

Comment Mining, Popularity Prediction, and Social Network Analysis

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

By

Salman Jamali
Bachelor of Science
FAST - National University of Computer & Emerging Sciences, 2006

Director: Huzefa Rangwala, Assistant Professor
Department of Computer Science

Spring Semester 2010
George Mason University
Fairfax, VA

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Allah Almighty for all of His blessings, and for making my life beautiful and bountiful. I beg for His mercy, I seek His protection, and I submit to His will. May He be with all of us in the most difficult times to come.

I am deeply indebted to my advisor, Dr. Huzefa Rangwala. His constant guidance, persistence, support, insightful feedback, constructive critique, and mentoring has been invaluable. I would not have envisaged a more dedicated mentor, where I relied on his expertise for a solid research direction, research writing etiquette, my deepest debugging troubles, and beyond.

Many thanks to Dr. Angelos Stavrou and Dr. Daniel Barbará for the time and effort they spent reviewing my research and the helpful comments they contributed to the success of this thesis. I am grateful to Dr. Gheorghe Tecuci and Dr. Mihai Boicu for dedicatedly mentoring me and helping me to exploit my research potential.

I express my deepest gratitude to my parents for their extraordinary and lively encouragement. This dissertation was simply impossible without the continuous support, love, and kindness of my sister, Maria Jamali, and my dearest friend and brother-in-law, Hamza Ahmad.

I cannot list all of my friends, and I simply cannot thank them enough even for just being my friends. For as long as I live, I wish to be with them in our happiest and saddest moments of life.

TABLE OF CONTENTS

	Page
List of Tables	v
List of Figures	vi
Abstract	vii
Chapter 1: Introduction	1
1.1. Motivation.....	2
1.2. Problem Hypothesis	4
1.3. Contributions.....	5
1.4. Thesis Outline	5
Chapter 2 : Background	6
2.1. Social Networks	6
2.2. Implicit versus Explicit Social Network Formations	7
2.3. Social Network Analysis versus Dynamic Network Analysis.....	8
2.4. Related Research.....	9
2.5. Social Bookmarking and Digg Network.....	11
Chapter 3: User Characterization	15
3.1. Network Description and Statistics.....	15
3.2. Degree Distribution.....	16
3.3. Egonet Analysis	17
3.4. User Membership Analysis.....	23
Chapter 4: Comparative Analysis of Network Formations	25
4.1. Introduction.....	25
4.2. Network Properties	28
4.3. Experimental Results	32
4.4. Discussion	34
Chapter 5: Popularity Prediction of Online Content	37
5.1. Introduction.....	37
5.2. Prediction Methods	37
5.3. Feature Description.....	38

5.4. Results and Discussion	41
Chapter 6: Opinion Mining.....	45
6.1. Introduction.....	45
6.2. Related Work	46
6.3. SentiWordNet	47
6.4. Method	48
6.5. Feature Description.....	50
6.6. Classification Results.....	53
6.7. Discussion	54
Chapter 7: Timed Egonets – Tracing Periodicity	56
7.1. Introduction.....	56
7.2. Process	57
7.3. Results and Discussion	59
Chapter 8: Applications	65
8.1. The Digg Effect.....	65
8.2. Evolution of Opinions.....	69
Chapter 9: Conclusion and Future Directions.....	73
Appendices.....	76
A.1 Digg Dataset Crawler.....	76
A.2 XML for Digg Story	82
References.....	86

LIST OF TABLES

Table	Page
Table 1.1: Ranking of social bookmarking services by EBizMBA.....	3
Table 2.1: Digg Dataset Statistics.....	12
Table 4.3: Social network statistics for co-participation and reply-answer network.....	32
Table 5.1: Performance for Digg-Score Prediction (Ten hours data only).....	41
Table 5.2: Performance for Digg-Score Prediction (Fifteen hours data only).....	41
Table 5.3: Performance for Digg-Score Prediction (All data).....	42
Table 6.1: Prediction Results with Sentiment Features	52
Table 6.2: Prediction Results without Sentiment Features	52
Table 7.1: Classification of Digg stories into constant sized bins	58
Table 8.3: Product popularity comparison based on summarized sentiments.....	71

