic detection (Pervin et al., 2013). However, a little attention has been paid on how information diffuses and

[1] every time a person close to another person u chooses an innovation, probability that u will adopt the innovation goes up

who participates in the diffusion process on Twitter, which demands further investigation. This is very significant, particularly in product advertising and campaigning through social media where a brand or a company seeks attention from large audiences very rapidly. This demands recognition of the potential and influential target audience on the Twitter network, who in turn can promote the product by tweeting / retweeting product related information to his friends and followers. Therefore, it is very important to identify the communicators in the diffusion process and investigate their roles in diffusion mechanism.

The principal factor of information diffusion on Twitter (Boyd et al., 2010; Suh et al., 2010), the so-called act of retweeting, allows users to broadcast someone else's tweet to their own set of friends and followers. In fact, the users can use the official retweet button to share the content just in one click. Though the practice of retweeting does not follow the standard rules, the most common practice of giving attribution to the user is adding "RT @" before the Twitter handle of the user. However, the construction and analysis of retweet network is not a straightforward task. Due to the limited 140 characters in a tweet, users frequently tend to delete or modify the tweet content to meet the character limit and this adds complications in the construction and analysis of retweet network.

Recently, a surge of interest has been observed to unfold the factors impacting the retweetability (probability of a tweet getting retweeted, which is usually measured by the retweet count of a tweet at a particular timestamp). Boyd et al. (2010) stated retweeting as a practice of participating in a conversation and studied the conventions and diverse reasons people retweet. On Twitter, information flows in a large network through the cascades of followers. To explode the social shares the tweets need to reach out the correct users timely and should attract their attention by its content. Suh et al. (2010) have shown that the inclusion of hashtags and URLs (Unified Resource Locator) in the tweet content increases its share count. While content features are important, retweetability mainly depends on who is seeing your tweet and eventually participating in the diffusion process. For instance, in the time of campaigning for new product launch the companies try to reach out the journalists and the celebrities to acquire involvement of more audiences in it.

While unstandardized retweet practices not only make the construction of retweet network non-trivial, consideration of only the tweet content to build retweet network also adds bias in the analysis. In this study we have focused to investigate the factors impacting retweetability considering both network variables (variables computed from *retweet network*[1]) and content variables of tweets. We present a systematic way to build the retweet network and then discuss the factors impacting retweetability. We define the retweet chain as the list of users in the retweet network arranged chronologically (according to the publication time of their tweets). The users in a retweet chain have been classified into three distinct classes, namely *information-starter, amplifier,* and *transmitter* according to their roles in information propagation (*Information Diffusion Impact*) and proposed scores of each user based on their roles. *Information Diffusion Impact* can be simply conceptualized as the number of people he makes aware of an information in

[1] By selecting the unique (tweeter, retweeter) pairs from the retweets, we obtained a network of tweeter → retweeter as edges and distinct users as nodes

the network. Finally, all the aforementioned factors, along with the three user scores are incorporated into the model.

In addition, we have investigated whether an external event can alter the probability of a tweet getting retweets. Our dataset (2011 Great Eastern Japan earthquake Twitter data, discussed in details later in the chapter) revolves around a major event and hence, allows us to specifically address this research issue. Moreover, we used another Twitter dataset from the Boston marathon bomb blast (April, 2013) to verify whether the results obtained in both the events follow a similar pattern. This in turn demonstrates the robustness of our findings.

For modeling the factors that affect retweetability, we used the regression technique. Furthermore, to check the effect of the event on retweetability we use the difference in difference estimator (DID) using three time windows centering the event in the dataset. The results obtained from both the datasets indicate that the user roles in information diffusion differs at the time of the event as compared to the pre-event time window. Users with comparatively less number of followers, i.e., not so famous on Twitter, participate in the information diffusion process during the event and play a significant role in the information diffusion process.

The contributions of our study are as follows:

i We define and classify user-roles in information diffusion directly grounded on the impact that the users have on the Twitter network. We also study the change of user roles at the time of the crisis.

ii We analyze the retweet network along with the follower network to understand the factors that has impact on retweetability. Herein, we check whether the user roles have a significant impact on retweetability along with the other factors.

iii We investigate the effect of a major event on these factors.

The rest of the chapter has been presented as follows: the following section presents a brief literature review, and then we describe the data collection and preparation. Next, we discuss the user classification process. Afterwards, we discuss the model predicting the retweetability of tweets. Finally, we summarized our findings.

3.2 Literature Review

A line of research focuses on understanding the communication during emergency situations prescribing to-be-done for the disaster relief management (Brashers, 2001; Guha-Sapir and Lechat, 1986; Hale et al., 2005; Pan et al., 2012; Pastor-Satorras and Vespignani, 2001; Richardson, 1994; Sellnow and Seeger, 2013; Wu et al., 2011). With the introduction of web 2.0 the communication medium is computer mediated. Vieweg et al. (2010) discussed how computer mediated communication and specifically microblog posts would be extractable for subsequent use in systems that support common situational awareness. A situational awareness perspective is helpful for anticipating how individuals, groups, and communities can use information contributed by others in a social media context (Vieweg et al., 2010). In fact, information technology played an important role in earlier disaster communications (Hughes and Palen, 2009; Zook et al., 2010; Vieweg et al., 2010). People adopt new technologies during the disaster and it spreads long term effect after the event also. Hughes and Palen (2009) observed that during the emergency events 13% of the tweets had URLs which increased up to 24.5% after the event. This demonstrates that because of the emergency event Twitter was adopted as a new medium of communication which gained popularity and sustained after the event. Moreover, they found users who adopted Twitter during an emergency event became long time adopters.

The Great Eastern Japan earthquake, happened on 11th March, 2011, was one of the five most powerful earthquakes in the world since modern record keeping began in the early nineteenth century[1]. When the earthquake occurred, there was no contact by cell phones due to a network outage, but people could still access the Internet through 3G services with smartphones such as iPhone. Reports (Tachiiri, 2011; Inose, 2011; Ogiue, 2011) show that Japan Government used Twitter to cope the crisis situation which helped to increase the awareness and reduce the anxiety level of the people in Tohoku area (Doan et al., 2012).

At the time of emergency situations number of tweets explodes. Here arises the simple question: "What does motivate people to share information, especially during the emergency situations", which demands thorough investigations? Sharing information with friends is considered to be a communal act in online social network sites. People share YouTube videos, Facebook posts, or tweets on Twitter. While a massive amount of information gets generated online, only a handful of them get noticed and shared. This

[1] http://www.webcitation.org/5xgjFTgf4

leads to the straightforward question, *what does make a piece of content more share-worthy than others?*.

In this study, we investigate information diffusion on Twitter. While a massive amount of information is available on Twitter, 40% of them are white noise (Chu et al., 2010). In the rest of the tweets many tweets are just the retweets of others. However, in practice only a small percentage of tweets get retweeted. What are the reasons for a tweet to get retweeted? What kind of content people share? The researcher has investigated that bad news travels faster on Twitter (Naveed et al., 2011).

In an early work, Kwak et al. (2010) have done a quantitative study of information diffusion on Twitter and investigated the relation between the author's in-degree and their reachability in the network. They argue that users with less than 1000 followers tend to have on average same number of additional recipients of the tweet. With the increase in the number of followers, the average amount of additional recipients increase. This suggests the clear correlation of in-degree of tweet author and the number of users reached on the network.

Suh et al. (2010) have examined a set of features that can predict the retweetability of a tweet. Applying Generalized Linear Model (GLM), they show that contextual features like hashtags, URLs or mentions affect the probability of a tweet getting retweeted. They also showed that if the original poster of the tweet has many followers and followees, the probability increases. Yang and Counts (2010) attempted to predict the information propagation considering properties like historical mentions of users using survival analysis modelling.

While the inclusion of features like URLs, hashtags, mentions or question marks in the tweet steer more attention, we claim that features like number of new people user makes aware of (not necessarily the number of followers), the position of the user in the retweet chain, and time of retweeting should also be considered. More importantly, users influence measure based on the information diffusion in the network has not been addressed in the prior literature. In this study, we define user impact score based on their role in the information diffusion process. In addition, the effect of a major event (like an earthquake) on the retweetability has previously not been investigated. The focus of this study is to investigate how a major event impacts the retweetability factors, particularly, how the user roles change due to a major event. This will be useful for strategy making in subsequent emergency situations, heretofore unexplored.

3.3 Solution Intuition

A new trend has emerged in product-advertising, marketing, political-campaigning through social media like Twitter as it is plausible to gain attention from large audiences very rapidly. While a massive amount of tweets is generated on Twitter, only a few of them gets retweeted and this spawns the age old query, "what makes it retweeted so widely?" To understand this retweet phenomena first we have built the retweet interaction graph or simply retweet network using both tweet content and follower information. Later, to unfold the role of users in information flow in retweet network, users are categorized into different roles based on their contribution in information diffusion in the network. Finally, a regression model has been constructed using the tweet related features (hashtags, URLs, etc.) and user score (calculated based on information diffusion impact of user in the network). The datasets used in this empirical study have been divided into three time-windows, pre-, during-, and post-event time windows to understand the effect of the event on retweetability factors.

3.4 Dataset Description

In this study, we have used two datasets from two separate emergency events - 2011 Japan earthquake and 2013 Boston Marathon bomb-blast. Both the datasets are described in turns.

3.4.1 2011 Great Eastern Japan Earthquake Dataset

2011 Great Eastern Japan earthquake at Tohuku area was of magnitude 9.0, which occurred on 11th March, 2011. The earthquake triggered powerful tsunami waves that reached heights of up to 40.5 metres in the Iwate Prefecture. The tsunami caused nuclear accidents in the Fukushima Daiichi Nuclear Power Plant complex. There were several (more than 1000) aftershocks in 2011 earthquake with magnitude of above 6.0. There were around 15,581 people were dead and 6,152 people were injured (Wikipedia, 2015a). The details of the twitter dataset collected during this time period is described below which shows a sudden increase in number of tweets during the event occurred.

I **Tweet Data:**

We used a Twitter dataset collected during the earthquake in 2011 described thoroughly in Toriumi et al. (2013). The dataset collection procedure has been discussed briefly here:

- First, a set of tweets has been collected from the Twitter streaming API during the event.
- Next, for all these tweets the user details along with the follower IDs have been crawled using the same API.
- For all these users the tweets were collected for 20 days of time period.

The dataset covers a period of 20 days (from 5th March, 2011 to 24th March, 2011), and consists of 362,435,649 tweets posted by 2,711,473 users in Japan. This dataset is remarkable by its completeness: 80% to 90% of all published tweets of these users were present in this dataset. It should be noted that the dataset consists of tweets of Japanese Twitter users only. A quick analysis of our dataset reveals that a major proportion of tweets (98%) in the dataset are written in Japanese.

Figure 3.1 shows the retweet count for a period of 20 days normalized to cut off daily variations. The first two major peaks represent the two big earthquakes on 11th and 12th March as reported in Wikipedia (2011). After the disaster, retweet count progressively returned to its normal average values.

II **Follower Network Data:** On Twitter, follower network depicts the social relationship between the users. Considering the Twitter API limit, collecting the follower information of all the users is time consuming. More importantly, if the users are not active in the time-frame of our study, we ignore those users. In order to select the active users, we have analyzed our dataset. On average, if a user is mentioned once everyday, he/she is active in the time frame. For our data collection we have chosen the same threshold. This will further reduce the cost of collecting data for not-so-useful users. Therefore, follower information has been collected by crawling Twitter API in May, 2013 for the active users who have been mentioned more than 20 times in 20 days.

Follower network dataset consists of 300,104 users and 73,446,260 relationship information. The degree distribution has been shown in Figure 3.2 by plotting the cumulative fraction of users against the number of followers / followees of the user. We acknowledge the fact here that the follower network information of the users is collected in a different time-frame. Also, we have the users' follower information only at one timestamp which restrain from studying the user behavior with the evolving follower network.

3.4.2 2013 Boston Marathon Bomb-blast Dataset

The 2013 Boston Marathon happened in Boston, Massachusetts on April 15, 2013. Almost two hours after the completion of the race, two explosions occurred near the finish line. Three spectators were killed and 264 others were injured. The bombs exploded about 12 seconds and 210 yards apart. The FBI led the investigation and on April 16th, the photographs of the suspects were released which resulted in killing 3 people (Wikipedia, 2015b). The details of the dataset is described below.

I **Tweet Data:** We have collected a month's Twitter data of Boston-marathon bomb-blast in 2013. For collecting this data we have used the following approach.

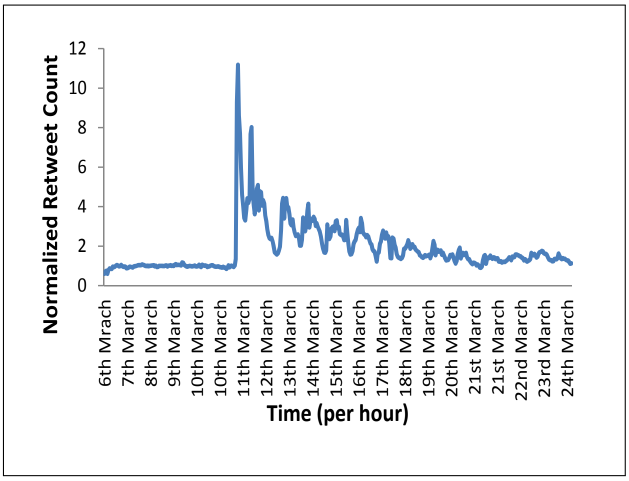
- Tweets were collected using the Twitter Search API using keywords like 'boston', 'bostonmarathon' etc. dated 15^{th} April, 2013.
- For all the above tweets, the profiles of the users (U) were tracked, (e.g., follower count, time zone, name, etc.).
- Next, for all the users in U, we collected the tweets for a month period, from April 1^{st} to April 30^{th}, 2013.

Figure 3.3 shows the tweet count for a period of 30 days (normalized to cut off the daily variations). Major peaks represent the high tweet count during the bomb blast. In this dataset, there are 112,93,215 tweets posted by 30,000 users.

II **Follower Network Data**

Investigating the tweet contents we found the users who participated in tweeting and retweeting at the time of crisis. The follower information of 30,000 active users has been collected using the Twitter API. Follower network consists of 30,000 users with 73475897 relationships.

Figure 3.1: Tweet Distribution over Days (Normalized), Japan Earthquake Data



3.5 Solution Details

In this work, we have investigated the retweet network (also referred to as activity network) and the static follower network of the Twitter users simultaneously. On Twitter, the retweet functionality allows users to share information with their friends and followers, generating a network of retweeters. Here, this retweet network is considered as the activity network. To analyze the information diffusion for each tweet, we are interested in the retweet sequence of each tweet, which we refer to as retweet chain.

3.5.1 How to find Retweet Chain

In recent works, particularly the work by Tinati et al. (2012) proposing a classification of user's roles, the diffusion of information is directly extracted from the content of the tweets. For instance, if a tweet published by user u_1 is composed of the following pattern:

"RT @u_0 tweet"

Figure 3.2: Cumulative Fraction of Users by Degree, Japan Earthquake Data

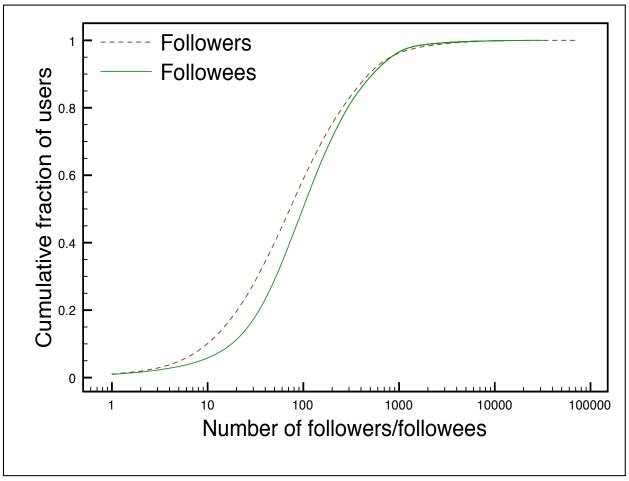
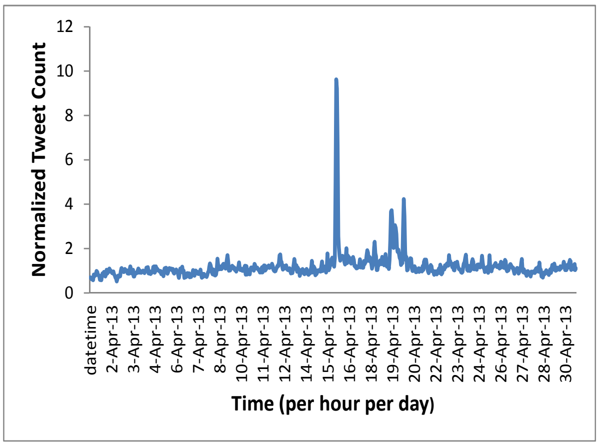


Figure 3.3: Tweet Distribution over Days (Normalized) Boston Marathon Bomb Blast



then one considers that the information diffused directly from u_0 to u_1. Retweet chains are identified by tweets containing several references, i.e *"RT @u_1 RT @u_0 tweet"*, or consecutive citations, such as u_1 posting the retweet: *"RT @u_0 tweet"* followed by u_2 posting *"RT @u_1 tweet"*.

However, in reality, after one step of citation, this has two important biases:

- users tend to keep only the original author of the tweet, and not intermediates, in particular to meet the 140 character limit of Twitter. Even when using the official retweet function of Twitter, only the initial poster is kept. This will strongly increase the number of direct retweets and in turn the apparent role of the original poster in the diffusion of information.

- users frequently retweet after seeing a tweet several times, as it has been shown in Leskovec et al. (2007). As a result, the user cited as the source might not be fully representative of the information diffusion.

In this work, to characterize the diffusion of information, we will therefore adopt a combination of both the follower network and the retweeter information from tweets. A retweet chain is simply defined as the sequence of all tweets published containing the original content, ordered by their publication time. To consider the information flow, we combine this information with the assumption that, each time a user publishes a tweet, all his followers can see the information. We can therefore know by whom the user might have been informed, independently of the user who appears as the source in the tweet itself.

3.5.2 User Classification

We classify user roles in the light of information propagation through retweeting. By analyzing the retweet chains the users are classified into three categories, "information starter", "amplifier", and "transmitter".

- *Information starters* are the users who are able to launch new information which will spread broadly in the network. They are the users whose information will reach many.

- *Amplifiers* are the users who do not publish interesting content by themselves, but who have the potential to diffuse information published by others to many new people.

- *Transmitters* are the users who act as bridges between several communities in the network. If an *information starter* publishes an interesting tweet in a given community of the network, the *amplifier* will spread this tweet in the same community, but the *transmitters* are necessary to reach other communities which in turn will result in transmission of the information broadly.