LIST OF FIGURES

Figure	Page
Figure 3.1: Distribution of Degree (Log Scale)	16
Figure 3.2: Egonets	18
Figure 3.3: Distribution of the ratio of within-category to the overall degree.....	21
Figure 3.4: Distribution of the ratio of within-topic to the overall degree	22
Figure 3.5: Hierarchical Entropy Distribution: User Entropy versus # of Users.....	24
Figure 4.1: A typical comment thread	26
Figure 4.2: Resultant reply network from the comment thread	26
Figure 6.3: Reply-network's edge between two users	54
Figure 7.2: Average Degree for the 16 most active users in two month periods.....	59
Figure 7.3: Timed Egonets.....	62
Figure 8.1: Unique visits per hour – Normal behavior without Digg Effect	66
Figure 8.2: Unique visits per hour – Slashdotted behavior with Digg Effect.....	67

ABSTRACT

COMMENT MINING, POPULARITY PREDICTION, AND SOCIAL NETWORK ANALYSIS

Salman Jamali, MS

George Mason University, 2010

Thesis Director: Huzefa Rangwala

With the growing number of online collaborative news aggregator social websites, we witness thousands of comments posted by the internet community on individual news items shared on such networks. We started out with an objective to exhaustively analyze these comments for extracting insightful information about their various collective aspects. For our study, we worked with the data of one of the most popular news aggregator websites, called Digg¹. Using Egonet analysis for projecting local neighborhoods, we identified the characteristics of highly active individual users with and without time constraints. The time-based egonets effectively improved our ability to visualize variations in user activity patterns. We proposed a framework to apply data mining techniques to these comments (and comment threads), which helped us in predicting the popularity of news stories. We reported a very small loss of 1.0-4.0% in

¹ <http://www.digg.com>

multiclass classification accuracy while predicting the popularity score using the first few hours of comment data in comparison to all the available comment data. We found that Digg community was highly active in posting comments and found their focus to be spread across a wide range of topics. We also performed a comparative analysis of two network formations: *co-participation* and *reply-answer*. This helped us in comparing these implicit networks that we derived with characteristic attributes of social networks. Further, we conducted preliminary experiments to improve the strength of a link in our co-participation network by analyzing the positive, negative or neutral sentiments expressed by users in their commentaries.

One important application of our work lies in a provision of unique and rich information to advertisers enabling them to target certain commenters as potential customers. Our framework can also be tweaked to forewarn web administrators against a potential *Digg Effect* (Section 8.1).

CHAPTER 1: INTRODUCTION

The past decade has seen a massive rise in web services and applications that allow users to create, collaborate, and share varied forms of data like articles (web-blogs), pictures (Flickr²), video (Youtube³), and status updates (Twitter⁴). Social bookmarking websites like Delicious⁵, Slashdot⁶, and Digg allow users to submit links to web content they find interesting along with a short description. Every user in these online communities can initiate or contribute to an ongoing discussion by providing comments for the posted content, and also rate the articles that they find interesting. Thus, social bookmarking sites serve as data aggregators, web-based discussion forums, and an online collaborative filtering system that can collectively determine popular online content.

Recently, there have been several studies [1], [2], [3] that have analyzed social networks generated from comment interaction between users. In this work we model a co-participation network similar to the co-authorship and citation networks [4], [5] where users are linked together if they comment on the same discussion thread or submitted story. This implicit relationship between users based on comment information provides an understanding of the complex underlying community structure. We use egonets [6] to

² <http://www.flickr.com>

³ <http://www.youtube.com>

⁴ <http://www.twitter.com>

⁵ <http://del.icio.us>

⁶ <http://www.slashdot.org>

capture the local neighborhoods of users within the derived social network, and provide an understanding of the community with multiple interests. We further extract several user-based and comment-based features, and train classification and regression models for predicting popular stories. We evaluate our methods to use features derived from comments that were posted within the first few hours of posting the story. Successful prediction of popular content allows users to sift through the vast amount of available online data and can also aid in the ranking algorithms pursued by social bookmarking websites.

For our analysis, we use Digg, a popular social bookmarking website that allows users to share comment, and rate on diverse online available information. Digg was founded in 2004 by Kevin Rose⁷. We found that the user community within Digg was highly active in posting comments and found their focus to be spread across a wide range of topics ranging from world business to entertainment. We also showed the ability to predict the popularity index using early available comment and user based features.

We continued our analysis and explored two related areas: Dynamic Network Analysis and Semantic Analysis. The experiments, results and discussions pertaining to these studies are present in chapter 5 and 6.

1.1. Motivation

Enormous amount of opinions and beliefs are being shared in online communities on a daily basis. This content provides useful information to marketers, policy makers, intelligence agencies, and election candidates. When people claim responsibility against a

⁷ <http://digg.com/about/kevin>

comment that they make on some internet community, they take great care in picking the best words for expressing themselves. They end up sharing ideas, which are valuable for people who conduct analysis in the respective domains. The very same people are responsible on behalf of their companies to make sure that the image of their object of interest is never spoiled. Further, it's essential that the internet community takes interest in their products, slogans, and/or objectives.

To assess the popularity that follows a certain news item about anything, fortunately, there is a new platform that can be utilized (ironically, manipulated⁸), called Social Bookmarking. Table 1.1 is a list of top 10 Social bookmarking services as per the criteria defined in [7]:

Table 1.1 - Ranking of social bookmarking services by EBizMBA

Rank	Service	Inbound Links	Monthly Visitors	Alexa Rank
1	Twitter	760,750,806	23,579,044	13
2	Digg	383,598,000	33,433,760	183
3	Yahoo! Buzz	20,031,000	8,119,906	NA
4	tweetmeme	422,863	18,244,542	1,898
5	StumbleUpon	234,000,000	4,418,609	362
6	Reddit	161,685,000	4,908,990	441
7	technorati	175,287,000	3,309,174	662
8	del.icio.us	427,665,000	1,623,083	2,476
9	kaboodle	2,600,000	3,941,212	1,694
10	Mix	16,005,000	879,108	645

⁸ <http://www.wired.com/techbiz/people/news/2007/03/72832>

Millions of comments are posted against the news stories, which are linked on these services. Just taking a glimpse at all the comments that are posted for some topics of interest in some particular time-slot is beyond the scope of manual analysis. Now considering the importance of this wealth of commentary that we discussed earlier, it was surprising to notice inconsiderate research focus on this topic. The most relevant research pertaining to our work was related to the analysis of Yahoo Questions & Answers [1] dataset and Slashdot news forum [2].

Essentially, we were interested in techniques and methods that could summarize few opinions in a way that would be helpful in forecasting the future and planning ahead of time. The forecast and planning could be for any object, be it a product, person, organization. We built a prediction model using standard classification and regression algorithms for predicting the popularity of links that were shared on Digg. We show promising results for predicting the popularity scores even after limiting our feature extraction to the first few hours of comment activity that follows a Digg submission. The results are evident of the usability of our framework.

1.2. Problem Hypothesis

There is a lot of opportunity in mining user comments, and harnessing the hidden structures in comment threads. An exhaustive analysis of such co-participation networks and/or reply-answer networks in static and dynamic setting would significantly help in studying diffusion of information, identifying hidden communities, tracing the evolution of communities, and in extracting insightful information in terms of user sentiments.

1.3. Contributions

- Popularity prediction of online content with limited visibility of events, and
- Insightful user characterization based on what, when, how, and where users comment against some online content.
- Publication: “*Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis*” [79], which was published by IEEE for WISM’09-AICI’09.

1.4. Thesis Outline

This thesis is divided into following 10 chapters: Chapter 2 provides the background information on the essential concepts used throughout the thesis. In Chapter 3, we provide discussions on the various experiments conducted for characterizing users. In chapter 4, we provide a description of our prediction model, the experiments we conducted, and a discussion of the results. Chapter 5 presents the preliminary experimentation we performed to quantify semantics of user comments. In Chapter 6, a comparative analysis of two network formations: *co-participation* and *reply-answer*, is discussed. Chapter 7 explains that evolution of user behavior captured using *Timed Egonets*. In chapter 8, we present two important future applications of our study. Finally, chapter 10 concludes the thesis and suggests some directions for future research.

CHAPTER 2: BACKGROUND

2.1. Social Networks

A network is a set of items called vertices or nodes, with connections between them called edges [8]. Such a network is the simplest form and more complex networks can be formulated with different types of nodes, and edges. Numerous systems around the world can be represented as networks, for e.g. network of friend-to-friend connections [9], or business-to-business connections [10, 11, 12], protein networks [13], the World Wide Web, network of paper citations [81], and email networks [80]. With the advancement in performance and storage capacity of computers, nowadays, we can analyze networks containing millions of nodes and edges. Consequently, the research methods pertaining to network analysis are now directed towards exploring and answering the same questions using commonly available network analysis tools.