We base our user role definitions on the concept of *Information Diffusion Impact (IDI)*, namely for a user u_1, the number of users he made aware of an information i. Therefore, making 10 people aware of one information and making one person aware of 10 different pieces of information result in the same IDI value. This notion is very important, as it allows us to compare the impact of different roles. For each user, we can compute a value of *IDI* for each behavior (*information starter, amplifier, and transmitter*), which represents the impact of the user on the diffusion of information: how many people were impacted by his publication of a tweet? How many people became aware of a tweet through his action of retweeting? And how many people could access the information because the user transmitted it to another community? These values are therefore comparable.
The notations used in this study have been listed in Table 3.1.

Table 3.1: Notation Table

Notation	Meaning
N^t	number of new people aware of tweet t
N_u	number of new people made aware by user u
$InformationStarter(u)$	Information starter impact of user u
$Amplifier(u)$	Amplifier impact of user u
$Transmitter(u)$	Transmitter impact of user u
C_i	community i
$follower_{C_i}(u)$	follower set of user u in community C_i
$order(u)$	position of user u in the retweet chain

Information Starter: *Information-starter* can be conceptualized as the one who creates the original information. *Information starters* are important as their information is retweeted by others and depending on the importance of the content, it is diffused further in the network. For each user u in the retweet chain, we compute the number of new people (N_u) u makes aware of, using u's follower information. The total number of people (N^t) in the network aware of the tweet t is given by

$$N^t = \sum_{\forall u \in retweet\ Chain} N_u$$

Here, N^t is the number of different users made aware of the tweet, which is the impact of *information starter u* for tweet t. Hence, the overall impact of u as an *information starter* is $InformationStarter(u)$ and is defined by,

$$InformationStarter(u) = \sum_{\forall t, u\ starts\ t} N^t$$

It is worthy to note that, a good *information starter* is not necessarily followed by many people.

Amplifier: *Amplifiers* are considered as the individuals who share others information and make many people aware of it. They are important as they are followed by many users and as a result, *amplifier* makes a large fraction of users aware of the information. To compute the power of the *amplifier*, unlike *information starter*, here we calculate the direct impact of the user in the network. For each tweet t, u participated, but not the *information starter*, we compute the number of new people u makes aware of, say N^t.

$$Amplifier(u) = \sum_{\forall t, u \text{ is not information starter of } t} N^t$$

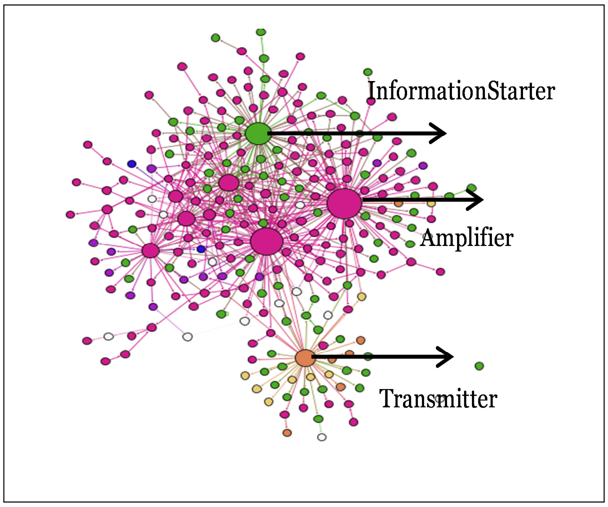
We should note that this value is usually less than the number of followers of u, as some of his followers are already aware[1] of the tweet. Therefore, the user who appears early in the retweet chain will naturally have a higher *amplifier score*.

Transmitter: It is now accepted that most social networks have a strong community structure (Girvan and Newman, 2002). The Twitter follower network is no exception, and its analysis reveals clearly defined modules. In this study, we used the Fast OSLOM algorithm (Lancichinetti et al., 2010) to detect communities in our follower network. This recent algorithm has several advantages:

- it is fast, which is important in our case, as our follower network contains more than 73 million edges.
- it allows overlapping of communities, an important property in this work, as we

[1] When we say a user is aware of a piece of information, we mean that the information is available to that user. It is possible that the information is available to him, but he did not consume it. In our measure we do not account this situation.

Figure 3.4: Retweet network of a popular tweet



want to find the users who might act as bridges between communities.

The algorithm found 8 communities in our follower network with an average size of 44668 nodes per community. By manual investigation, we found obvious meaning for some communities, such as a community of foreigners and a community of users related to nightlife (disc jockeys, hip-hop celebrities, etc.).

We observed that communities play an important role in the diffusion of information. Maximum number of tweets are diffused only in a fraction of one community, and some of the tweets get retweeted widely, but still confined in the same community. Therefore, we identify a user as a *transmitter* who spreads a tweet initially stuck in a community A to another community B. We consider that a tweet is stuck in a community if the first 20 retweets are in the same community. This number has been chosen experimentally, as we observed that the tweets which get retweeted 20 times in the same community, they tend to be stuck there. Therefore, the first user from a different community B to retweet is considered as a *transmitter* to B if again gets retweeted by other people from B.

The impact of a *transmitter* for one tweet is simply the number of people who gain access to the information by his retweet. More formally, the effective number of users informed about the transmission of tweet i in community C_j is the summation of the number of followers of retweeters in community C_j.

$$Transmitter^i_{C_j}(u) = |follower_{C_j}(u)| + \sum_{u_k \in C_j, order(u_k) > order(u)} |follower_{C_j}(u_k)|$$

where $order(u)$ is the position of user u in the retweet chain of tweet i and $follower_{C_i}(u)$ represents the number of followers of user u in community C_i.

If a user belongs to several communities, he can be a transmitter to different communities for a single tweet.

$$Transmitter^i(u) = \sum_{\forall j,\ u\ transmits\ to\ C_j} Transmitter^i_{C_j}(u)$$

Hence, an overall transmitter score of u, for all tweets he is transmitter, can be given by

$$Transmitter(u) = \sum_{\forall i,\ u\ is\ a\ trasmitter\ of\ i} Tranmitter^i(u)$$

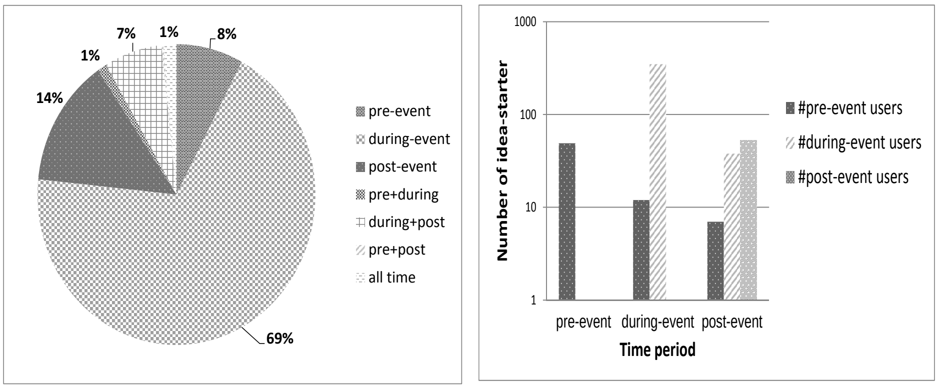
Figure 3.4 shows the retweet network corresponding to a popular tweet, where each node represents a user and an edge $B \rightarrow A$ exists if B follows A and $order(B) > order(A)$ in the retweet chain. Node color represents the community he belongs to and the size of a node is an indicator of the number of followers of that user. By our metric we identify *information-starter*, *amplifier*, and *transmitter* in the retweet chain. One can note that the *information-starter* is not followed by many people, as the size of the node is moderately small. The *amplifier* is the one with many followers and well-connected in the network. On the other hand transmitter is the node from a different community where he diffused information.

3.5.3 Evolution of User Roles over Time

In different tweets, one user might have different roles, *information-starter / amplifier/ transmitter*. We have measured the individual impact for each role. In Figure 3.5 and Figure 3.6 we have computed the percentage of users who retained and disappeared as an *information-starter* and an *amplifier* respectively in the three time windows. Figure 3.5a shows the overall distribution of the *information-starter* in the three time-windows. A hopping 69% of the total *information-starters* emerged only during the earthquake and 7% of the popular *information-starters* remained popular after the event also. To understand the transition of the *information-starter* from one time-window to another, we analyzed the proportion of the *information-starters* in pre-event time-window, who retained in other time-windows. From Figure 3.5b, one can see that out of 49 users in pre-event time-window, only 12 retained during the event. After the event, it was only 7. Also, a large number (349) of *information-starters* emerged during the event and 38 of them retained and 53 new users emerged after the event.

Similar analysis has been carried out for amplifiers in Figure 3.6a and Figure 3.6b.

Figure 3.5: Distribution of Role Retention as the Information-starters in Pre-, During- and Post-event Time Windows



(a) In all time windows

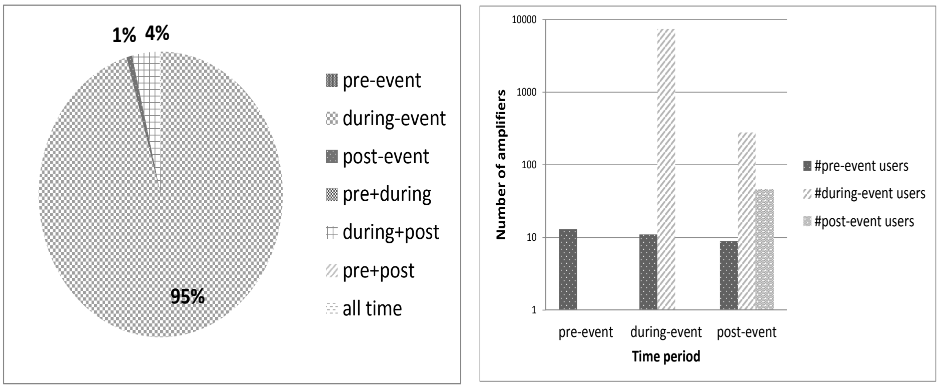
(b) Pre-and during event windows

The overall distribution of the *amplifiers* in Figure 3.6a shows that the number of new *amplifiers* who emerged during the event and disappeared after the event is very high (95%) and 4% of the *amplifiers* emerged during the event continued to contribute after the event also. Figure 3.6b shows that popular *amplifiers* in pre-event time-window (= 13) tends to be popular during the event (= 11) and after the event (= 9) though a high number of *amplifiers* appeared only during the event (= 7400). Comparing Figure 3.5 and Figure 3.6, one can note that a large number *information-starters* and *amplifiers* in the post-event time-window were from during-event time-window.

3.5.4 Associations of User Roles

In Figure 3.7, we plotted (in log scale) *Information-Starter Impact* against *Amplifier Impact* for each user for three time-windows. We have divided the region into four quadrants - clockwise from the origin they are named as average users, high-impacted *amplifiers*, super users, and high-impacted *information-starters*. The points on the x-axis and the

Figure 3.6: Distribution of Role Retention as Amplifier in Pre-, During- and Post-event Time Windows



(a) In all time windows

(b) Pre-and during event windows

y-axis represents the pure *information-starter* and *pure amplifier* respectively (as shown in Figure 3.7d). We have plotted only the users for whom the sum of *information-starter impact* and *amplifier impact* is at least 100,000, which means that overall, the user has impacted 100,000 users in the network as an *information-starter* or *amplifier*, which is basically the user's *IDI* value. In pre-event time-window (Figure 3.7a), number of high-impacted *information starters* (in quadrant 4) is comparatively larger than high-impacted *amplifiers*(quadrant 2). The number of super-users in pre-event time-window is comparatively lower than other time-windows. In during-event time window (Figure 3.7b), there is a gradual increase in the number of users in all quadrants and number of super-users are maximum during the event who contribute a lot in launching important information and spreading to others in the network. We have also observed that many *information-starters* started behaving as amplifier during the disaster. For instance, the user 'earthquake_jp' was a bot in the pre-event time window and acted as only good *information-starter*. However, during the event, it started retweeting other's tweet and became potential *amplifier*, as commonly referred as cyborg in the literature (Chu et al., 2010). Interestingly, after the event (Figure 3.7c) it became again a bot. Unlike 'earthquake_jp', user 'nhk_pr' was an information-starter as well as an amplifier in all time. Particularly during the disaster, he became very popular both as an *information starter* and an *amplifier* and also remains as a potential *information-starter* and an *amplifier* after the event. Interestingly, in the post-event time-window, many users were observed with high impact in dual roles and some new users emerged as potential *information-starters* and *amplifiers* after the event and a number of super-users increases compared to pre-event time-window.

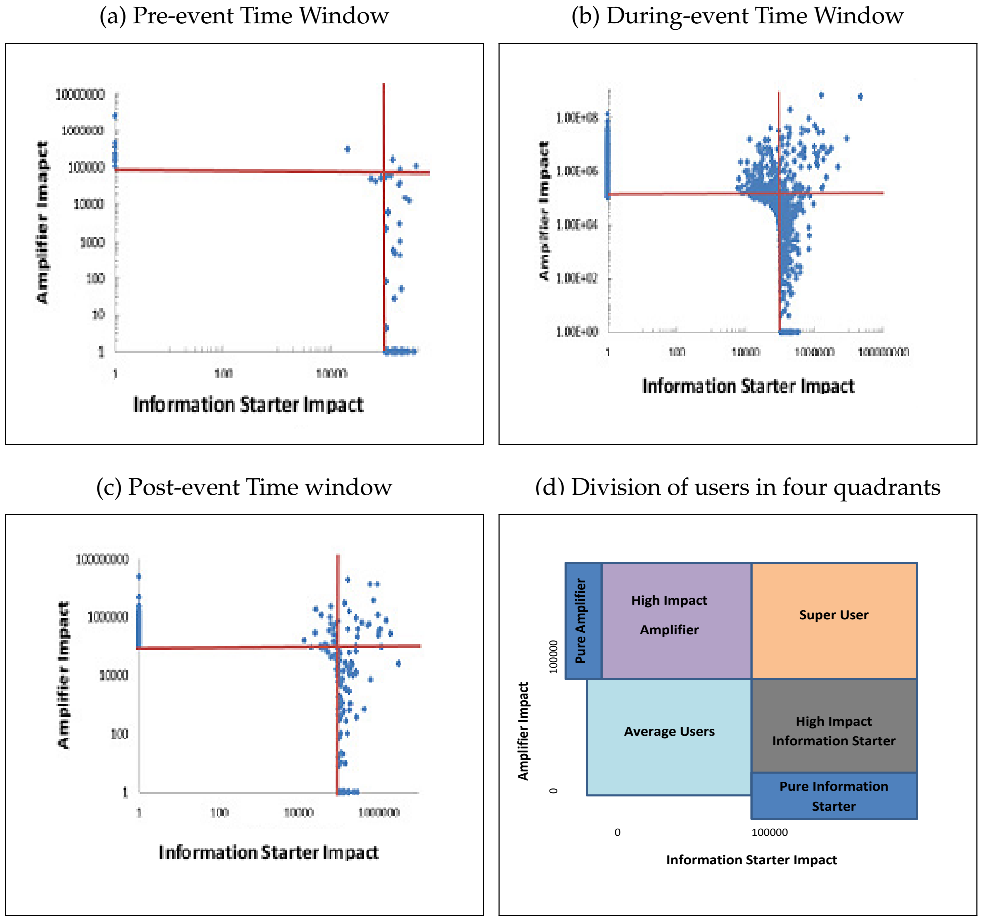
Scoring high *transmitter IDI* value is rarer than other two metrics. However, a comparison of the top 100 *information-starters*, *amplifiers*, and *transmitters* is carried out, which reveals that 21 users were listed in both *top-information starter* and *top-amplifier*, 7 were listed in both *top-amplifier* and *transmitter* and 1 was in *top-information starter* and *transmitter*. The popular celebrity with Twitter id 'ayu_19980408' was there in all three top-lists.

3.5.5 Transmitter's Topology

According to raw *IDI* values, *transmitters* do not have an impact as high as the two other roles. However, many *transmitters* had a strong impact on the diffusion of information with 15 users having an overall *transmitter score* above 100,000 and 538 users with a score above 10,000.

We can identify two categories of *transmitters*. The first category corresponds to

Figure 3.7: Information-starter vs. Amplifier Impact in Pre-, During- and Post-event Time Windows



(a) Pre-event Time Window

(b) During-event Time Window

(c) Post-event Time window

(d) Division of users in four quadrants

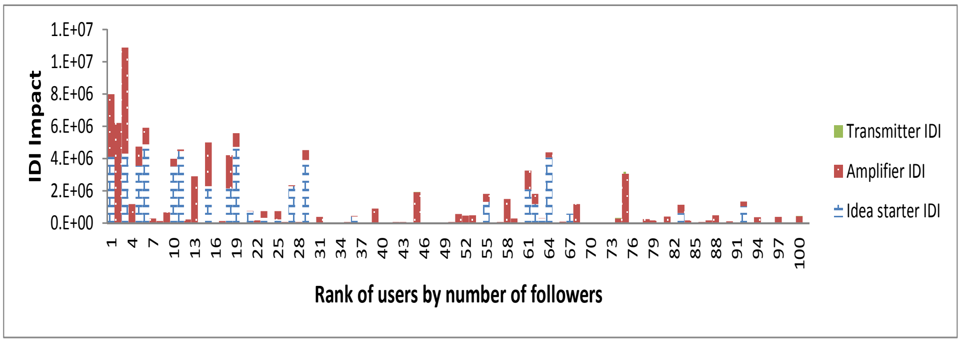
users who are frequent transmitters to small communities. They have been transmitters for several dozens of tweets, but to the community they transmitted information is not very large, resulting in relatively low *IDI* scores. On the contrary, some of the users with top transmitter scores are transmitters for less than 5 tweets; but they were transmitting an information from small communities to the largest ones. Therefore, a tweet which could have reached only a fraction of all users without transmission, it reaches most of the network after transmission. The impact of the transmission is therefore very high in this case.

3.5.6 IDI of User Role and Number of Followers

We investigated the correlation between the overall *information-starter, amplifier*, and *transmitter* impact of each user with that of their number of followers. We found that number of followers is not correlated with that of *information-starter impact* (correlation = 0.2827), *amplifier impact* (correlation = 0.4352), and *transmitter impact* (correlation = 0.0273).

Figure 3.8 shows the contrast of 100 top-followed users with *information starter, amplifier*, and *transmitter impact*. One can note that amongst the top-followed users, the roles are very different and they have very different *IDI* impacts. Hence, metrics like number of followers cannot determine the user-roles we discussed.

Figure 3.8: Comparison of number of followers with IDI impact of three roles



3.5.7 What Factors to Consider?

In this study we want to model retweetability. To do that we need to find the factors that might affect retweetability. Particularly, we want to investigate the user roles on retweetability. Therefore, we consider the user's *IDI* value for three different roles as discussed above. Along with those we use the following variables:

Network Variables

Number of followers: Similar to previous works, we have also validated the correlation of in-degree of user with retweet count. A preliminary analysis of our dataset shows that the average retweet count increases with the number of followers of the original poster of the tweet.