In [14], networks from different branches of science were compared and the commonalities were highlighted. One of the types of network explored was Social Network. A social network is defined between persons or groups of persons with some pattern of interactions or connections amongst them [15, 16].

Traditionally, analysis of these social networks has suffered from limited capabilities of resources and small sample sizes. Surveys and interviews were conducted

in hope of collecting data but all such approaches were labor intensive. [17] provides a review of the issues that were faced in early social network analysis. When these problems were overcome, it resulted in network analysis of well-documented and reliable networks, like actor-to-actor network using IMDB database [82], friend-to-friend network [9], reply-network [2], and Question Answer network [1].

2.2. Implicit versus Explicit Social Network Formations

Each collection of objects that is used to formulate a social network carries certain features and limitations with it. Networks like Facebook⁹, Orkut¹⁰, and LinkedIn¹¹, have one limitation in common – every user must be an *explicitly defined social contact* of any other network user in order to interact using the relative service’s facilities and for sharing interests with him or her. Contrary to this, paper citation networks, co-participation networks [1, 2], and other such networks do not restrict interactivity and hence, people assume hidden, also called implicit links amongst themselves by following each other in respective ways. Such linkages can also be called *implicitly defined social contacts* [2]. So, although in a network of implicit edges, people might not be close or real-world friends, but, they share something in common, and most probably it’s the opinions on topics of shared interests [18, 19]. In this research, we have extracted implicit networks from the dataset of a social bookmarking service, known as Digg.

⁹ <http://www.facebook.com>

¹⁰ <http://www.orkut.com>

¹¹ <http://www.linkedin.com>

2.3. Social Network Analysis versus Dynamic Network Analysis

In Social Network Analysis, ideas derived from graph theory are applied to a social network for studying the underlying relationships. In contemporary research, the term SNA refers to analysis of any network where nodes are of one or two types only [20]. Therefore, the SNA tools are effective across many domains of networks regardless of the nature of nodes. Typical analysis of social networks is conducted in a static setting, which means that irrespective of time, all of the available edges and nodes are consumed in forming the social network. However, extraction of communities, clustering of nodes, tracing diffusion of information, and statistical analysis of a network can be done in a dynamic setting as well [21, 22, 23].

Social user groups or communities evolve over time and the evolution can be traced if we split the social network across a sequence of time slots. Such a study can result in insightful suggestions about the behavior of individual nodes, the influence of certain nodes on other nodes, and the diffusion of information. The application of this information is highly valued in epidemiology (e.g. tracing pandemic viruses) and viral marketing (e.g. propagation of an idea of innovation).

NAACSOS¹² define Dynamic Network Analysis as follows: Dynamic Network Analysis (DNA) varies from traditional social network analysis in that it can handle large dynamic multi-mode, multi-link networks with varying levels of uncertainty. DNA, like quantum mechanics, is a theory in which relations are probabilistic, acts of measurement change the network, and movement in one part of a network propagates through the

¹² <http://www.casos.cs.cmu.edu/naacsos/sections/dna.php>

entire system. Network properties change over time and the actors can drift from one behavior, rank or opinion to next [20].

There is a lot of research interest in DNA and few interesting frameworks were developed for such analysis [21, 22, 23, and 24]. In this thesis, we have tried to visualize and describe behavioral changes adapted by Digg users over time. We introduced a concept of *Timed Egonets* where for each of the active commenters, we generated different mutually exclusive egonets in a sequence defined by 8 time bins. Chapter 7 provides a complete description of the method, experiments and results pertaining to timed egonets.

2.4. Related Research

USENET was one of the first web based message forum developed in 1979 and has seen several works related to development of tools for visualizing the structure of the discussions within these forums [75]. Statistical analysis methods [76] and network analysis [77] methods were developed to understand the characteristics of the different discussion forums.

Recently, researchers have used comment information to define implicit relationships between users, and then used social network analysis methods to understand the characteristics and interaction patterns of several communities and groups [3], [78]. Implicit relationships or links are defined between users who comment or reply on discussion threads to a particular user [3]. Within the context of individual web-blogs, a relationship was defined between the author of the blog and the commenter [78].