PageRank: Each user on Twitter has a number of followers and followees which can be thought of as incoming and outgoing links from a web page. Similar to web pages, we can also compute the PageRank of a user to enumerate the popularity of the user.

Content Variables

Hashtag inclusion: In previous works, particularly Suh et al. (2010) have shown that having hashtags in the tweet increases the probability of retweeting greatly. The hashtags have been extracted from the tweet contents by searching words that start with "#" symbol. An indicator variable has been used to specify whether the tweet has hashtags or not. The value of the variable = 1 if it contains hashtags, 0 otherwise.

URL inclusion: Suh et al. (2010) have also checked the inclusion of URL increases the probability of getting retweeted. Similar to hashtag, regular expression has been used to extract URLs from tweets. An indicator variable has been used to specify whether the tweet has URLs or not. The value of the variable = 1 if it contains URLs, 0 otherwise.

Table 3.2: Factors Affecting Retweetability

	Variable	**Meaning**
Dependent Variable	Retweet frequency	Number of rewteets by a user per unit time
Network variable	Number of followers	Number of followers of the user/retweeter
	PageRank	Calculated PageRank for the user using in-degree and out-degree information
	Amplifier score	How many new people he can make aware of
	Information-starter score	How many people made aware of the tweet he is the author
	Lag people aware at (t-1)	Number of people aware in previous time window in the retweet chain
User specific variable	Tweet Count	Total count of tweets by a user
	Average position (Early Retweeters)	Position of the user in the retweet chain, taking all tweets by a user we compute the average position of the user in the retweet chain to indicate early/late retweeters
Content variable	Hashtag	Indicator variable to specify whether a tweet has hashtags
	URL	Indicator variable to specify whether a tweet has URLs
Control Variable	Time of the day	24 hours have been divided into 5 time-windows, morning (7am-10am), noon (11am-3pm), afternoon (4pm-7pm), evening (8pm-11pm), night (12am -6am). This has been coded as a dummy variable indicating the 5 time-windows.
	Day of the week	Day of the week is coded as a dummy variable
	Tweet Age	Time since the tweet is composed (in hour)

Other factors

Day of Week: Day of the week might have impact on retweetability depending on it is a weekend or weekday. TweetSmarter (2011) found that day of the week controls traffic on Twitter, while Monday to Thursday the tweet volume increases, Friday it slows down. In our model we have included this as a control variable.

Time of the day: Reports show that Twitter gets the most traffic during 9am-3pm from Monday to Thursday (TweetSmarter, 2011). We also include time of the day as a control variable in the retweet model.

Tweet Count: Users who are active are the only ones to retweet more (Sysomos, 2009). If the user participates in writing, commenting, or sharing tweet, it shows his activity in the network. We counted the number of tweets (*tweetCount*) each user has participated either by tweeting or retweeting.

Tweet Age: The lifetime of a tweet is very short (less than 48 hours GaggleAMP (2013); Frederic (2010)), usually with time the retweetability first increases and then decreases. Particularly, in our dataset we have also observed that in the beginning the frequency of retweets is high, which decreases slowly with time. On average, the lifetime of a tweet is 24 hours. However, a few tweets were retweeted more than 10 days. Most of these tweets were about the earthquake which started on 11^{th}-13^{th} March and were retweeted till the last date of our dataset.

Early retweeters: There are some users who like to retweet very early. These are the users who make many people aware of the tweet for the first time through their follower network. For each user we find their position in the retweet chain. If the user has many followers and he is in the beginning of the retweet chain he can make a large number of people aware of the tweet.

In Table 3.2 we present all the variables we have considered for modeling retweetability.

3.6 Data Analysis and Findings

3.6.1 Data Preparation

Using the Twitter dataset described in Section 3.4, we randomly selected 10,000 widely retweeted tweets. For all these 10,000 tweets, retweet chains have been formed, which are basically the chains of users in the chronological order of their retweet of the original post. Our tweet dataset (5^{th}-24^{th} March) has been divided into three time windows, pre-earthquake ($5^{th}-10^{th}$ *March*), during-earthquake ($11^{th}-18^{th}$ *March*), and post-earthquake ($19^{th}-24^{th}$ *March*). For each tweet, we first build the retweet network, i.e., we identify the users who retweeted the tweet along with the timestamp of their retweet actions. Next, the retweet frequency has been computed per minute. On average, the lifetime of a tweet is very short, less than 48 hours (GaggleAMP, 2013; Frederic, 2010) and there is a handful of tweets which get retweeted for more than 5 days.

Figure 3.9: Retweet Frequency Distribution by Day of the Week

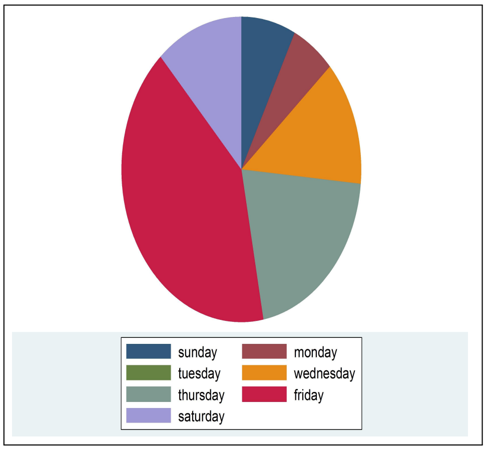
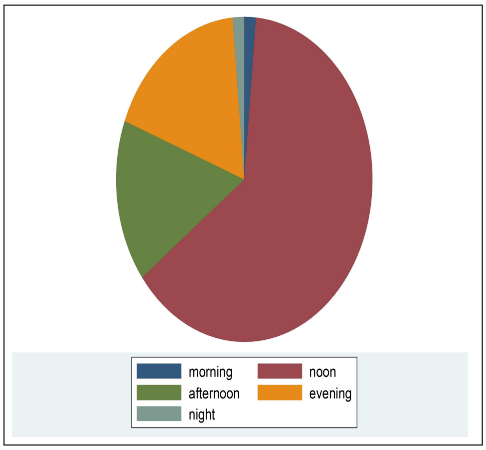


Figure 3.10: Retweet Frequency Distribution with Time of the Day

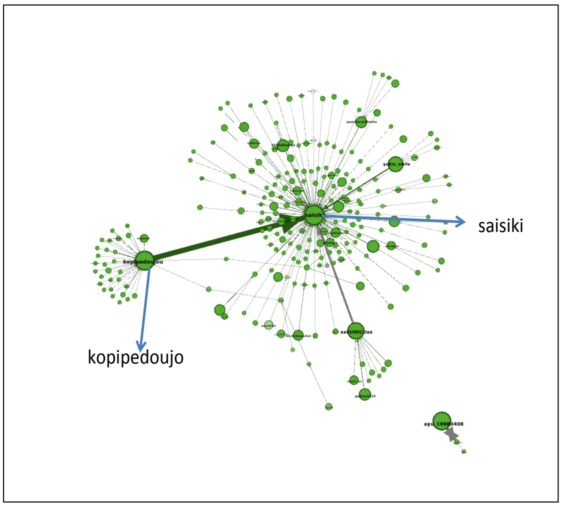


3.6.2 Data Analysis

Using the follower network information the number of followers, PageRank, and number of new people users make aware of in a retweet chain have been computed. The PageRank of a user estimates the popularity of a user. The number of new people he makes aware of determines his own contribution in the retweet process. Notably, the number of followers of a user and the number of new people he makes aware of are not the same because there will be overlap among the followers of the retweeters. For instance, a user u_1 can have thousands of followers, but if he retweets after user u_2 and all the followers of u_1 are included in the set of follower of u_2, then u_1 cannot make any new people aware of the tweet. Thus, the user's action of retweeting will contribute to the awareness of the tweet depending on the position of the user in the retweet chain.

We analyze the retweet frequency over the day of week and observed that throughout the week tweets get retweeted, but on the Friday retweet frequency seems to be much higher (Figure 3.9). Retweet frequency of tweet is also monitored round the clock. In general, maximum retweet happens during the noon time (between 12 noon to 3pm)(Figure 3.10) which is in inline with earlier findings (TweetSmarter, 2011). For obvious reason in the morning and night time the retweet frequency is the lowest. We have used these variables as controls in our model.

Figure 3.11: Example of retweet chain of a widely retweeted tweet, clearly the tweet was retweeted widely after the amplifier retweeted it



Users who are participating in the retweet process also play very crucial role. The one who starts the tweet (or "information-starter" as defined earlier) does not necessarily have many followers. But if the tweet gets noticed by a highly influential user it will be retweeted by many. As shown in Figure 3.11 the tweet was first tweeted by “kopipedoujou” and he was retweeted less time, however, while retweeted by “saisiki” the tweet exploded in a bigger network.

Besides network structure and the users’ participation in the retweet actions, tweet content also needs to be considered to understand retweetability. Usage of the hashtags is very common and it allows the user to follow or search related information regarding the topic of the hashtag on Twitter (Tsur and Rappoport, 2012). Previous researchers (Boyd et al., 2010; Suh et al., 2010) have found evidences that inclusion of URLs and hashtags increases the chance of retweetability. In our dataset among the retweeted tweets 26.5% of the tweets have a URL and 10.3% of the tweets contain hashtags. We have revisited the impact of the URLs and hashtags in retweetability.

3.6.3 Retweet Model

To model the factors affecting the retweetability of tweets, we have considered the variables described in Table 3.2. For randomly selected 10,000 tweets (each tweet was retweeted at least once) we have constructed the retweet chain with the chronological sequence of the users who retweeted. The regression technique has been used to model retweet frequency of a tweet and hence the dependent variable considered is computed as the number of times a tweet gets retweeted per minute (retweet count of a tweet per minute). The retweet model is given below and correlations among the variables are reported in Table 3.6.

$$
\begin{aligned}
RetweetCount_{it} = {} & \beta_1 peopleAware_{i,t-1} + \beta_2 NumFollowers_i(u) + \beta_3 AmplifierScore_i(u) \\
& + \beta_4 InformationStarterScore_i(u) + \beta_5 TransmitterScore_i(u) + \beta_6 isHashTag_i \\
& + \beta_7 isURL_i + \beta_8 isURL_i \times isHashTag_i + \beta_9 Age_{it} + \beta_{10} TweetCount_{it} \\
& + \beta_{11} DayOfWeek_i + \beta_{12} TimeOfDay_i
\end{aligned}
$$

We want to estimate the effect of the independent variables on retweet count per unit time. In the model we have used the user's score based on their roles in information diffusion. In section 3.5.7, we have already discussed the user roles, *information-starter, amplifier,* and *transmitter scores* and we want to examine whether these user roles are important in order to get the higher retweet frequency. On Twitter, users can follow tweets of a specific topic by following hashtags, or in other words hashtags make a tweet discoverable. On the other hand, since a tweet can contain a maximum of 140 characters, users tend to include shortened URLs to add more information to the tweet. While the content of the tweet is important, who is tweeting or retweeting a tweet is also important. Twitter users with a high PageRank or a large number of followers are classified as influential persons by researchers. Retweeters with large number of followers help a tweet to get spread in a bigger community. We want to investigate whether these effects contribute to the retweet frequency and we use a panel regression model to estimate the effects.

Next, we investigate whether there is an effect of the event on retweetability. We used the Japan earthquake data to investigate the effect of the event, earthquake being the event in this case. Afterwards we examine how the factors discussed earlier affect differently on retweetability at the time of the event as well as in the post-event time window. With a different dataset of Boston marathon bomb blast in 2013, we have replicated the experiment similar to Japan earthquake.

3.6.4 Findings and Discussion

For the three distinct time-windows in Japan earthquake dataset, the model has been tested using Generalized Least Square (GLS) regression model and the results are shown in Table 3.3. Number of people the users make aware of in the previous time-units (here previous minutes), i.e., $PeopleAware(t-1)$ does not have a significant impact on retweet frequency in the pre-event time window. However, in the during-event time window and post-event time window this impact became positive. On Twitter, same tweets get retweeted from several sources (followees), and people might retweet it after seeing it from more than one source. In normal situation (when there is no event), people may not retweet it immediately. However, in the time of emergency if more users see the tweet (more people are aware of the event) it increases its retweetability, whereas in normal situation this effect is much more complex.

Interestingly, users with high number of followers have a positive impact on the retweet frequency for both pre- and post-event window, but during the event the effect is opposite. This indicates that more users with low indegree (number of followers) participated in the retweet process. Also, the impact of *amplifier score* and *information-starter score* on retweet frequency is very high during the event. This suggests that during the event the highly retweeted tweets were retweeted mostly by the low in-degree users. However, by the definition of our *amplifier* and *information-starter score* of the users, the users having low in-degree can have a high *amplifier score* or an *information-starter score* if he makes a large number of audience aware of the information. To put it in simpler words, in the time of crisis the users who usually creates and shares tweets are not so famous Twitter users.

Surprisingly, *transmitter score* has negative impact on retweet frequency. This result is non-intuitive as we hypothesized that if a tweet is transmitted to many communities the tweet will be retweeted more. It might be due to the fact that retweetability of a tweet increases when the same content is presented to user's timeline multiple times. If the tweet does not get retweeted in the same community many times it reduces its probability to be retweeted in that community.

Like previous works (Tsur and Rappoport, 2012), our model suggests that inclusion of hashtags and URLs have significant positive impact on the retweet frequency in pre-event time window. However, the effect does not hold at the time of crisis. This might be due to the fact that the tweets get retweeted based on the actual content of the tweet rather than trending hashtags or URLs in it. If the tweet really contains some important information in it, it gets retweeted regardless of whether the tweet contains hashtags or URLs. In case of URLs in a tweet, the effect can be explained in a similar way. The effect persists even in the post-event time window. However, when we considered the effect of the interaction term HashTagXURL, the effect was positive in all the time windows, i.e., inclusion of both hashtag and URL in the tweet increases its probability of getting

Table 3.3: Regression Result with the Japan Earthquake Dataset

Variable	Pre-event	During event	Post-event
Number of Followers	.0002**	-.0002**	.000014
PeopleAware(t-1)	.00003	.00008***	.00013***
AmplifierScore	-1.10e-08	2.20e-06 ***	5.79e-07***
InformationStarterScore	-.0095***	.098***	.0254***
TransmitterScore	-0.127***	-0.036***	-0.040***
Early Retweeterers	-.698***	-4.11e-08***	-.614***
HashTag	1.646***	-2.55***	-.658***
URL	.849***	-1.708 ***	-2.053***
HashTagXURL	39.9 ***	2.646 ***	2.097***
Age	Present		
Tweet Count	Present		
Time of day	Present		
Day of the week	Present		

* - p < 0.10 , ** - p < 0.05, *** - p < 0.01

retweeted. In our dataset we observed that there were some hashtags, which got widely retweeted during the time of the event. For obvious reason, as a tweet grows older (i.e., age of a tweet), the retweet frequency decreases and it has been controlled in the model. We observed in our dataset that on average, the lifetime of a tweet is 1 day.

The event-centric (here earthquake) nature of our dataset allows us to systematically partition the time range into three distinct time windows (pre-, during-, and post-earthquake) and enables to understand the changes in the effects of the factors inherently for these three time periods. To check the effect of the event we investigate whether there is a significant difference in the retweet frequency in the three time periods. We used difference in difference estimator to compare the effect of the factors in different time windows. In the model we have considered the pre-event time period as the base for comparison. Compared to pre-event time window, number of followers have a negative impact on retweetability for both during and post-event time windows. On the other

Table 3.4: Effect of Event on Retweetability - the Japan Earthquake Dataset

Variable	Coefficient	$P > \|Z\|$	95% Confidence Interval	
FollowersXduringEvent	-.0005	0.00	-.0006	-.0004
FollowersXpostEvent	-.0002	0.005	-.0003	-.00006
AmplifierScoreXduringEvent	1.77e-06	0.000	1.74e-06	1.79e-06
AmplifierScoreXpostEvent	9.55e-07	0.000	9.28e-07	9.83e-07
Information-starterScoreXduringEvent	.153	0.001	.059	.246
Information-starterScoreXpostEvent	.0719	0.139	-.0233	.1673
HashTagXduringEvent	-14.151	0.000	-14.661	-13.640
HashTagXpostEvent	-11.528	0.000	-12.082	-10.975
URLXduringEvent	-6.985	0.000	-7.382	-6.589
URLXpostEvent	-8.697	0.000	-9.109	-8.284
Followers	.0003	0.000	.0001	.0004
HashTag	12.424	0.000	11.919	12.929
URL	5.882	0.000	5.489	6.274
AmplifierScore	2.86e-07	0.000	2.62e-07	3.10e-07
Information-starterScore	-.0909	0.082	-.1936	.0116
duringEvent	1.0859	0.000	.889	1.283
postEvent	2.834	0.000	2.623	3.044
Age	Present			
TweetCount	Present			
TimeOfDay	Present			
DayOfWeek	Present			

hand, in the during and post-event time window the *amplifier score* and *information-starter score* have a higher positive impact on retweetability. Similarly, inclusion of hashtags and URLs have negative impact in during-event time window.

All these aspects give us a signal that during the event the impacts of the factors affecting retweetability are very different in comparison with normal time. Interestingly, some of these effects have long term impact on retweetability in the post-event time window, e.g., inclusion of hashtags and URLs.

Furthermore, we have reexamined our model using a different Twitter dataset of the

Table 3.5: Regression Result with the Boston Marathon Bomb Blast Dataset

Variable	Pre-event	During event	Post-event
NumberFollowers	6.61e-08**	-5.91e-08	5.91e-08***
PeopleAware(t-1)	.00003***	.0002***	.00002***
AmplifierScore	2.45e-07***	3.10e-06***	1.77e-07***
InformationStarterScore	-2.88e-08***	-2.13e-06***	-1.61e-08***
TransmitterScore	-0.127***	-0.036***	-0.040***
EarlyRetweeterers	-.698***	-.736***	-.136***
HashTag	.01399***	-.609	-.02896***
URL	.0605***	-.0578	.00917
HashTagXURL	-.01399***	7.0403***	.02556***
Age	Present		
TweetCount	Present		
TimeOfDay	Present		
DayOfWeek	Present		

* - $p < 0.10$, ** - $p < 0.05$, *** - $p < 0.01$

Boston marathon bomb blast, which happened on 15th April, 2013. The dataset description for the event has been discussed in Section 3.4. Table 3.5 suggests that inclusion of hashtags and URLs have significant positive impact on the retweet frequency in the pre-event time window. However, this effect is not significant at the time of crisis. In the post-event time window the effect varies.

To verify the effect of the event we have investigated whether there is a significant difference in the retweet frequency in the three time periods (pre-, during-, and post-bomb blast). The impact of the *amplifier* is positive and impact of *information-starter* is negative for all the three time periods. Another interesting finding is that follower count of a user at the time of crisis is not of much importance. Users with a comparatively low number of users tend to participate in the information diffusion significantly.