Our work is closely related to the analysis of the community participating in the Yahoo Question and Answer forum (Yahoo QA) [1]. In case of the Yahoo QA forum a user posts a question and several users provide an answer which are rated by the community. The work analyzed the interaction patterns between the various users belonging to multiple categories. An interaction or relationship was defined as a directed edge between the user who initiated a question and the users who replied with an answer. Using egonets [6] to characterize the local neighborhood of users within the derived social network, differences in the interaction patterns between users belonging to the technical and advice forums was observed. In our work, we define a weaker undirected interaction between two users who comment on the same story.

Recently, a social network was modeled [2] for the user community in Slashdot (another online bookmarking site). The implicit relationship was defined similar to the reply-answer network above, where an edge was defined between users who would comment directly to posted comments. Thus, if user A posts a comment, and user B replies to the comment, a relationship exists between users A and B. However, if a user C comments to the story but not to A's comment then there exists no relation between user A and C. Our definition of the implicit relationship between user follows the more traditional definition in co-authorship network [4], [5] and will results in relationships between the three users A, B, and C in the above example.

2.5. Social Bookmarking and Digg Network

The Federal Depository Library Program (FDLP¹³) defines Social Bookmarking as follows: “A Web-based service where users can create and store links. It is an increasingly popular way to locate, classify, rank, and share internet resources. While Web browsers have the ability to bookmark pages, those links are tied to that browser/computer. Social Bookmarking, on the other hand, is tied to an online account, which can be publicly or privately accessible. Based on the viewing properties and tags/categorization, these bookmarks can be shared and discovered by others. It is a way to share news, sites, and much more with a broader audience.”

Digg is one of the most active social bookmarking website where registered users submit news articles, videos, and images along with an optional short description. Submissions can lead to a discussion amongst the registered users who may post a series of comments regarding the material posted. A registered Digg user can rate the submissions (referred to as stories in this work), and support the stories that they find interesting by providing a positive rating referred to as a *digg*. On the other hand users can also provide negative rating known as a *bury*. Using the collaborative effort of millions of registered users, stories get rated to have a Digg-score, which is defined as

$$\text{Digg Score} = \text{Sum of Diggs} - \text{Sum of Buries}$$

Digg-score serves as a popularity index. The exact algorithm that decides which stories would show up on the front page of Digg website is not revealed, but stories that

¹³ <http://www.fdlp.gov/home/tutorials/fdlpdesktop/246-what-is-social-bookmarking>

achieve a high Digg-score from a diverse group of users are, almost always, promoted to the popular section of Digg [25].

Table 2.1 - Digg Dataset Statistics

Category	S	U	M	C/S
World & Business	7341	133468	84220	252
Technology	7536	117441	48567	135
Offbeat	4715	118446	51111	205
Entertainment	3850	90414	19634	150
Science	4924	82575	14765	113
Lifestyle	4221	93161	16465	143
Gaming	2399	69110	13331	177
Sports	2199	51257	5753	90

In Table 2.1, S denotes the total number of stories within the categories. U indicates the total number of users who commented at least once for the stories within the categories. M indicates the total number of users assigned to the categories (members). C/S denotes the average number of comments per story within the category.

Users also have the option to provide a rating for the individual comments. A positive rating for a comment is an *up score* whereas a negative rating is a *down score*. We used the Digg API to crawl 37,185 popular stories from November 16, 2007 to March 10, 2009. The total number of comments in our dataset is 6,188,266, and the total number of users who posted at least one comment is 253,846. The Digg-score for the

crawled stories ranged from 86 to 37,947 with a mean of 1,204 and a standard deviation of 1,122. The average number of comment made by a user is 24.

As shown in Table 2.1, stories at Digg are classified hierarchically into two levels, namely eight categories and 51 topics (not shown in Table 2.1) within the different categories. The eight categories include

1. *World Business,*
2. *Technology,*
3. *Science,*
4. *Gaming,*
5. *Sports,*
6. *Entertainment,*
7. *Life Style, and*
8. *Offbeat.*

There were a total of 51 topics when we crawled the data. Examples of topics include “Apple”, “Microsoft”, and “Linux” within “Technology”, “Football” and “Basketball” within “Sports”, and “2008 US Elections” (one of the most popular topic) within “World Business”. At the time of this writing however the topic “2008 US Elections” was no longer present. Table 2.1 provides general statistics about the dataset divided across the eight categories. The table shows the number of stories (S), total number of users who at least commented (U) once, and the average number of comments per story within the eight categories.