Table 3.6: Correlation of Factors

		1	2	3	4	5	6	7	8	9	10
NumFollowers	1	1									
PageRank	2	**0.855**	1								
InformationStarter	3	0.453	**0.549**	1							
Amplifier	4	0.025	0.01	-0.009	1						
PositionInChain	5	0	-0.022	-0.031	0.229	1					
TweetCount	6	0	0.013	0.07	-0.041	-0.152	1				
NumDays	7	0.035	0.019	-0.006	0.468	0.114	-0.069	1			
HashTag	8	0.004	0.004	0.002	-0.065	-0.012	-0.019	-0.023	1		
URL	9	0.006	0.012	0.002	-0.067	-0.037	-0.023	-0.143	0.085	1	
peopleAware(lag 1)	10	-0.017	-0.02	-0.024	-0.097	-0.127	0.018	-0.112	-0.004	0.049	1

Table 3.7: Effect of Event on Retweetability - the Boston Marathon Bomb Blast Dataset

Variable	**Coefficient**	$P > \|Z\|$	**95% Confidence Interval**	
FollowersXduringEvent	-1.97e-07	0.319	-5.85e-07	1.91e-07
FollowersXpostEvent	-5.11e-09	0.980	-4.14e-07	4.03e-07
AmplifierScoreXduringEvent	2.96e-06	0.002	1.11e-06	4.82e-06
AmplifierScoreXpostEvent	-4.24e-09	0.996	-1.77e-06	1.76e-06
Information-starterScoreXduringEvent	-2.38e-06	0.000	-3.43e-06	-1.33e-06
Information-starterScoreXpostEvent	-9.73e-09	0.985	-9.92e-07	9.73e-07
HashTagXduringEvent	5.2802	0.000	4.5448	6.0156
HashTagXpostEvent	-.0695	0.839	-.7406	.6016
URLXduringEvent	4.6999	0.000	3.9857	5.4141
URLXpostEvent	-.0232	0.946	-.6920	.6456
Followers	7.91e-08	0.649	-2.61e-07	4.19e-07
HashTag	.0307	0.906	-.4805	.5419
URL	.0435	0.868	-.4686	.5556
AmplifierScore	2.07e-07	0.784	-1.27e-06	1.69e-06
Information-starterScore	-2.38e-08	0.951	-7.76e-07	7.28e-07
duringEvent	-.5541	0.066	-1.1456	.0372
postEvent	.1427	0.536	-.3089	.5943
Age	Present			
TweetCount	Present			
TimeOfDay	Present			
DayOfWeek	Present			

A direct comparison of the outcomes from both the datasets is shown in Table 3.8. From Table 3.8 it is apparent that most of the variables have similar effects in retweetability in all the time windows for both the datasets except for the HashtagXURL in the pre-time window and for *information-starter* in during- and post-event time windows. This agreement is quite significant considering the vast disparity in the nature of the two events.

Table 3.8: Comparison of the Japan Earthquake (E) and the Boston Blast (B)

Variable	**Pre-event**		**During event**		**Post-event**	
	E	**B**	**E**	**B**	**E**	**B**
NumFollowers	+	+	-	NS	NS	+
PeopleAware(t-1)	NS	+	+	+	+	+
AmplifierScore	NS	+	+	+	+	+
InformationStarterScore	-	-	+	-	+	-
TransmitterScore	-	-	-	-	-	-
PositionInChain	-	-	-	-	-	-
HashTag	+	+	-	NS	-	-
URL	+	+	-	NS	-	NS
HashTagXURL	+	-	+	+	+	+
Age	Present					
TweetCount	Present					
TimeOfDay	Present					
DayOfWeek	Present					

*NS = not significant $p >= 0.10$

Discussion on External Validity Twitter is one of the most popular microblogging services till date. While we can theoretically accumulate data for a million users base using this services Application Program Interface (API) (tools for software applications) on an international level, the key challenges for academicians are operationalization of empirical study using large datasets, and deciding appropriate sampling frame for studies. While Twitter provides clean and well-documented API for developers, the current rate-limit (15 requests per window per leveraged access token, refer Twitters REST API v1. 1 for more information) puts a boundary of accumulating the correct sample for the emerging research topic. As a result it becomes challenging to attain external validity.

According to Campbell and Stanley (1966) external validity can be conceived as “External validity asks the question of generalizability: To what populations, settings, treatment variables and measurement variables can this effect be generalized?

In our current research on Twitter, we cautiously performed the data collection and sampling. the data set collected is a big population of Japanese Twitter users which has been collected by tracking the users who have participated in the given time frame of our study. Our results can be generalized to the online behavior (in terms of the information diffusion on Twitter network) of the users in the three different time periods (as seen from two different datasets from Japan and Boston).

3.7 Summary

Retweet is the core mechanism for information diffusion on Twitter. In this work we have studied the retweet phenomenon to understand the factors affecting retweetability. Earlier research has shown that content factors like hashtags or URLs increase the likelihood for a tweet to get retweeted. However, our findings reveal that along with these content features, network features like how many people in the network are made aware (people aware) are very crucial. Users who are present at the beginning of the retweet chain (early retweeters) can make aware most of their followers for the first time and hence contribute largely in the diffusion process. Using datasets of two distinct events, the Great Eastern Japan earthquake and the Boston marathon bomb blast, we examine the effect of these factors in pre-, during-, and post-event time windows and the results obtained from both the datasets are in good agreement. While hashtags and URLs have significant positive impacts in pre-event time-window, during the event the effects are opposite. However, the inclusion of hashtags and URLs both in the tweet increases the probability of getting retweeted. These changes of effect of the factors in three time-periods demonstrate the influence of the event on retweetability and difference and difference estimator (DID) supports these findings. Further, the results show that during the event people do not necessarily retweet the users who have high in-degree. In fact, during the event low-indegree users participate in information diffusion significantly as compared to users with large number of followers.

The findings can be very much useful for targeting users in emergency events. Our results show that the users with less number of followers are the one participating actively in the time of the event, which is useful piece of information while targeting users. Moreover generalization of the results on a different dataset strengthen the usefulness of the results in a different country as well as users. The users in the network can be categorized well advanced in time (on regular basis) as discussed in this chapter to target and track the user activities during catastrophic events.

Chapter 4

Hashtag Popularity on Twitter: Analyzing Co-occurrence of Multiple Hashtags

4.1 Introduction

In August 2007, Chris Messina tweeted on his Twitter account ”how do you feel about using # (pound) for groups? As in #barcamp [msg]?” It was claimed as the first ever hashtag (Sweeney, 2012) on Twitter and since then this became a unique strategy for categorizing messages which can properly lead individuals to conversations and discussions pertaining to a specific topic (Doctor, 2012; Shirley, 2014). Social media is fast paced and no one has the time all day long to sift through his timeline to read everything being posted. That is where hashtags are significant. It can generate immediate, live, and interactive reactions and responses to specific topics. People use hashtags while watching their favorite TV program, listening to a debate on the radio, promoting a product, or running a campaign. It has been shown that when individuals used a hashtag within their tweet, engagement can increase as much as 100% and for brands it could get an increase of 50% (Cooper, 2013). This is because a hashtag immediately expands the reach of the tweet beyond followers of the tweet author and hence is reachable to anyone interested in that hashtag phrase or keyword.

A hashtag can collate similar ideas under one thread so that Twitterers get a more targeted user experience instead of just running through thousands of random unrelated tweets. It can be used to run a contest on Twitter (or other social networks) and create a wider market for brands. Moreover, creating hashtags on Twitter can improve one's 'following to follower' ratio, which has been extensively considered as a measure of user influence in the Twitter world. Originally this hashtag concept was Twitter exclusive, but the popularity surrounding such a small symbol has made other platforms, such as Facebook, Instagram, etc., realize its significance. Images in Instagram that include hashtags get more likes than the ones with no hashtags (Zarella, 2014).

During the World Cup, Olympics, and World series hashtags are considered as valuable as 30-second commercials (Fixmer, 2014). Some of the biggest advertisers like Kia, Volkswagen, Marriott, Johnson & Johnson, etc., created hashtag campaigns to reach viewers during sports. TV shows also promote their own hashtags and the world leaders use them to rally conversations. In U.S.A. Twitter is now charging companies $200,000 a day to buy a promoted trend (Kafka, 2013). This amount is more than twice the amount ($80,000) when promoted trends were introduced back in 2010. Companies who purchase a promoted trend get a customized hashtag placed at the top of the list of trending topics (Fiegerman, 2013; Doctor, 2013). Clicking on this hashtag shows a tweet from that company at the top of the search result page. Big brands like Coke, Disney, and Hyundai have purchased promoted trends over the years. The promoted trend lets an advertiser insert its own message atop the "trending topics" list on Twitter.com home pages and also on Twitter apps. During the 2012 presidential election, both Obama and Romney used hashtags to campaign through social media. The craze for hashtags is so high that people are willing to pay even $3,000 to rent a "social media wedding concierge" (Hathaway, 2014).

Can anybody legally own a hashtag? Till date the answer is no (Sweeney, 2012). However, one can register a hashtag in Twubs. Twubs.com is an online directory of hashtags (Twub, 2013). The registry at Twubs helps to minimize the possibility that your newly created hashtag is already in use by some other organizations and also to prevent another company from using the hashtag you 'own'. However, Twub cannot guarantee that your hashtag will not be squatted on. The popularity of the social network sites ensures that Twitter, Google+, or Facebook will hardly disappear in the near future. Similarly, it will be hard to believe that any of these social media platforms will turn off the hashtag functionality. Mostly for the younger generation hashtag use is as natural and common as typing their query in Google.

From the preceding discussion, it is transparent that the importance of hashtags is enormous, which motivates us to investigate the characteristics of these hashtags.

The abundance of information to which we are exposed through online social networks exceeds the amount of information we can consume. Hence, the hashtags compete with each other to attain our limited attention. Users can remember a bounded number of different hashtags at a time, which suggests that one hashtag is remembered by the users at the expense of others (Weng et al., 2012). How many users will adopt a hashtag determines its popularity. This adoption solely depends on how people find it meaningful and attractive which is concluded from their metacognitive experience. Metacognitive experiences are those experiences that are related to the current, on-going cognitive endeavor while metacognition refers to a level of thinking that involves active control over the process of thinking that is used in learning situations (Reber et al., 1998a; Reber and Schwarz, 1999). The detailed discussion of metacognition can be found in the literature review section.

On inspecting tweets containing hashtags, one can notice that hashtags usually come in groups, i.e., a single tweet contains more than one hashtag. Also a preliminary analysis of our data set reveals that tweets containing multiple hashtags get diffused more compared to tweets having a single hashtag. Here the decisive question arises whether the characteristics of the hashtags appeared together are random or it carries certain pattern, which is the focus of this study. So the first research question addressed in this chapter is

Does co-occurrence of hashtags increase popularity of a focal hashtag?

The popularity of one hashtag might boost the popularity of others when they appear together. For instance, say hashtag *h* becomes trendy on Twitter. Now, users start using

h with h_1 which increases the discoverability of h_1 also. In such circumstances, there can be three main possibilities: a) popularity of h takes off further, b) hashtag h_1 becomes more popular, and c) hashtag h_1 replaces hashtag h. To understand this phenomenon, it is necessary to investigate the change of popularity of a hashtag h when co-appeared with other hashtags h_1, h_2, etc. We investigate the popularity of a hashtag measured by the number of distinct users who have adopted / used it and model the popularity using regression technique considering both network variables and content variables of hashtag.

If co-occurrence of hashtags increases a focal hashtags's popularity, the second question arises, which hashtags should co-occur together? So the second research question addressed herein is

Which hashtags should co-occur together?

Here we investigate the nature of these co-appearing hashtags in terms of similarity. Moreover, we want to investigate if any additional information like URL moderates the effect of similarity/dissimilarity on hashtag popularity. So we posit our third research question as

How does presence of URL moderate hashtag popularity?

To address the above mentioned research questions, our study models hashtag popularity and investigates the moderating effect of URL on hashtag popularity. Drawing from the concept from metacognitive experience, we explained the moderating effect of URL inclusion. Dissimilar hashtags increase the metacognitive difficulty of the users (Pocheptsova et al., 2010), but when used with URLs it adds more information and brings surprisingness to the tweet, which in turn increases the popularity of hashtags. Earlier studies (Hughes and Palen, 2009) have shown that when hashtags appeared with a URL in a tweet, retweetability of that tweet escalates which is in line with our hypothesis.

This study makes several empirical contributions in the literature of product marketing. First, to the best of our knowledge to examine the effect of hashtag co-occurrence on its popularity. Secondly, the moderating effect of URL on the relationship of dissimilarity and hashtag popularity has been realized. The findings will be helpful for the product advertisers to implement effective marketing strategy while broadcasting product related tweets. Moreover, since hashtags are now popular on other social medias, these findings will help to promote even in other social medias.

The rest of the chapter has been presented as follows: the next section presents a brief literature review, and then we describe the data collection and preparation. Afterwards, we discuss the model and finally, we summarized our findings.

4.2 Literature Review

What does motivate people to share information? Sharing information with friends is considered to be a communal act in online social network sites. People share YouTube videos, Facebook posts, or tweets on Twitter. While a massive amount of information gets generated online, only a handful of them get noticed and shared. This leads to the straightforward question, "what makes a piece of content more share-worthy than others". Researches have been carried in the viral-marketing area to unfold the characteristics of the content that goes viral (Aral et al., 2009; Berger and Milkman, 2012, 2010). However, the main query lies in why people share information in the first place and what type of content gets shared. Consumers might share some content online for several reasons, e.g., altruistic reasons (e.g., to help others) or for self-enhancement purposes (e.g., to appear knowledgeable, see Wojnicki and Godes (2008)). Herein, we discuss the literature on social influence and self-presentation followed by word-of-mouth communication and viral marketing. Finally, we discuss literature related to meta-cognitive experience.

Social Influence and Self-presentation

Toubia and Stephen (2012) experimented the image-related vs. intrinsic motivations to contribute content in social media like Twitter. Intrinsic motivation is defined as "the doing of an activity for its inherent satisfactions rather than for some separable consequence" (Ryan and Deci, 2000). Image-related motivation, on the other hand, assumes users are motivated by the perception of others. Image-related motivation is also related to status seeking or prestige motivation (Glazer and Konrad, 1996; Fershtman and Gandal, 2007; Lampel and Bhalla, 2007). It was shown that on Twitter, intrinsic motivation to post content predicts that users post more as their numbers of followers (i.e., their audience) increase. On the other hand, image-related motivation leads to the prediction that users should derive less marginal utility from additional followers as their numbers of followers (a measure of stature) increase, and therefore users should have less motivation to post content. However, as group size grows, individual contribution levels decline (Zhang and Zhu, 2011). Theoretical models based on pure altruism generally support the hypotheses. Chen et al. (2012) have shown that indegree of a user have a positive impact on the intrinsic interest in broadcasting information more. The information sharing theory posits that in order to increase organizational or personal benefits people share information (Constant et al., 1994). Using information sharing theory Jarvenpaa and Staples (2000) claimed that a user's perceived usefulness of the information arouses information sharing behavior on collaborative electronic media because a user's expectation of the beneficial outcomes from the information (i.e., usefulness of the information) escalates the amount it is used and shared. Continuing the same line, Ha and Ahn (2011) showed that individuals' perceptions of the argument quality and source credibility of a received tweet play a major role in their information sharing behavior via the perceived level of usefulness of the information. Additionally, a URL in a tweet moderates the impact of argument quality on users' attitudes toward received tweets. On the other hand, some people rely on other's action to reiterate. People are characterized by herd behavior; i.e., people will be doing what others are doing rather than using their own information (Banerjee, 1992; Zhang, 2010; Asch, 1956).

Word-of-mouth Communication and Viral Marketing

Word-of-mouth (WOM) plays an important role in driving sales. Godes and Mayzlin (2004a) found that WOM is helpful for driving sales that occur between acquaintances (not friends) and is created by non-loyal customers rather than loyal. WOM is commonly measured by counting, i.e., volume of WOM generated. The authors have shown that dispersion is a good predictor of future sales, where dispersion has been measured by the entropy instead of variance (Godes and Mayzlin, 2004b). They also find that higher volume has no impact on TV show ratings, but a higher WOM dispersion is associated with higher future ratings for the show. In some cases, the impact of negative reviews is greater than the impact of positive reviews (Chevalier and Mayzlin, 2006). In the era of microbloging, Twitter became one of the popular sites for campaigning, product advertisement (Jansen et al., 2009; Shi et al., 2014), etc. Jansen et al. (2009) found that 19% of the tweets contain brand mentions and 20% of the tweets contain brand sentiments, which suggest that microblogs can be good resources for brand imaging and influencing a large population through a microblog. Maintaining the presence in the microblogs and managing the brand perception are very important for brand campaign. Researches (Canright and Engø-Monsen, 2006; Kossinets and Watts, 2006; Chwe, 2000) suggest that while the content of the tweet is important, the network structure and the positions of the users in the network play a critical role in information diffusion. Low dimension or strong link networks are better for coordination than the high dimension or weak link networks (Chwe, 2000). Also, these network structures evolve over time, which is dominated by the network topology and organizational structure (Kossinets and Watts, 2006). On Twitter, eWOM diffusion happens through retweet mechanism. Retweet of a tweet refers to the re-sharing of the same content by the followers of the users. A line of research has been carried out to understand the retweet functionality and its effect on Twitter users, especially who are the users having much impact on the information diffusion process (Cha et al., 2010; Watts and Dodds, 2007; Shuai et al., 2012; Aral and Walker, 2012; Bakshy et al., 2011). Cha et al. (2010) have measured the influence of users over a variety of topics and showed that users can hold influence in several topics. According to the findings of this study if a user has millions of followers, that does not mean that the user is influential in Twitter world. Overall, indegree measures user's popularity while retweet and mention shows the user's influence in the network. However, where influential users are important in the diffusion process, large cascades of diffusion happen by a critical mass of easily influenced users (Watts and Dodds, 2007; Shuai et al., 2012). Aral and Walker (2012) have carried out an experiment with Facebook users to find the influential users and the users susceptible to influence. Their findings show that highly influential individuals tend not to be susceptible, highly susceptible individuals tend not to be influential, and almost no one is both highly influential and highly susceptible to influence. This implies that influential individuals are less likely to adopt the product as a consequence of natural influence processes (i.e., in the absence of targeting).

Metacognitive Experience

Human reasoning is accompanied by metacognitive experiences. The assumptions about what makes it easy or difficult to think of certain things or to process new information contribute to what exactly people conclude from their metacognitive experiences. Researches showed evidence that people are more likely to advocate a statement as true when the color in which it is printed makes it easy to read (e.g., Reber and Schwarz

(1999); Reber et al. (1998a)). Schwarz (2004) describes that accessibility and processing fluency both pertain to the ease of recalling and processing new information. Moreover, repeated exposures lead to the subjective feeling of perceptual fluency, which in turn influences liking (Reber et al., 1998b). On the other hand, Pocheptsova et al. (2010) experimentally showed that metacognitive difficulty increases the attractiveness of a product by making it appear unique or uncommon.