We also assign a user membership to one of the eight categories. This is done by assigning the user the category where he/she comments the most. In Table 2.1, we report the total members per category (M). We similarly assign a user to belong to one of the topics within the categories. From columns U and M we notice that there is a large overlap in the categories that users comment.

CHAPTER 3: USER CHARACTERIZATION

Motivated by the work involved with co-authorship and citation networks [26], [27] we define a co-participation network to model the relationships between different users in the Digg community.

3.1. Network Description and Statistics

An undirected graph $G = (V, E)$ is used to represent the co-participation network. The set of vertices V represent the set of users commenting across the different stories. The sets of edges E represent the interaction between the different users, and an edge E_{ij} exists between users V_i and V_j if the pair of users co-participate by commenting on n or more stories. We experimented with the threshold parameter n used to define the presence or absence of an edge or relationships between users. The average degree (i.e., number of edges per node) was 2414.5, 114.4, and 26.8 for threshold values of n equal to 1, 4, and 8, respectively. For the results reported here we use a threshold value of $n = 4$ i.e., a pair of users are considered to be connected if they both comment on at least four different stories.

A pair of users commenting on the same story may have differing or even opposing views. In the future (refer to chapter 5), we aim to refine our relationship definition between the users based on the polarity of the comments i.e., perform

sentiment analysis or opinion mining [28], [29] using text information of the comment.

3.2. Degree Distribution

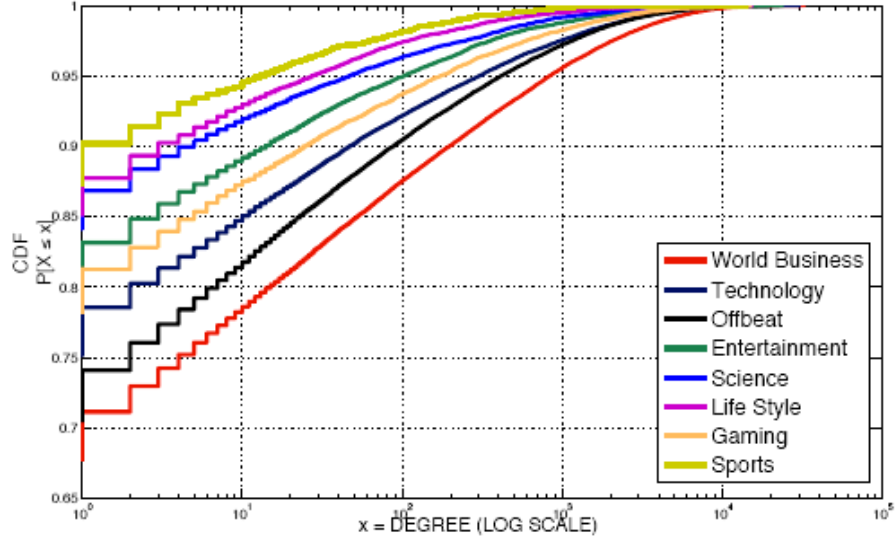


Figure 3.1 - Distribution of degree (Log Scale)

In Figure 3.1 we present the cumulative distribution (CDF) of the degrees per user separated based on membership to one of the eight categories. Nodes with more degrees indicate the user participating with several other users.

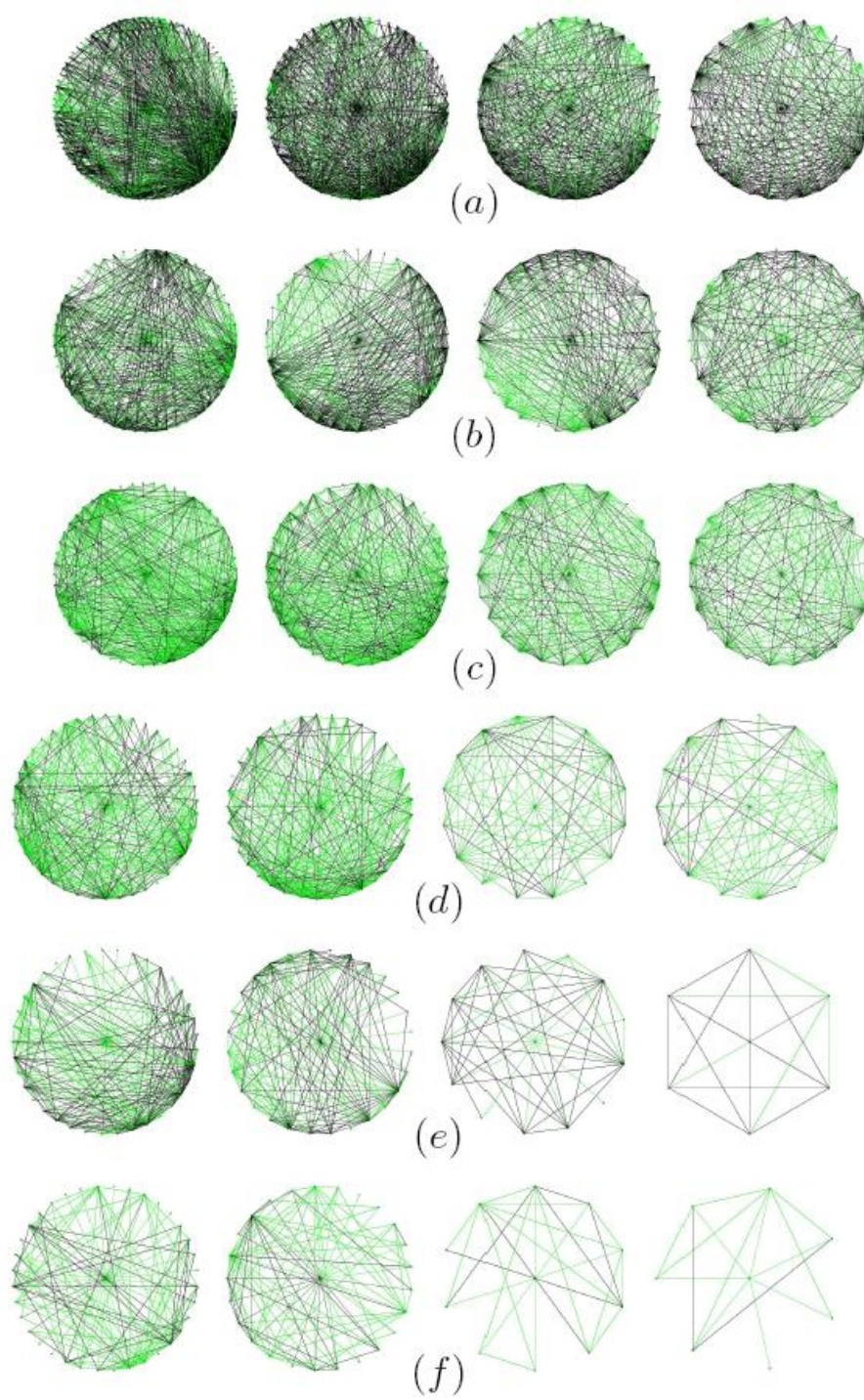
From this graph, we observe the difference between the users in the eight categories. The degree is plotted using a logarithmic scale and indicates a heavy tailed distribution, referring to the high levels of co-participation activity for the various categories. The category *World & Business* shows the highest participation amongst the

users as seen from the CDF plot. This is primarily because of the high level of activity in terms of posting stories, and discussions due to the *2008 US Presidential Elections* (a topic under World & Business), a popular topic when we downloaded the data. We can also see the differences between the other categories.

A user was assigned a category membership based on the category in which he/she would post the maximum comments. A user was free to comment across various categories, and though we compute the degree per user and analyze by category, we do not restrict the neighbors to be in the same topic or category.

3.3. Egonet Analysis

One technique to visualize the local neighborhood of individual users is by using Egonets. We also use egonet analysis to understand the relationships amongst different users within the different categories. Such an egonet analysis was done previously [1, 30] to differentiate between community of users that were discussion prone or not. A one level egonet for a user is defined as the user, the set of users who interact directly with the user (neighbors), and the relationships between those users. We can extend the definition of egonet to have neighbors who are N hops (links) away from the user in consideration.