In this work we want to investigate why some hashtags go more viral than others? A hashtag is a word or phrase preceded by a hash sign (#), used on Twitter to identify messages on a specific topic. This works as a user-defined index term to link several topics or events together. Yang et al. (2012a) examined the dual effect of hashtags on Twitter: a) a symbol of a community membership and b) a bookmark. In this paper they investigated which of the two reasons strive people to adopt a hashtag. The prediction using SVM technique incorporates social network variables like indegree, outdegree of nodes (number of people retweeted the hashtag), relevance, popularity of the hashtags, length (number of characters), age of the hashtag. The dataset used in this study was Twitter data on politics.

Popularity of the hashtag determines how many users will adopt a particular hashtag. Using a 25 week Twitter data, Tsur and Rappoport (2012) reported hashtag frequency prediction on a weekly basis using regression technique. Features used in the regression model were extracted from the hashtag itself (e.g., number of characters in the hashtag) and their experiment shows that hashtag popularity can be predicted using only the content features of the hashtags instead of using the costly graphical features extracted from tweets. However, Ma et al. (2013) claimed that contextual features are more effective than content features which can be explained by the fact that community graph plays an important role in information diffusion. Clarity of hashtag, number of words in a hashtag, user count, tweet count, etc. were used to predict the popularity. However, on Twitter a large number of hashtags are generated every day and people cannot remember all of them. Using an agent-based simulation model, Weng et al. (2012) claimed that the users can remember a bounded number of different memes at a time, which suggests that one meme is remembered by the users at the expense of others. The proposed retweet model assumes the finite memory of the users where memes are registered and by the friend and follower links, some other users can read the meme posted. However, a careful investigation of the usage of hashtags needs to be done. On inspecting tweets containing hashtags, one can notice that hashtags usually come in groups, i.e., a single tweet contains more than one hashtag. A preliminary analysis on our dataset reveals that tweets containing multiple hashtags get diffused more than tweets having a single hashtag. It will be interesting to investigate "Are these characteristics of the hashtags appeared together random or does it carry certain patterns?" Moreover, in the time of emergency, the adoption of hashtags might change. Using a 2011 Japan earthquake data, this chapter investigates what factors impact popularity of hashtags; more importantly the moderating effect of URL inclusion on the relationship of hashtag dissimilarity and hashtag popularity.

4.3 Solution Intuition

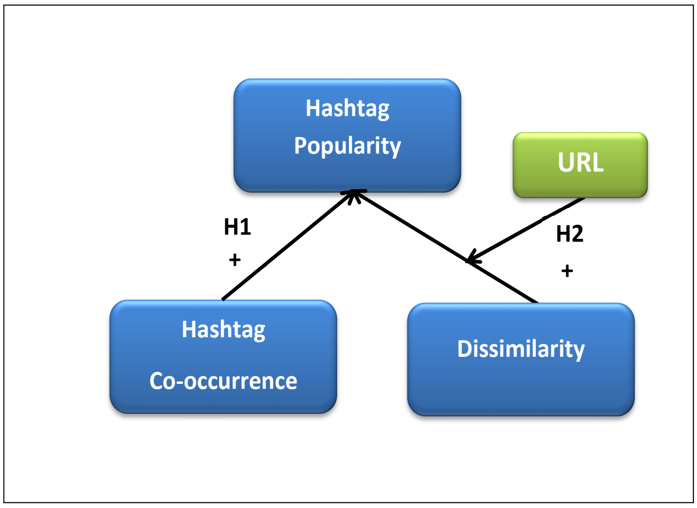
Hashtags used on Twitter are keywords or phrases preceded by # character, which help to categorize tweets into different topics. Moreover, a hashtag facilitates to get the tweet discoverable in the Twitter search result unless otherwise set private. While researchers have focused on finding the popularity of a single hashtag, an intent investigation reveals that when a hashtag appears with other hashtags, it inflates its popularity. In order to understand the moderating effect of URL on the relationship of hashtag similarity/dissimilarity on popularity, we measured the distance of a focal hashtag with the ones it appeared with in a tweet. Network variables along with the hashtag content variables are used to model the popularity of hashtags using regression technique. Retweet network among the users is constructed to calculate the network variables.

4.4 Dataset Description

In this study, we have used data set from the 2011 Japan earthquake. The details of the dataset description are described in Chapter 3.

The dataset covers a period of 20 days (from 5^{th} March, 2011 to 24^{th} March, 2011), and consists of 362,435,649 tweets posted by 2,711,473 users in Japan.

Figure 4.1: Research Model for Hashtag Popularity



4.5 Solution Details

4.5.1 Building Research Model and Hypotheses

In this section, we propose the research framework as in Figure 4.1 to examine the hashtag co-occurrence phenomenon. Further, we develop the research hypotheses to explain the factors affecting hashtag popularity. It has been seen that when more than one hashtag appear together in a tweet, the retweetability of the tweet is more compared to when the tweet contains only one hashtag. With this observation, we want to investigate whether the popularity of a hashtag is influenced by the co-occurrence of multiple hashtags.

Moreover, it has been observed that the hashtags usually comes in groups. However, it is not known whether it is effective to add more hashtags in a tweet. Intuitively, it can be conceptualized that more hashtags makes a tweet more discoverable and hence, the resultant popularity of a hashtag increases. Therefore, we posit our first hypothesis as follows:

Hypothesis 1. *Hashtag popularity increases when it appears with other hashtags.*

Hashtags, when appeared together, it will increase the visibility of the tweet to manyfold. However, when the hashtags are similar, the metacognitive difficulty decreases, hence the conclusion drawn from the metacognitive experience results in positive outcome Schwarz (2004). This suggests that when hashtags are similar the hashtag popularity increases. On the other hand, dissimilar hashtags will increase the metacognitive difficulty of the users (Pocheptsova et al., 2010), hence the hashtag popularity decreases; but when used with URLs it would add more information, bring surprisingness to the tweet, and could increase the popularity of hashtags. Thus, we postulate our second hypothesis:

Hypothesis 2. *Presence of URLs positively moderates the relationship between dissimilarity of co-appearing hashtags and hashtag popularity.*

We have investigated what happens when the hashtags co-appear. First, we examine whether the co-occurrence of hashtags plays any role in hashtag popularity and then we calculate the distance among the co-appearing hashtags to test whether the distance among the tags has any impact on its popularity. Additionally, we have also considered the interaction effect of the URL and the distance among the co-appearing hashtags. We have modeled hashtag popularity using the content variables of hashtags and the user specific variables. We have presented two models, one with only the hashtag specific variables, and another model with the hashtag specific and dyad specific variables. The first model examines the hashtag popularity where popularity has been defined as the total number of distinct users who used the hashtag. However, to verify whether this adoption is user-specific, we have modeled hashtag popularity at the dyad level. Both these models are described in the following section.

4.5.2 Factors considered for hashtag popularity

To investigate the factors impacting popularity hashtag specific, dyad specific, and control variables are considered as follows:

Hashtag Specific Variables

Length of hashtag: The hashtag has been extracted from the tweet content by searching words that start with #. For all hashtags we counted, the number of characters in that hashtag. Very long hashtags are not economical in the Twitter perspective as tweets are limited to only 140 characters. On the other hand, very small hashtags (e.g., abbreviated hashtags containing only 2 or three letters) do not contain sufficient information to understand.

Number of words: Clarity of the hashtag is important for its adoption. Hashtags, which contain multiple words are easy in order to follow the context from the hashtag itself. However, finding the word segments from a hashtag in the Twitter context is not straightforward as Twitter users use Twitter specific lingual. For the same reason, we counted the number of words in the hashtag by separating the capital letters or other special separator characters (e.g., underscore (_), plus (+) etc.).

Contains Capital Letters: This is a boolean variable computing the presence of capital letters in the hashtag. The value of the variable = 1, if the hashtag contains capital letters and 0 otherwise.

Contains Digits: This is a boolean variable denoting the presence of digits in the hashtag. The value of the variable = 1, if the hashtag contains digits and 0 otherwise.

Contains Other Separators: This is a boolean variable computing the presence of other separators in the hashtag, e.g., underscore (_), plus (+). The value of the variable = 1, if the hashtag contains other separators and 0 otherwise.

Appeared with Other Hashtag: This determines whether a hashtag appeared with other hashtags or not. If the hashtag appears with other hashtags then the value of the variable is the number of hashtags it appeared with and 0 otherwise. This is a time series variable indicating that the value of the variable determines whether the hashtag appeared with others or not in a particular time unit.

Distance: For the co-appearing hashtags, we compute the distance between the hashtag pairs. If more than two hashtags appear with the focal hashtag then the average distance of the hashtag pairs are considered. We describe the distance calculation between a hashtag pair.

Distance Calculation: For calculating distance we measured the distance in two different ways. Each of them are discussed in turn.

Levenshtein distance: “The Levenshtein distance is a string metric for measuring the difference between two sequences. Informally, the Levenshtein distance between two words is the minimum number of single-character edits (i.e., insertions, deletions, or substitutions) required to change one word into the other.” (Wikipedia, 2014a)

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases} \tag{4.5.1}$$

where $1_{(a_i \neq b_j)}$ is the indicator function equal to 0 when $a_i = b_j$ and equal to 1 otherwise.

Contains Other: It has been observed in our dataset that many co-occurred hashtags are substring of another. We have checked if string a contains string b and then measured the distance in number of characters. The pseudo code is given in Algorithm 4.

Finally, distance considered here is calculated as the minimum of the two distances discussed above, i.e.,

$distance(a,b) = Minimum(LevenshteinDistance(a,b), Contains(a,b))$

Inclusion of URLs: Earlier studies (Suh et al., 2010) have shown that inclusion of URLs in the tweet increases a tweet’s retweetability. Our previous study also supports this

ALGORITHM 4: Distance calculation: Contains other

```
Input: String a, String b
Output: Distance d
Initialize $d \leftarrow 250$ /*sufficiently large in the Twitter context*/
a = a.toLowerCase()
b = b.toLowerCase()
if (length(a) == length(b) and a==b) then
    d = 0;
end
if length(a) < length(b) and b contains a then
    d = length(b) - length(a);
end
if length(a) > length(b) and a contains b then
    d = length(a) - length(b);
end
return d
```

finding (as described in the previous chapter, Chapter 3). Moreover, we also observed that the presence of both hashtags and URLs in the tweet increases its popularity. Therefore, we compute this variable as a boolean variable denoting the presence of URLs in the tweet. URL = 1 if the tweet contains a URL, 0 otherwise. If a hashtag appears in more than one tweet, we compute the average number of times the focal hashtag appeared with URLs. We place URL = 1 if the average number of tweets > 0, 0 otherwise.

DistanceXURL: To examine the moderating effect of the URL on distance of co-appearing hashtags, we compute the interaction variable of distance and boolean URL.

$$DistanceXURL = distance \times URL$$

Frequency of Hashtag: For each hashtag *h* we calculate the frequency of hashtag as the number of times *h* has been retweeted per minute.

Age of Hashtag: For each hashtag h we compute the age of the hashtag since it has been used by some user. Unit of time used here is an hour.

Dyad Specific Variables

Frequency of Dyad: For each hashtag h we calculate the frequency of hashtag at the dyad level. Hence, dyad frequency (per minute) is computed as the number of times user $u_{retweeter}$ retweets a tweet by u_{author}, that contains hashtag h.

PageRank of Author and Retweeter: Each user on Twitter has a number of followers and followees which can be thought of as incoming and outgoing links from a web page. Similar to web pages, we can also compute the PageRank of a user to enumerate his popularity. However, in our case we formulated the retweet network of the users where direction indicates the reverse direction of information flow from (retweeter → author). Instead of using the PageRank computed on the follower-followee network, we computed the PageRank based on the retweet network. In this case, unlike otherwise, computed PageRank determines the activeness and actual influence of the users.

For both author and retweeter of the tweet, we compute the PageRank ($PageRank_{author}$ and $PageRank_{retweeter}$).

Betweenness Centrality of Author and Retweeter: Betweenness centrality is a measure of a node's centrality in a network. It is equal to the number of shortest paths from all vertices to all others that pass through that node. We have measured the betweenness centrality of the users on the retweet network.

For both author and retweeter of the tweet, we compute the betweenness centrality

($betweenness_{author}$ and $betweenness_{retweeter}$).

Relationship between dyad (Author and retweeter): On Twitter, a tweet can be retweeted by author's followers or friends. However, if a tweet becomes popular, this can be retweeted by retweeters even if they do not have any relationship with the author of the tweet.

Below are two control variables used in the model:

Day of Week: Day of the week (TweetSmarter, 2011) might have an impact on the popularity of the hashtag. TweetSmarter (2011) finds that day of the week controls traffic on Twitter, while Monday to Thursday the tweet volume increases, Friday it slows down. On Mondays users usually use Monday specific hashtags more frequently (#Monday, #mondayfever). On the other hand, on Saturdays and Sundays people write more fun-filled hashtags like #supersunday, #saturdaysale.

Time of the day: Twitter gets the most traffic during 9am-3pm from Monday to Thursday (TweetSmarter, 2011). We also include this as a control variable in the popularity model.

Model Specifications

Model 1: To model the factors affecting the popularity of the hashtags, we have considered the variables described in Table 4.1.

Here, we define the popularity of a hashtag as *total number of distinct users who have adopted/ used the hashtag.*

The regression technique has been used to model popularity of a hashtag. The

Table 4.1: Variables Affecting Hashtag Popularity

	Variable	**Meaning**
Dependent Variable	numDistinctUsers	number of distinct users who adopted/used the hashtag
Network Variables	PageRank	average PageRank of the users using the hashtag in retweet network
	betweenness	average betweenness centrality of the users using the hashtag in retweet network
Content Variables	hasCaps	value = 1 if the hashtag contains capital letters, = 0, otherwise
	hasDigits	value = 1 if the hashtag contains digits, = 0, otherwise
	hasOther	value = 1 if the hashtag contains other separators, = 0, otherwise
	numWords	number of words in the hashtag
	length	length of the hashtag
	appearedWithOthers	number of hashtags #h appeared with
	distance(h,H)	average distance of #h with all hashtags in H
	isURL	boolean variable indicating if the tweet contains URLs
Control Variables	timeOfDay	24 hours have been divided into 5 time-windows, morning(7am-10am), noon (11am-3pm), afternoon (4pm-7pm), evening (8pm-11pm), night (12am -6am)
	dayOfWeek	day of the week is coded as dummy variable
	tagAge	time since the tweet is composed (in hour)
	tagFreq	frequency of the hashtag per unit time (minute)
Dyad Specific Variables	dyadFreq	dyad (retweeter → author) frequency per unit time (measured in minute)
	relationship	boolean variable denoting follower-followee relationship, value = 1 if relationship exists and 0 otherwise

dependent variable in the model has been computed as the number of distinct users who have used the hashtag in their tweet (per hour). Popularity model is given below and correlations among the variables are reported in Table 4.9.

$$numberDistinctUsers_{i,t} = \beta_1 appearedWithOther_{i,t} + \beta_2 PageRank_{i,t}(u) + \beta_3 betweenness_{i,t}(u)$$

$$+\beta_4 hasDigits_i + \beta_5 hasCaps_i + \beta_6 numWords_i + \beta_7 numWords_i{}^2 + \beta_8 length_i + \beta_9 length_i{}^2$$

$$+\beta_{10} distace_{i,t} + \beta_{11} isURL_{i,t} + \beta_{12} isURL_{i,t} \times distace_{i,t} + \beta_{13} age_{i,t} + \beta_{14} timeOfday_{i,t} + \beta_{15} dayOfWeek_{i,t} + \epsilon$$

Model 2: In this model the dependent variable is the retweet count of tweets containing a specific hashtag for a specific dyad (retweeter → author pair).

$RetweetCount_{i,j,t} = \alpha + \sum_i H(i,t) + \sum_j D(j,t) + \sum_k C(k,t) + \epsilon$

$H(i,t), D(j,t), C(k,t)$ refer to the vector of hashtag specific variables, dyad specific variables, control variables respectively.

$$H(i,t) = [hasDigits,\ hasCaps,\ numWords]'_i + [distance,\ appearedWithOthers,\ isURL,\ distanceXisURL]'_{i,t}$$

$$D(j,t) = [Pagerank_{author},\ betweenness_{author},\ Pagerank_{retweeter},\ betweenness_{retweeter}]'_{j,t} + [relationship]_j$$

$$C(k,t) = [dayOfWeek,\ timeOfDay,\ age]'_{k,t}$$

4.6 Data Analysis and Findings

The Great Eastern Japan earthquake dataset has been used to examine this phenomena. The dataset consists of 1.3 million observations with 521028 hashtags from 0.1 million users. The model investigates the effect of URLs in hashtag popularity at two levels - first at the hashtag level and second at the dyad (user-retweeter pair) level. After examining the model at the hashtag level, we wanted to verify if user-specific variables have any impact on the adoption of the hashtags. This is the reason we have used the dyad level model as well.

4.6.1 Data Preparation

Using the Twitter dataset described in Chapter 3, we found the hashtags from each tweet by simply searching words that start with "#". From the primary tweet dataset we prepared a dataset where each row contains the timestamp of the tweet, the list of hashtags in the tweet, author of the tweet, boolean variable indicating whether the tweet contains URLs. Our tweet dataset (5^{th}-24^{th} March) has been divided into three time windows, pre-earthquake ($5^{th} - 10^{th}\ March$), during-earthquake ($11^{th} - 16^{th}\ March$), and post-earthquake ($17^{th} - 24^{th}\ March$).

For our analysis we have prepared two sets of data, one to understand the popularity (measured by distinct number of users who have used the hashtag) of hashtag and second to analyze the impact at a granular level, i.e, at the dyad (retweeter-user pair) level (here the popularity of hashtags is measured by the retweet count at the dyad level). Overall, two models have been verified, one with the dependent variable as the number of distinct users using the hashtag and in the other model, retweet count of the focal hashtag from a retweeter to a user.

Table 4.2: Summary Statistics in Pre-event Time Window

Variable	Mean	Std. Dev.	Min	Max
numDistinctUsers	2.839	20.006	1	3156
length	8.268	3.565	1	139
numWords	1.494	0.678	1	33
appearedWithOthers	2.001	26.064	0	4657
$appearedWithOthers^2$	683.354	48094.31	0	2.17E+07
hasCaps	0.233	0.423	0	1
hasDigits	0.272	0.445	0	1
hasOther	0.088	0.284	0	1
$numWords^2$	2.690	3.534	1	1089
$length^2$	81.060	127.287	1	19321
PageRank	7.72E-07	0.0002	0	0.188
betweenness	90.785	12536.74	0	5929623
isURL	0.624	0.484	0	1
distance	3.278	4.112	0	97

Number of Observations = 1687085

4.6.2 Data Analysis

We formed the retweet network from the tweets in our database, where the nodes represent the users and directed links represent the reverse of direction (retweeter → user) of information flow. Network variables such as PageRank, betweenness centrality are measured using the retweet network.

Besides network variables, hashtag contents have been analyzed. Since tweets are limited to 140 characters, each character in the tweet is very costly. Therefore, very long hashtags are not preferable. Moreover, long and complex hashtag increases the cognitive load and are not easy to understand (Song and Schwarz, 2008, 2009). For the same reason, the hashtag is analyzed and the number of words in the hashtag is counted. The intuition behind this is that the number of words in a hashtag increases its clarity and the hashtag itself carries more contextual information about the tweet

Table 4.3: Summary Statistics in During-event Time Window

Variable	Mean	Std. Dev.	Min	Max
numDistinctUsers	6.712	99.008	1	19683
length	8.344	4.188	1	139
numWords	1.476	0.679	1	31
appearedWithOthers	4.835	83.043	0	13817
$appearedWithOthers^2$	6919.431	564206.2	0	1.91E+08
hasCaps	0.222	0.416	0	1
hasDigits	0.234	0.424	0	1
hasOther	0.110	0.313	0	1
$numWords^2$	2.639	3.800	1	961
$length^2$	87.155	243.944	1	19321
PageRank	0.000	0.000	0	0.252
betweenness	367.534	59587.630	0	3.29E+07
isURL	0.583	0.493	0	1
distance	3.440	4.166	0	112

Number of Observations = 1298383

itself. If the words are separated by special characters or by capital letters it is easy to determine the words in the hashtag.

There tends to be more than one hashtag in a tweet. A preliminary analysis has shown that if a hashtag appears with others, popularity of the focal hashtag increases. Moreover, we included the distance among the co-appearing hashtags to examine the effect of distance on its popularity. Previous studies have experimented that inclusion of URLs and hashtags increase the chance of retweetability (Boyd et al., 2010). As mentioned in the earlier chapter (Chapter 3), in our dataset among the retweeted tweets, 26.5% of the tweets have URLs and 10.3% of the tweets contain hashtags. In this study, we have included the boolean variable for URL to examine its effect on similarity/dissimilarity of hashtags.

Next, we examine whether there is an effect of an event on popularity and Japan

Table 4.4: Summary Statistics in Post-event Time Window

Variable	Mean	Std. Dev.	Min	Max
numDistinctUsers	3.565	31.705	1	14207
length	8.418	3.812	1	139
numWords	1.506	0.681	1	31
appearedWithOthers	2.439	28.376	0	6213
appearedWithOthers^2	811.172	56732.450	0	3.86E+07
hasCaps	0.226	0.418	0	1
hasDigits	0.270	0.444	0	1
hasOther	0.101	0.302	0	1
numWords^2	2.731	3.493	1	961
length^2	85.389	159.338	1	19321
PageRank	0.000	0.000	0	0.005915
betweenness	4.891	1433.555	0	988199.9
isURL	0.617	0.486	0	1
distance	3.429	4.221	0	129

Number of Observations = 2171903

earthquake data is used for that reason. The effects of the variables are investigated in all the three time periods (pre-, during- and post-event time-windows) independently.

Summary statistics for the three time windows are shown below: Table 4.2, Table 4.3, Table 4.4.

4.6.3 Findings and Discussion

We describe the findings from both the models in turn.

Discussion of Model 1: To model the popularity of the hashtags, random effect GLS regression model is used (Table 4.5). Both the content variables and network variables are included in a hierarchical way to address our research questions. From Table 4.5 we can see that in general, when a hashtag appeared with other hashtags, then the popularity increases significantly (coefficient = 0.5479).

Table 4.5: Regression Results with Content and Network Variables

Variable	Coef.	P>z	[95% Conf.	Interval]
appearedWithOthers	0.5479	0.000	0.5471	0.5487
hasCaps	-0.5770	0.000	-0.6356	-0.5184
hasDigits	-0.9209	0.000	-0.9997	-0.8422
hasOther	0.9999	0.000	0.9054	1.0944
numWords	0.4197	0.000	0.3253	0.5140
length	-0.0059	0.237	-0.0158	0.0039
$numWords^2$	-0.0474	0.000	-0.0608	-0.0340
$length^2$	-0.0001	0.377	-0.0004	0.0001
PageRank	-5.4366	0.927	-121.203	110.3302
betweenness	2.2E-06	0.009	5.51E-07	3.92E-06

Table 4.6: Regression Results Examining Hashtag Similarity

Variable	Coef.	P>z	[95% Conf.	Interval]
appearedWithOthers	0.5481	0.000	0.5473	0.5489
hasCaps	-0.5725	0.000	-0.6311	-0.5139
hasDigits	-0.9271	0.000	-1.0059	-0.8483
hasOther	0.9798	0.000	0.8850	1.0746
numWords	0.4152	0.000	0.3209	0.5096
length	-0.0053	0.293	-0.0152	0.0046
$numWords^2$	-0.0468	0.000	-0.0602	-0.0334
$length^2$	-0.0001	0.369	-0.0004	0.0001
PageRank	-5.2308	0.929	-120.997	110.5352
betweenness	2.23E-06	0.009	5.45E-07	3.91E-06
distance	-0.0132	0.000	-0.0184	-0.0080

It is also clear from our dataset that length does not have a significant impact on popularity, however, the number of words in a hashtag has inverse u-shaped impact. While the number of words has a positive impact on popularity, too many words in a hashtag have a negative impact. Intuitively, this is comprehensible, as the number of words increases, the clarity of hashtag at first, but as the number of words grows in abundant the hashtag becomes complex. Further, the presence of digits or capital letters in the hashtag has negative impact on popularity, but popular hashtags mostly contain

Table 4.7: Regression Results Examining Inclusion of URLs on Similarity

Variable	Coef.	P>z	[95% Conf.	Interval]
appearedWithOthers	0.5470	0.000	0.5462	0.5479
hasCaps	-0.4396	0.000	-0.4983	-0.3809
hasDigits	-1.4176	0.000	-1.4983	-1.3370
hasOther	0.9241	0.000	0.8293	1.0188
numWords	0.4823	0.000	0.3880	0.5766
length	0.0015	0.761	-0.0083	0.0114
$numWords^2$	-0.0437	0.000	-0.0571	-0.0304
$length^2$	-5.9E-05	0.663	-0.0003	0.0002
PageRank	2.9426	0.960	-112.706	118.5914
betweenness	2.33E-06	0.007	6.48E-07	4.01E-06
distance	-0.0713	0.000	-0.0796	-0.0630
isURL	1.1646	0.000	1.1073	1.2218
distanceXisURL	0.0632	0.000	0.0525	0.0738

other separators to segregate the words in hashtag phrases.

Two network variables, namely PageRank and betweenness centrality have negligible impact. PageRank does not have a significant impact on popularity, but the betweenness centrality has significant positive impact, though the coefficient is negligible (coefficient = 2.2E-06,Table 4.5). This finding is inline with earlier work by Tsur and Rappoport (2012), where the authors have shown that network variables do not have a significant impact on the hashtag popularity, instead the content of the hashtag plays a vital role.

Next, we investigate the effects of similarity / dissimilarity of co-appearing hashtags on the popularity of the focal hashtag. For example, Twitter users include hashtags like "#HappyFriendshipDay" and "#FriendshipDay" together in a tweet and hence the tweet can be discoverable by more than one hashtag. It is easy to note that these two hashtags are similar. On the other hand, hashtags like "#HappyFriendshipDay" and

“#ContestAlert” are dissimilar and it is not easy to derive the context of the tweet from the pair of hashtags. We hypothesized that when similar hashtags appear together the popularity of the hashtag increases. From our findings, we can see that with the increase of distance with other co-appearing hashtags, the popularity of the focal hashtag decreases (coefficient = -0.0132), or in other words when a hashtag appears with similar hashtag, the popularity increases.

Afterwards, we examined the moderating effect of URLs on the similarity/ dissimilarity of co-appearing hashtags (Table 4.7). Inclusion of URLs in a tweet increases the popularity of the focal hashtag that appears with dissimilar hashtags. This phenomenon can be explained from the fact that dissimilarity among the hashtags introduces curiosity among the Twitter users, and the addition of more information through URLs clarify the meaning of the dissimilarity and in turn it appears surprising to the Twitter users. As a result, it gains more popularity.

Further, for the three distinct time periods, pre-event, during-event, and post-event time-windows regression models have been tested and the results are shown in Table 4.8. In all the time-windows, the effects of the variables on popularity (number of distinct people who adopted/used the hashtag) have similar trends. However, in the time of the event the coefficients of the independent variables are larger compared to pre-event time-window, which indicates that during the event there was a stronger effect of the independent variables on popularity.

Table 4.8: Interaction Effect of Dissimilarity and URL on Hashtag Popularity in Three Time Windows

Variable	Pre-event	During event	Post-event
appearedWithOthers	.547***	1.108***	0.773***
hasCaps	-.439***	-.203***	-0.702**
hasDigits	-1.417***	-.814 ***	-1.837***
hasOther	.924***	3.365***	.955***
length	0.002	.020	-.012**
$length^2$	-.00006	.0004**	.0004**
numWords	0.482***	.523***	.944***
$numWords^2$	-.044***	-.075 ***	-.068***
PageRank	2.943	103.43	1472.106***
betweenness	2.33e-06***	6.71e-08	1.14e-06
distance	-.0713 ***	-.146***	-.108***
isURL	1.165 ***	.585***	1.130***
distanceXisURL	.063 ***	.236***	.141**

* - $p < 0.10$, ** - $p < 0.05$, *** - $p < 0.01$

Table 4.9: Correlation Among the Variables

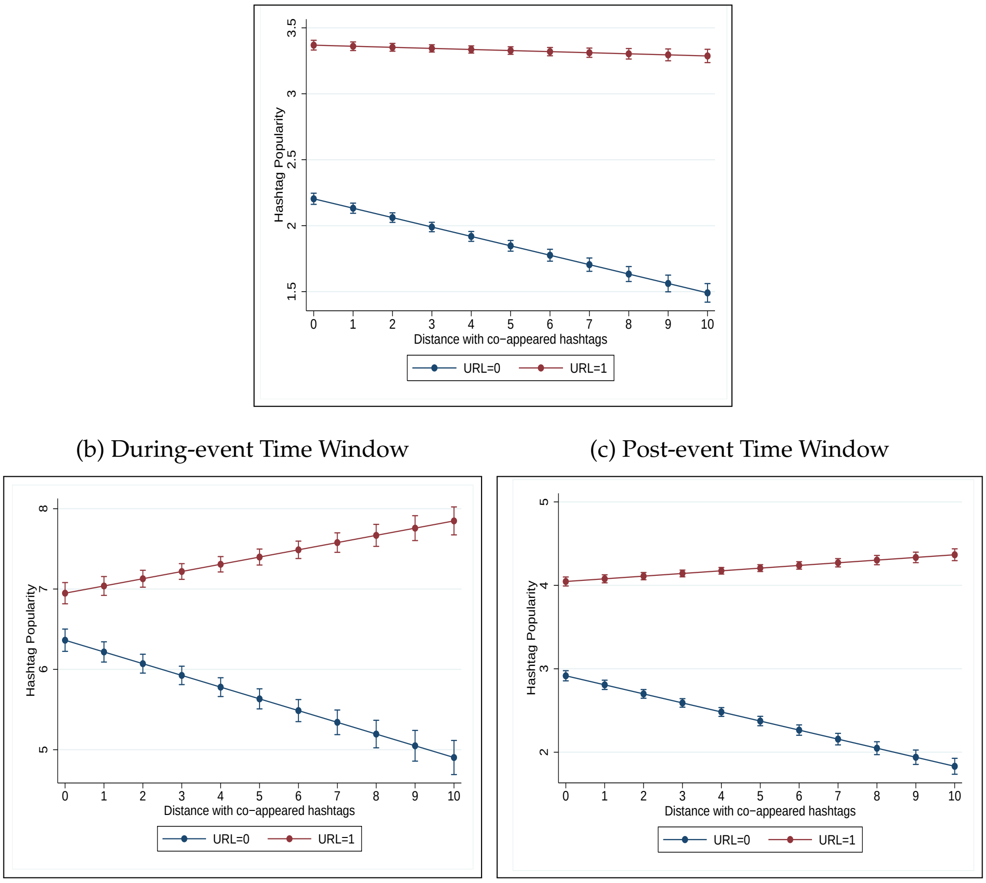
		1	2	3	4	5	6	7	8	9	10	11	12	13
appearedWithOthers	1	1												
hasCaps	2	-.014	1											
hasDigits	3	-.025	-.211	1										
hasOther	4	.001	-.049	-.096	1									
numWords	5	-.021	.131	.561	.396	1								
length	6	-.015	-.045	.137	.259	.445	1							
numWords^2	7	-.014	.122	.345	.289	.85	.416	1						
length^2	8	-.007	-.003	.021	.129	.247	.715	.352	1					
PageRank	9	0	.0001	-.0001	-.0004	-.0003	.0004	-.0002	0.0001	1				
betweenness	10	.0001	-.001	-.001	.003	.001	.001	.001	.001	0.0004	1			
distance	11	.054	.035	-.085	-.064	-.066	.009	-.028	.017	.0002	-.001	1		
isURL	12	.033	-.152	.308	-.083	.082	-.025	.019	-.043	-.001	-.001	.091	1	
distanceXisURL	13	.066	-.033	.027	-.085	-.054	-.047	-.045	-.029	-.0004	-.001	.710	.491	1

To determine whether the patterns characterizing the significant interactions conform to the directions as proposed in the research hypotheses, we have plotted the interaction effects (Figures 4.2a,4.2b, and 4.2c) for all three time-windows. This procedure was introduced by Cohen et al. (1983) for all interaction cases. Figures 4.2a, 4.2b, and 4.2c show the disordinal (or crossover) interaction of URLs on the relationships of hashtag similarity with hashtag popularity.

Figure 4.2a plots the interaction effect of URLs on distance in the pre-event time-window. The main effect of the presence of URLs can be seen by calculating the mean points in both red and blue lines (URL = 0 and URL = 1 respectively). It shows that when there is a URL in a tweet (along with a hashtag), the popularity of the hashtag is more compared to when there is no URLs in the tweet. To check the main effect of the hashtag similarity the mean points between the two lines are considered for high and low distance, which indicates that when the distance is low (i.e., the hashtags are similar), popularity is higher compared to when the hashtags are dissimilar. Examination of the interaction effect between the two reveals that the absence of a URL in a tweet when the hashtags are dissimilar leads to low popularity compared to the addition of a URL in it. However, the effect of URLs (presence or absence) in the popularity of hashtags does not differ much for similar hashtag co-occurrence.

Figure 4.2b plots the interaction effect of URLs on similarity with hashtag popularity. During earthquake both the main effects of URLs and hashtag similarity are significant as seen in pre-event time window. However, when the hashtags are similar, presence or absence of URLs does not have significant difference, but when the co-appearing

Figure 4.2: Interaction Plot on Distance and URLs in Pre-, During-, and Post-event Window (Hashtag Level)



(a) Pre-event Time window

(b) During-event Time Window

(c) Post-event Time Window

hashtags are dissimilar URLs play the critical role (and statistically significant) in popularity of hashtags. This asserts our hypothesis that the dissimilarity of hashtag increases the meta-cognitive load of Twitter users, which adversely affect the hashtag popularity. Tweets are of limited characters and shortened URLs provide more information about the tweets as well as the hashtags. Inclusion of URLs with dissimilar hashtags probably decreases the meta-cognitive load and helps in adoption/usage of the focal hashtag.

Similarly, Figure 4.2c shows the interaction effect in the post-event time-window.

As can be seen from the graph, the direction of the interaction is the same as in the two other windows. However, the trend of the effect of the presence of URLs on hashtag similarity tends to go back as in pre-event time-window.

Discussion of Model 2: To understand the effect of hashtag popularity at user level, we modeled retweetability of a hashtag for dyads, where each dyad consists of the user who tweeted and the one who is retweeted. The model is run with the hashtag (which we used in our previous model also) and dyad specific variables along with the control variables as listed earlier. Retweet count (per hashtag per dyad) is considered as the dependent variable. We have divided the dataset into three different time-windows and regression technique has been used in all cases. Results for the three time-windows have been shown in Table 4.10. The findings show that dyad specific and hashtag specific variables considered in the model have significant impacts on retweet count. The dyad frequency have significant positive impact on retweet count per unit time, which suggests that the users retweet hashtags from Twitter users they usually retweet from. Moreover, in the Twitter world if the user and retweeter has follower-followee relationship, then retweet count of a hashtag increases opposed to retweet practices from non-follower/friend relationship. Content variables of hashtags, like number of words in the hashtag, length, presence of digits and capital letters have similar impacts as we have observed in our previous models. On the other hand, while hashtag frequency (for a specific dyad) has a positive impact, which suggests that user retweets tweet containing specific hashtags many times. Above all, we have examined the interaction effect of hashtag dissimilarity with presence of URLs and we receive similar impact as

Table 4.10: Hashtag Popularity Model at the Dyad Level

Variable	Pre-event	During-event	Post-event
$dyadFreq$	0.001***	0.002***	0.005***
$length$	-0.004***	-0.004***	-0.004***
$numWords$	-0.006***	0.002	0.017***
$length^2$	0.001***	0.001***	0.001***
$numWords^2$	0.001**	0.000	-0.005***
$tweetCount$	0.009***	0.046***	0.016***
$hasCaps$	0.013***	0.010***	0.029***
$hasDigits$	-0.028***	-0.027***	-0.039***
$hasOther$	0.005***	0.024***	0.029***
$appearedWithOthers$	0.008***	0.003***	0.001***
$tagAge$	-7.7E-05***	-7E-05***	-5.3E-05***
$tagFreq$	3.63E-05***	1.07E-05***	9.83E-06***
$distance$	-0.001***	-0.004***	-0.002***
$isURL$	0.085***	0.093***	0.104***
$distancesXisURL$	0.004***	0.012***	0.006***
$PageRank_{author}$	0.037	0.214	-0.672
$betweenness_{author}$	0.001E-10***	0.001E-10***	-3.01E-10***
$PageRank_{retweeter}$	0.317	-2.173***	-1736.72***
$betweenness_{retweeter}$	0.001E-09***	0.001E-09***	-1.21E-09***
$relationship$	0.772***	0.868***	0.911***

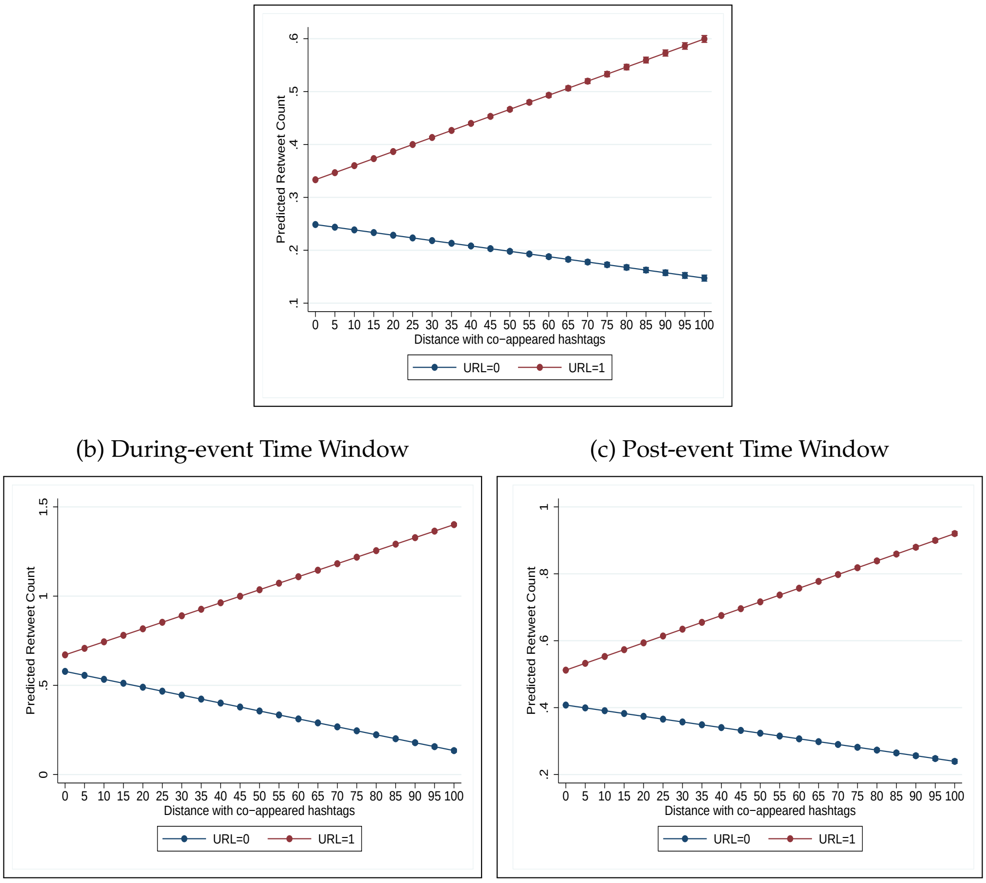
seen in model 1 at hashtag level.

Figure 4.3a plots the interaction effect of URLs on distance in the pre-event time-window at the dyad level. Similar to hashtag level analysis, in the dyad level also we notice the similar effect in the pre-event time window. It shows that when there is a URL in a tweet (along with a hashtag), the retweet count of that hashtag by a specific dyad is more compared to when there is no URLs in the tweet. In this case, when the distance among the co-appearing hashtags is higher, introduction of a URL results in higher retweet count, compared to when a URL appears with similar hashtags.

Figure 4.3b and Figure 4.3c plot the interaction effect of URLs on similarity with retweetability of a hashtag at dyad level in the during-event and post-event windows

Figure 4.3: Interaction Plot on Distance and URLs in Pre-, During-, and Post-event Window (Dyad Level)

(a) Pre-event Time Window



(b) During-event Time Window

(c) Post-event Time Window

respectively. Similar to hashtag level analysis, one can note that the retweet count at dyad level has similar result as in Figure 4.2b. The appearance of URLs when the hashtags are similar has significantly less impact compared to when the hashtags are dissimilar.

Overall, we can see that the presence of URLs with similarity (or dissimilarity) of hashtags has significant impact at dyad level, which suggests that choice of hashtag is driven by individual metacognitive experiences.

4.7 Summary

Hashtag in a tweet starts with a # symbol and is used before a relevant keyword or phrase in a tweet, which facilitates to categorize the tweets into different topics. Consequently, it becomes convenient to search them in a Twitter search. However, in practice hashtags mostly come in groups, i.e., one can find more than one hashtag in a tweet. Are these co-appearing hashtags random or do they carry certain patterns? In this study, we have investigated the characteristics of the co-appearing hashtags. Findings show that the popularity of a hashtag increases when a hashtag appears with other hashtags. Moreover, the similarity / dissimilarity of the hashtags plays crucial role in hashtag popularity. Results indicate that when similar hashtags appear together the hashtag popularity increases as opposed to dissimilar hashtags. To our surprise when the dissimilar hashtags appear with a URL, then the effect is reversed. This phenomenon can be explained by the fact that when dissimilar items co-appear it increases the meta-cognitive load and introduces confusion, but with the provision of extra information (e.g., URL), this becomes surprising and interesting to users resulting adoption of those hashtags together.

These findings can help to diffuse new hashtags by coupling with similar popular hashtags or adding the pinch of surprise with dissimilar hashtags and a URL. It also can help the practitioners implement efficient policies for product advertisement with brand hashtags.

Chapter 5

Conclusion

Electronic word of mouth (eWOM) has various prototypes. In this thesis, it has been investigated in two different contexts: a) product recommendation and b) information diffusion on social media. User reviews have been used to generate recommendations for emerging classes of products like mobile applications and Twitter data have been examined to understand the information diffusion on Twitter.

First, a novel approach has been described to generate mobile app recommendations for users. The proposed approach has been verified using a real world dataset of mobile applications, collected from Mobilewalla. To the best of our knowledge, this work is one of the first mobile app recommendation technique proposed. Results achieved from the algorithm ascertain the huge applicability of our system for marketing. Diversity of the mobile apps increases the quality of mobile applications. On Twitter, everyday a massive amount of tweets are generated, however, only a handful of them gets retweeted widely. The information primarily propagates through the retweet mechanism on Twitter. Understanding the factors contributing to the retweet phenomena is the key to address this issue. The impacts of these factors, specifically the user roles, have been investigated in Chapter 3. The concept of *Information Diffusion Impact (IDI)* has been introduced and three important user roles, namely "information starter", "amplifier", and "transmitter" have been identified. The effect of a major event on the factors affecting retweetability has also been investigated. The findings demonstrate that retweetability is significantly affected by *amplifiers* and *information-starters*. Further, due to an event, like earthquake, these effects change substantially.

In the third study, we have examined the Twitter dataset to investigate hashtag popularity. A hashtag in a tweet that starts with a "#" symbol was introduced originally by the Twitter users. Users use the hashtag symbol, "#" before a relevant keyword or phrase in their tweet, which facilitates to categorize the tweets into different topics. Consequently, it becomes convenient to search them in a Twitter search. If a hashtag is used extensively, it becomes a trending topic. However, this becomes possible only for selective hashtags. In Chapter 4, the evolution of these hashtags have been investigated to understand what contributes to its popularity. Findings show that when a hashtag appears with other hashtags popularity increases. On investigating the similarity of the co-appearing hashtags, it has been observed that when the hashtags are similar people use the hashtag more compared to when they are dissimilar. Interestingly, when the dissimilar hashtags are accompanied with extra information, e.g., URL, popularity again escalates.

Overall, this thesis deals with three independent studies pertaining eWOM diffusion in two different domains, mobile applications and social media analysis under the umbrella of Information System. With the emergence of web2.0, traditional WOM turned into even more effective channel of information broadcast. As compared to traditional WOM dissemination, a piece of information can be broadcasted very quickly through online social medias to a larger number of audiences. However, from a individual's perspective, the outcome is two folded - at one hand information is received at a rapid rate, on the other hand the information processing becomes a tedious process because of its enormous amount. In such circumstances, a system is necessary to get assistance to consume the information needed. A novel recommendation algorithm has been proposed in the domain of mobile applications. Herein, this proposes an effective way to solve the issue of information overload in a new domain (here we have used mobile apps domain, which can be easily adapted to other domains as well). We believe that the proposed technique will be useful for achieving a sustainable marketing strategy for online recommendation. However, in the proposed algorithm, additional user information like social network information about users have not been incorporated. Thus, it would be interesting to investigate how social information can be integrated into the user profiles to understand their product preferences. This would lead us to find the users in the community who share similar taste with the active user, for which there are now limited methods available, but will be very important in the future.

Moreover, in eWOM diffusion another important aspect is to identify the important users in the network for different objectives. Plethora of research has been carried out to find the influential users in the network for effective dissemination of information for product advertisement, crisis information information broadcast, etc. In recent times, product advertisement on social media has gained a lot of popularity simply because it is easier to achieve a wide market online. We have identified users using two emergency event-centric datasets. Here the question arises "how is this behavior different when non-emergency event becomes trending?" As a future direction it will be interesting to investigate the impact of user roles on non-emergency events like Christmas or FIFA World Cup. Moreover, the users' location information is not available in our dataset which can provide us the freedom for in-depth analysis of the user roles. It has been seen that multiple information from "eyes-on-the-ground" provides more detailed local context and frequent updates, useful for the ones who need to make decisions on how to act (Vieweg et al., 2010).

While influential users are essential to reach out the correct audience, it is also important to understand which kind of information attains larger attention. This thesis aims to provide guidelines in these two directions, which has been experimented and validated using big Twitter data set. Findings from this thesis will be beneficial for marketing strategy developer to optimize audience targeting and brand management through efficient hashtag inclusion in Twitter like miroblogs.

This page is left intentionally blank.

Bibliography

Abel, F., Gao, Q., Houben, G.-J., and Tao, K. 2011. "Analyzing user modeling on twitter for personalized news recommendations," in *Proceedings of the 19th international conference on User modeling, adaption, and personalization*, , UMAP'11, Berlin, Heidelberg: Springer-Verlag. 56

Adam Lella, A. L. 2014. "The U.S. Mobile App Report," .
URL `http://venturebeat.com/2013/07/10/state-of-the-apposphere/` 14

Adomavicius, G., and Kwon, Y. 2014. "Optimization-Based Approaches for Maximizing Aggregate Recommendation Diversity," *INFORMS Journal on Computing* (26:2), pp. 351–369. 20

Agrawal, R., Imieliński, T., and Swami, A. 1993. "Mining association rules between sets of items in large databases," *SIGMOD Rec* (22:2), pp. 207–216. 20

Agrawal, R., and Srikant, R. 1994. "Fast Algorithms for Mining Association Rules in Large Databases," in *Proceedings of the 20th International Conference on Very Large Data Bases*, , VLDB '94, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 20

Apache 2001. “Lucene,” . 27

Aral, S., Muchnik, L., and Sundararajan, A. 2009. “Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks,” *Proceedings of the National Academy of Sciences* (106:51), pp. 21,544–21,549. 106

Aral, S., and Walker, D. 2012. “Identifying influential and susceptible members of social networks,” *Science* (337:6092), pp. 337–341. 109

Asch, S. E. 1956. “Studies of independence and conformity: I. A minority of one against a unanimous majority,” *Psychological monographs: General and applied* (70:9), pp. 1–70. 107

Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. 2011. “Everyone’s an influencer: quantifying influence on twitter,” in *Proceedings of the fourth ACM international conference on Web search and data mining*, , ACM. 109

Banerjee, A. V. 1992. “A simple model of herd behavior,” *The Quarterly Journal of Economics* pp. 797–817. 107

Bartosz Ziolko, S. M., and Wilson, R. 2007. “Fuzzy Recall and Precision for Speech Segmentation Evaluation,” in *Proceedings of 3rd Language & Technology Conference, Poznan, Poland, October*, .

URL "http://www-users.cs.york.ac.uk/~suresh/papers/FRAPFSSE.pdf" 43

Bennett, J., Lanning, S., and Netflix, N. 2007. “The Netflix Prize,” in *In KDD Cup and Workshop in conjunction with KDD*, . 17

Berger, J., and Milkman, K. 2010. “Social transmission, emotion, and the virality of online content,” *Wharton Research Paper* . 106

Berger, J., and Milkman, K. L. 2012. “What makes online content viral?” *Journal of Marketing Research* (49:2), pp. 192–205. 106

Boyd, D., Golder, S., and Lotan, G. 2010. “Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter,” in *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, , HICSS ’10, Washington, DC, USA: IEEE Computer Society. 57, 88, 125

Brashers, D. E. 2001. “Communication and uncertainty management,” *Journal of Communication* (51:3), pp. 477–497. 60

Campbell, D. T., and Stanley, J. C. 1966. “Experimental and Quasi-Experiment Designs for Research,” *Rand Mc-Nally* . 97

Canright, G. S., and Engø-Monsen, K. 2006. “Spreading on networks: a topographic view,” *Complexus* (3:1-3), pp. 131–146. 108

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P. K. 2010. “Measuring User Influence in Twitter: The Million Follower Fallacy.” *ICWSM* (10), pp. 10–17. 109

Charlett, D., Garland, R., and Marr, N. 1995. “How damaging is negative word of mouth,” *Marketing Bulletin* (6:1), pp. 42–50. 2

Chatterjee, P. 2001. “Online reviews: do consumers use them?” *Advances in consumer research* (28:1). 2

Chen, Q., Phan, T. Q., and Goh, K.-Y. 2012. "Do Pepsi Drinkers Talk About Sleepwalker? The Effects of Self-Presentation and Conformity in Competing Word-of-Mouth," in *ICIS*, , Association for Information Systems. 107

Chevalier, J. A., and Mayzlin, D. 2006. "The effect of word of mouth on sales: Online book reviews," *Journal of marketing research* (43:3), pp. 345–354. 1, 108

Chu, Z., Gianvecchio, S., Wang, H., and Jajodia, S. 2010. "Who is tweeting on Twitter: human, bot, or cyborg?" in *Proceedings of the 26th Annual Computer Security Applications Conference*, , ACSAC '10, New York, NY, USA: ACM. 56, 62, 79

Chwe, M. S.-Y. 2000. "Communication and coordination in social networks," *The Review of Economic Studies* (67:1), pp. 1–16. 108

Cohen, J., Cohen, P., West, S., and Aiken, L. 1983. "Applied multiple regression/correlation analysis for the social sciences," *L Erlbaum Associates, Hillsdale, NJ* . 132

Constant, D., Kiesler, S., and Sproull, L. 1994. "What's mine is ours, or is it? A study of attitudes about information sharing," *Information systems research* (5:4), pp. 400–421. 107

Cooper, S. 2013. "Big Mistake: Making Fun Of Hashtags Instead Of Using Them," http://www.forbes.com/sites/stevecooper/2013/10/17/big-mistake-making-fun-of-hashtags-instead-of-using-them/ (last accessed 2nd August 2014). 101

Costello, S. 2014. "How Many Apps Are in the iPhone App Store?" http://ipod.about.com/od/iphonesoftwareterms/qt/apps-in-app-store.htm. 3, 14

Datta, A., Dutta, K., Kajanan, S., and Pervin, N. 2011. "Mobilewalla: A Mobile Application Search Engine," in *MobiCASE*, . 3, 14

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., and Sampath, D. 2010. "The YouTube video recommendation system," in *Proceedings of the fourth ACM conference on Recommender systems*, , RecSys '10, New York, NY, USA: ACM. 13, 17

Doan, S., Vo, B.-K. H., and Collier, N. 2012. "An analysis of Twitter messages in the 2011 Tohoku Earthquake," in *Electronic Healthcare*, , Springer, pp. 58–66. 61

Doctor, V. 2012. "What's The Point Of All These Hashtags?" http://www.hashtags.org/platforms/twitter/whats-the-point-of-all-these-hashtags/ (last accessed 2nd August 2014). 100

Doctor, V. 2013. "What Does It Take for a Hashtag to Trend?" http://www.hashtags.org/business/management/what-does-it-take-for-a-hashtag-to-trend/(last accessed 2nd August 2014). 102

Fershtman, C., and Gandal, N. 2007. "Open source software: Motivation and restrictive licensing," *International Economics and Economic Policy* (4:2), pp. 209–225. 106

Fiegerman, S. 2013. "Report: Twitter Now Charges $200,000 For Promoted Trends," http://mashable.com/2013/02/11/report-twitter-now-charges-200000-for-promoted-trends/ (last accessed 2nd August 2014). 102

Fixmer, A. 2014. "World Cup Advertisers Trade Commercials for Hashtags," http://www.socialmediatoday.com/content/can-you-legally-own-twitter-hashtag (last accessed 2nd August 2014). 101

Frederic, L. 2010. "The short lifespan of a tweet: retweets only happen in the first hour," . 84, 85

GaggleAMP 2013. "The 48 hours life of a tweet," . 84, 85

Girvan, M., and Newman, M. E. J. 2002. "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences* (99:12), pp. 7821–7826. 73

Glazer, A., and Konrad, K. A. 1996. "A signaling explanation for charity," *The American Economic Review* pp. 1019–1028. 106

Godes, D., and Mayzlin, D. 2004a. "Firm-created word-of-mouth communication: A field-based quasi-experiment," *HBS Marketing Research Paper* (16:04-03). 107

Godes, D., and Mayzlin, D. 2004b. "Using online conversations to study word-of-mouth communication," *Marketing Science* (23:4), pp. 545–560. 108

Goldberg, D., Nichols, D., Oki, B. M., and Terry, D. 1992. "Using collaborative filtering to weave an information tapestry," *Commun ACM* (35:12), pp. 61–70. 17

Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A. 2004. "Information diffusion through blogspace," in *Proceedings of the 13th international conference on World Wide Web*, , ACM. 56

Guha-Sapir, D., and Lechat, M. F. 1986. "Information systems and needs assessment in natural disasters: an approach for better disaster relief management," *Disasters* (10:3), pp. 232–237. 60

Ha, S., and Ahn, J. 2011. "Why Are You Sharing Others' Tweets?: The Impact of Argument Quality and Source Credibility on Information Sharing Behavior," in *ICIS*, , Association for Information Systems. 107

Hale, J. E., Dulek, R. E., and Hale, D. P. 2005. "Crisis Response Communication Challenges Building Theory From Qualitative Data," *Journal of Business Communication* (42:2), pp. 112–134. 60

Hathaway, J. 2014. "Now You Can Get a $3,000 "Social Media Concierge" For Your Wedding," . 102

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., and Gremler, D. D. 2004. "Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet?" *Journal of interactive marketing* (18:1), pp. 38–52. 1, 2

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. 2004. "Evaluating collaborative filtering recommender systems," *ACM Trans Inf Syst* (22:1), pp. 5–53. 17

Huang, Z., and Zeng, D. D. 2011. "Why does collaborative filtering work? transaction-based recommendation model validation and selection by analyzing bipartite random graphs," *INFORMS Journal on Computing* (23:1), pp. 138–152. 20

Hughes, A., and Palen, L. 2009. "Twitter adoption and use in mass convergence and emergency events," *International Journal of Emergency Management* (6:3), pp. 248–260. 56, 60, 61, 105

Inose, N. 2011. "How Tokyo responded to request for help on Twitter," http://www.nikkeibp.co.jp/article/column/20110314/263638/ (last accessed 18th April 2011). 61

Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A. 2009. "Twitter power: Tweets as electronic word of mouth," *Journal of the American society for information science and technology* (60:11), pp. 2169–2188. 108

Jarvenpaa, S. L., and Staples, D. S. 2000. "The use of collaborative electronic media for information sharing: an exploratory study of determinants," *The Journal of Strategic Information Systems* (9:2), pp. 129–154. 107

Jiang, H., and Sun, H. 2012. "Choice-Based Recommender Systems: A Unified Approach to Achieving Relevancy and Diversity," *Available at SSRN 1989238* . 20

Kafka, P. 2013. "Twitter Hikes Its Promoted Trend Prices Again, to $200,000 a Day," http://allthingsd.com/20130209/twitter-hikes-its-promoted-trend-prices-again-to-200000-a-day/ (last accessed 2nd August 2014). 101

Khoshneshin, M., and Street, W. N. 2010. "Incremental collaborative filtering via evolutionary co-clustering," in *Proceedings of the fourth ACM conference on Recommender systems*, , RecSys '10, New York, NY, USA: ACM. 18

Kim, Y. S., and Yum, B.-J. 2011. “Recommender system based on click stream data using association rule mining,” *Expert Syst Appl* (38:10), pp. 13,320–13,327. 15, 20

Koren, Y. 2010. “Factor in the neighbors: Scalable and accurate collaborative filtering,” *ACM Trans Knowl Discov Data* (4:1), pp. 1:1–1:24. 18

Kossinets, G., and Watts, D. J. 2006. “Empirical analysis of an evolving social network,” *Science* (311:5757), pp. 88–90. 108

Krishnamurthy, B., Gill, P., and Arlitt, M. 2008. “A few chirps about twitter,” in *Proceedings of the first workshop on Online social networks*, , ACM. 4

Kwak, H., Lee, C., Park, H., and Moon, S. 2010. “What is Twitter, a social network or a news media?” in *Proceedings of the 19th international conference on World wide web*, , WWW ’10, New York, NY, USA: ACM. 4, 62

Lampel, J., and Bhalla, A. 2007. “The role of status seeking in online communities: Giving the gift of experience,” *Journal of Computer-Mediated Communication* (12:2), pp. 434–455. 106

Lancichinetti, A., Radicchi, F., Ramasco, J. J., and Fortunato, S. 2010. “Finding Statistically Significant Communities in Networks,” *CoRR* (abs/1012.2363:5). 73

Lekakos, G., and Caravelas, P. 2008. “A hybrid approach for movie recommendation,” *Multimedia Tools Appl* (36:1-2), pp. 55–70. 13

Leskovec, J., Adamic, L. A., and Huberman, B. A. 2007. “The dynamics of viral marketing,” *ACM Transactions on the Web (TWEB)* (1:1), p. 5. 70

Leung, C. W.-k., Chan, S. C.-f., and Chung, F.-l. 2006. “A collaborative filtering framework based on fuzzy association rules and multiple-level similarity,” *Knowl Inf Syst* (10:3), pp. 357–381. 20

Li, B., Yang, Q., and Xue, X. 2009. “Can movies and books collaborate?: cross-domain collaborative filtering for sparsity reduction,” in *Proceedings of the 21st international jont conference on Artifical intelligence*, , IJCAI’09, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 19

Linden, G., Smith, B., and York, J. 2003. “Amazon.com Recommendations: Item-to-Item Collaborative Filtering,” *IEEE Internet Computing* (7:1), pp. 76–80. 13, 17

Ma, Z., Sun, A., and Cong, G. 2013. “On predicting the popularity of newly emerging hashtags in twitter,” *Journal of the American Society for Information Science and Technology* (64:7), pp. 1399–1410. 111

Mendoza, M., Poblete, B., and Castillo, C. 2010. “Twitter under crisis: can we trust what we RT?” in *Proceedings of the First Workshop on Social Media Analytics*, , SOMA ’10, New York, NY, USA: ACM. 56

Miller, B. N., Riedl, J. T., and Konstan, J. A. 1997. “Experiences with GroupLens: Making Usenet Useful Again,” in *Proceedings of the 1997 Usenix Winter Technical Conference*, . 17

mobiForge 2014. “Global mobile statistics 2014 Home: all the latest stats on mobile Web, apps, marketing, advertising, subscribers, and trends," http://mobiforge.com/research-analysis/. 14, 18

Naveed, N., Gottron, T., Kunegis, J., and Alhadi, A. C. 2011. "Bad News Travel Fast: A Content-based Analysis of Interestingness on Twitter," in *WebSci '11: Proceedings of the 3rd International Conference on Web Science*, . 62

Ogiue, K. 2011. "An examination of rumors about the Great East Japan Earthquake. Kobunsha, Tokyo," . 61

Okada, I., and Yamamoto, H. 2011. "Effects of Information Diffusion in Online Word-of-Mouth Communication Among Consumers." *JACIII* (15:2), pp. 197–203. 2

Pan, S. L., Pan, G., and Leidner, D. 2012. "Crisis response information networks," *Journal of the Association for Information Systems* (13:1), pp. 31–56. 60

Papagelis, M., Rousidis, I., Plexousakis, D., and Theoharopoulos, E. 2005. "Incremental Collaborative Filtering for Highly-Scalable Recommendation Algorithms," in *Proceedings of the 15th International Symposium on Methodologies of Intelligent Systems (ISMIS'05)*, . 18

Pastor-Satorras, R., and Vespignani, A. 2001. "Epidemic spreading in scale-free networks," *Physical review letters* (86:14), p. 3200. 60

Pavlov, D. Y., and Pennock, D. M. 2002. "A maximum entropy approach to collaborative filtering in dynamic, sparse, high-dimensional domains," in *Advances in Neural Information Processing Systems*, . 45

Peres, R., Shachar, R., and Lovett, M. J. 2011. "On brands and word-of-mouth," *Available at SSRN 1968602* . 1

Perez, S. 2014. "Mobile App Usage Increases In 2014, As Mobile Web Surfing Declines," http://techcrunch.com/2014/04/01/mobile-app-usage-increases-in-2014-as-mobile-web-surfing-declines/. 14

Pervin, N., Fang, F., Datta, A., Dutta, K., and Vandermeer, D. 2013. "Fast, scalable, and context-sensitive detection of trending topics in microblog post streams," *ACM Transactions on Management Information Systems (TMIS)* (3:4), p. 19. 56

Pocheptsova, A., Labroo, A. A., and Dhar, R. 2010. "Making products feel special: When metacognitive difficulty enhances evaluation," *Journal of Marketing Research* (47:6), pp. 1059–1069. 5, 104, 110, 114

Prawesh, S., and Padmanabhan, B. 2014. "The Most Popular News Recommender: Count Amplification and Manipulation Resistance," *Information Systems Research* (25:3), pp. 569–589. 20

Reber, R., and Schwarz, N. 1999. "Effects of perceptual fluency on judgments of truth," *Consciousness and cognition* (8:3), pp. 338–342. 103, 109

Reber, R., Winkielman, P., and Schwarz, N. 1998a. "Effects of perceptual fluency on affective judgments," *Psychological science* (9:1), pp. 45–48. 103, 110

Reber, R., Winkielman, P., and Schwarz, N. 1998b. "Effects of perceptual fluency on affective judgments," *Psychological science* (9:1), pp. 45–48. 110

Richardson, B. 1994. “Socio-technical disasters: profile and prevalence,” *Disaster Prevention and Management* (3:4), pp. 41–69. 60

Rui, H., and Whinston, A. 2011. “Designing a social-broadcasting-based business intelligence system,” *ACM Transactions on Management Information Systems (TMIS)* (2:4), p. 22. 50

Ryan, R. M., and Deci, E. L. 2000. “Intrinsic and extrinsic motivations: Classic definitions and new directions,” *Contemporary educational psychology* (25:1), pp. 54–67. 106

Sarwar, B., Karypis, G., Konstan, J., and Rield, J. 2000. “Analysis of Recommendation Algorithms for E-Commerce,” . 15, 20

Schubert, P., Uwe, L., and Risch, D. 2006. “Personalization beyond recommender systems,” in *Project E-Society: Building Bricks*, , Springer, pp. 126–139. 18

Schwarz, N. 2004. “Metacognitive experiences in consumer judgment and decision making,” *Journal of Consumer Psychology* (14:4), pp. 332–348. 110, 114

Sellnow, T. L., and Seeger, M. W. 2013. *Theorizing crisis communication*, vol. 4, John Wiley & Sons. 60

Shannon, C. E. 2001. “A mathematical theory of communication,” *ACM SIGMOBILE Mobile Computing and Communications Review* (5:1), pp. 3–55. 45

Shi, Z., Rui, H., and Whinston, A. B. 2014. “Content sharing in a social broadcasting environment: evidence from twitter,” *Mis Quarterly* (38:1), pp. 123–142. 108

Shirley, T. 2014. “Why Hashtags Are So Important,” http://www.thelasthurdle.co.uk/hashtags-important/ (last accessed 2nd August 2014). 100

Shuai, X., Ding, Y., and Busemeyer, J. 2012. “Multiple spreaders affect the indirect influence on Twitter,” in *Proceedings of the 21st international conference companion on World Wide Web*, , ACM. 109

Song, H., and Schwarz, N. 2008. “If It’s Hard to Read, It’s Hard to Do Processing Fluency Affects Effort Prediction and Motivation,” *Psychological Science* (19:10), pp. 986–988. 10, 124

Song, H., and Schwarz, N. 2009. “If It’s Difficult to Pronounce, It Must Be Risky Fluency, Familiarity, and Risk Perception,” *Psychological Science* (20:2), pp. 135–138. 124

Statistica 2014. “Number of apps available in leading app stores as of July 2014,” http://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/. 14

Suh, B., Hong, L., Pirolli, P., and Chi, E. H. 2010. “Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network,” in *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, , SOCIALCOM ’10, Washington, DC, USA: IEEE Computer Society. 57, 58, 62, 82, 88, 117

Sweeney, D. 2012. “Can You Legally Own a Twitter Hashtag?”

http://www.socialmediatoday.com/content/can-you-legally-own-twitter-hashtag (last accessed 2nd August 2014). 100, 102

Sysomos 2009. "An In-Depth Look at the 5% of Most Active Users," http://www.sysomos.com/insidetwitter/mostactiveusers/. 84

Tachiiri, K. 2011. "An examination of the Great East Japan Earthquake: What did social media report? Discover Twenty One, Tokyo," . 61

Takács, G., Pilászy, I., Németh, B., and Tikk, D. 2009. "Scalable Collaborative Filtering Approaches for Large Recommender Systems," *J Mach Learn Res* (10), pp. 623–656. 18

Tinati, R., Carr, L., Hall, W., and Bentwood, J. 2012. "Identifying communicator roles in twitter," in *Proceedings of the 21st international conference companion on World Wide Web*, , WWW '12 Companion, New York, NY, USA: ACM. 68

Toriumi, F., Sakaki, T., Shinoda, K., Kazama, K., Kurihara, S., and Noda, I. 2013. "Information sharing on Twitter during the 2011 catastrophic earthquake," in *Proceedings of the 22nd international conference on World Wide Web companion*, , WWW '13 Companion, Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. 65

Toubia, O., and Stephen, A. T. 2012. "Intrinsic versus Image-Related Motivations in Social Media: Why Do People Contribute Content to Twitter?" Tech. rep., working paper, Columbia University. 106

Tsur, O., and Rappoport, A. 2012. "What's in a Hashtag?: Content Based Prediction of

the Spread of Ideas in Microblogging Communities," in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, , WSDM '12, New York, NY, USA: ACM.
URL http://doi.acm.org/10.1145/2124295.2124320 88, 91, 110, 128

TweetSmarter 2011. "The Ultimate Guide to Finding the Best Time to Tweet," http://blog.tweetsmarter.com/retweeting/when-is-the-best-time-to-tweet/. 84, 87, 120

Tweney, D. 2013. "Mobile app growth exploding, and shows no signs of letting up," .
URL http://venturebeat.com/2013/07/10/state-of-the-apposphere/ 14

Twub 2013. "Twubs-Register your hashtag," http://twubs.com/ (last accessed 2nd August 2014). 102

Vargas, S., and Castells, P. 2011. "Rank and relevance in novelty and diversity metrics for recommender systems," in *Proceedings of the fifth ACM conference on Recommender systems*, , RecSys '11, New York, NY, USA: ACM. 15

Vieweg, S., Hughes, A. L., Starbird, K., and Palen, L. 2010. "Microblogging during two natural hazards events: what twitter may contribute to situational awareness," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, , ACM. 60, 61, 141

Watts, D. J., and Dodds, P. S. 2007. "Influentials, networks, and public opinion formation," *Journal of consumer research* (34:4), pp. 441–458. 109

Weng, L., Flammini, A., Vespignani, A., and Menczer, F. 2012. "Competition among memes in a world with limited attention," *Scientific Reports* (2). 103, 111

Wikipedia 2011. "List of foreshocks and aftershocks of the 2011 Tohoku earthquake," . 65

Wikipedia 2014a. "Levenshtein distance," . 117

Wikipedia 2014b. "Netflix," http://en.wikipedia.org/wiki/Netflix. 17

Wikipedia http://en.wikipedia.org/wiki/2011_Tohoku_earthquake_and_tsunami, visited on March. 5, 2015a. "2011 Great Eastern Japan Earthquake," . 64

Wikipedia http://en.wikipedia.org/wiki/Boston_Marathon_bombings, visited on March. 5, 2015b. "2013 Boston Marathon bombings," . 67

Wojnicki, A. C., and Godes, D. 2008. "Word-of-mouth as self-enhancement," *HBS marketing research paper* . 106

Wu, S., Hofman, J. M., Mason, W. A., and Watts, D. J. 2011. "Who says what to whom on twitter," in *Proceedings of the 20th international conference on World wide web*, , ACM. 60

Yang, J., and Counts, S. 2010. "Predicting the Speed, Scale, and Range of Information Diffusion in Twitter," . 62

Yang, L., Sun, T., Zhang, M., and Mei, Q. 2012a. "We know what@ you# tag: does the dual role affect hashtag adoption?" in *Proceedings of the 21st international conference on World Wide Web*, , ACM. 110

Yang, X., Zhang, Z., and Wang, K. 2012b. “Scalable collaborative filtering using incremental update and local link prediction,” in *Proceedings of the 21st ACM international conference on Information and knowledge management*, , CIKM ’12, New York, NY, USA: ACM. 18

Zarella, D. 2014. “New Data Shows the Importance of Hashtags on Instagram,” http://danzarrella.com/new-data-shows-the-importance-of-hashtags-on-instagram.html (last accessed 2nd August 2014). 101

Zhang, J. 2010. “The sound of silence: Observational learning in the US kidney market,” *Marketing Science* (29:2), pp. 315–335. 107

Zhang, M., and Hurley, N. 2008. “Avoiding monotony: improving the diversity of recommendation lists,” in *Proceedings of the 2008 ACM conference on Recommender systems*, , RecSys ’08, New York, NY, USA: ACM. 15, 44

Zhang, M., and Hurley, N. 2009. “Novel Item Recommendation by User Profile Partitioning,” in *Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, , WI-IAT ’09, Washington, DC, USA: IEEE Computer Society. 15

Zhang, X., and Zhu, F. 2011. “Group size and incentives to contribute: A natural experiment at Chinese Wikipedia,” *The American economic review* (101:4), pp. 1601–1615. 107

Ziegler, C.-N., Lausen, G., and Schmidt-Thieme, L. 2004. “Taxonomy-driven computation of product recommendations," in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, , CIKM '04, New York, NY, USA: ACM. 15, 19, 20

Zook, M., Graham, M., Shelton, T., and Gorman, S. 2010. "Volunteered geographic information and crowdsourcing disaster relief: a case study of the Haitian earthquake," *World Medical & Health Policy* (2:2), pp. 7–33. 60

This page is left intentionally blank.

Appendix A

List of Publications

Refereed Conference and Workshop Publications

- Datta, A., Dutta, K., Kajanan, S., **Pervin, N.**, "A Mobile App Search Engine", *Mobile Computing, Applications, and Services (MobiCASE)*, 2011
- Fang, F., **Pervin, N.**, Datta, A., Dutta, K., VenderMeer, D. "Detecting Twitter Trends in Real Time", *Workshop on Information Technology and Systems (WITS)*, 2011
- Kajanan, S., **Pervin, N.**, Ramasubbu, N., Dutta, K., Datta, A., "Takeoff and Sustained Success of Apps in Hypercompetitive Mobile Platform Ecosystems: An Empirical Analysis", *International Conference on Information Systems (ICIS)*, 2012
- Ramasubbu, N., Kajanan, S., **Pervin, N.**,Dutta, K., Datta, A."Surviving Hyper-Competitive, Unforgiving Platform Ecosystems: Examining Developer Strategies in iOS and Android Marketplaces." *Workshop on Information Technology and Systems (WITS)*, 2012

- **Pervin, N.**, Datta, A., Dutta, K. "Towards Generating Recommendations on Large Dynamically Growing Domains.", *Pacific Asia Conference on Information Systems (PACIS)*, 2013

- Cazabet, R.,**Pervin,N.**, Toriumi, F., Takeda, H. "Information Diffusion on Twitter: everyone has its chance, but all chances are not equal", *International Conference on Signal-Image Technology and Internet-Based Systems (SITIS)*, 2013

- **Pervin, N.**, Cazabet, R., Datta, A., Toriumi, F., Takeda, H. "User Roles in the time of Crisis: A social Media Analysis", *Workshop on Information Technology and Systems (WITS)*, 2013

- **Pervin, N.**, Takeda, H., Toriumi, F. "Factors Affecting Retweetability : An Event-Centric Analysis on Twitter", *International Conference on Information Systems (ICIS)*, 2014

- **Pervin, N.**, Phan, T.Q.,Datta, A., Takeda, H., Toriumi, F. "Hashtag Popularity on Twitter: Analyzing Co-occurrence of Multiple Hashtags", *HCI International*, 2015, *Forthcoming*

Refereed Journal Publications

- **Pervin, N.**, Fang, F., Datta, A., Dutta, K., VanderMeer, D. "Fast, Scalable, and Context-Sensitive Detection of Trending Topics in Microblog Post Streams." *ACM Transactions on Management Information Systems*, Vol. 3, No. 4, Article. 19, 2013

- Datta, A., Dutta, K., Kajnan, S., **Pervin, N.** "A mobile app search engine." *ACM*

Mobile Networks and Applications (MONET), Vol. 18, No. 1, pages: 42-59, 2013

Submitted Journal Publications

- **Pervin, N.**, Kajanan, S., Ramasubbu, N., Dutta, K., Datta, A.,"Takeoff and Survival of Apps in Mobile Platform Ecosystems: An Empirical Analysis of iOS and Android Apps" 2014. Submitted to *Journal of Management Information Systems* (*Revise and Resubmit*).

- **Pervin, N.**, Dutta, K., Datta, A.,"Generating Scalable and Diverse Recommendations for Mobile Apps" 2014. Submitted to *INFORMS Journal on Computing*.