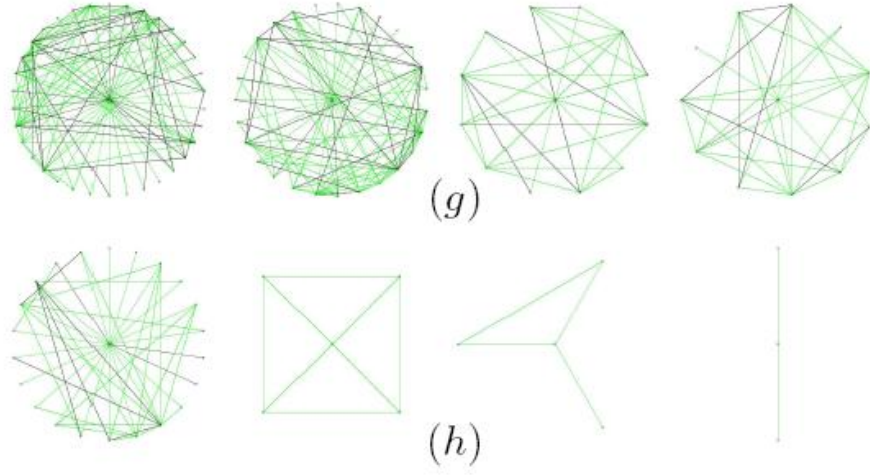


Figure 3.2 - Egonets

In Figure 3.2, each row from left to right shows egonets for 1st, 20th, 80th, and 100th most active users in Categories: (a) World & Business, (b) Technology, (c) Offbeat, (d) Entertainment, (e) Science, (f) Life Style, (g) Gaming, and (h) Sports. We show the egonets for a set of four active users within each of the eight categories. For each category we identify the most active users, i.e., users who have commented the most on stories posted within a particular category. Figure 3.2 shows only the 1st, 20th, 80th, and 100th most active users per category. We have up to 200 egonets per category at the project website <http://www.cs.gmu.edu/mlbio/digg-ego/>.

The egonets we present have a two color coding scheme where a co-participation edge $E_{i,j}$ is colored black if both V_i and V_j have the same category membership, whereas edge V_i and V_j is colored green if users V_i and V_j belong to different categories.

Analyzing these egonets helped us to differentiate between users based on their focus and participation level. If the egonet is very dense, it infers that

- the user in consideration has many neighbors, and/or
- these neighbors are highly connected amongst each other.

By analyzing the color density we can infer two things based on the two colors that we used. A blackish egonet means that many of the neighbors belong to the same category as the category of the user in consideration; hence he can be categorized as focused user. A greenish Egonet on the other hand implies that the user is not focused because the neighbors belong to different categories. So, we can say that a blackish dense egonet represents a highly active and focused user as compared to a blackish sparse egonet, which will be moderately active and focused user. Whereas a greenish dense egonet would represent a highly active but unfocused user; and, a greenish sparse egonet belongs to a moderately active and unfocused user.

We have ordered the categories from top to bottom in decreasing order of the densities of egonets. The egonets of users in categories like “Sports”, “Gaming”, and “Life Style” (Figure 3.2 (f)-(h)) have smaller and less denser neighborhood in comparison to categories like “World Business”, “Technology”, and “Offbeat” (Figure 3.2 (a)-(c)). The dense nature of egonets for the “World Business” category can be explained by the large number of stories that became popular due to the 2008 US Elections. From this data we can also infer that within the Digg community stories within the Sports and Gaming categories do not lead to large user interaction and discussion.

The egonets also suggests that users within “Technology” participate by way of commenting in much larger volumes in comparison to “Science”.

The egonets for the “World Business” category (Figure 3.2(a)) show more black edges in comparison to the corresponding egonets for the “Offbeat” category (Figure 3.2(c)). This suggests that the “World Business” community users are focused and involved with discussing stories that are posted in that category. The “Offbeat” category is a collection of diverse topics (e.g., “comedy”, “pets”) that do not fit within the other seven categories. As such it is expected that users within the “Offbeat” category are loosely coupled i.e., not focused to comment in the same categories as their category membership indicates. We observe that users in the “Offbeat” category also comment on stories posted in the “Life Style” and “Entertainment” categories.

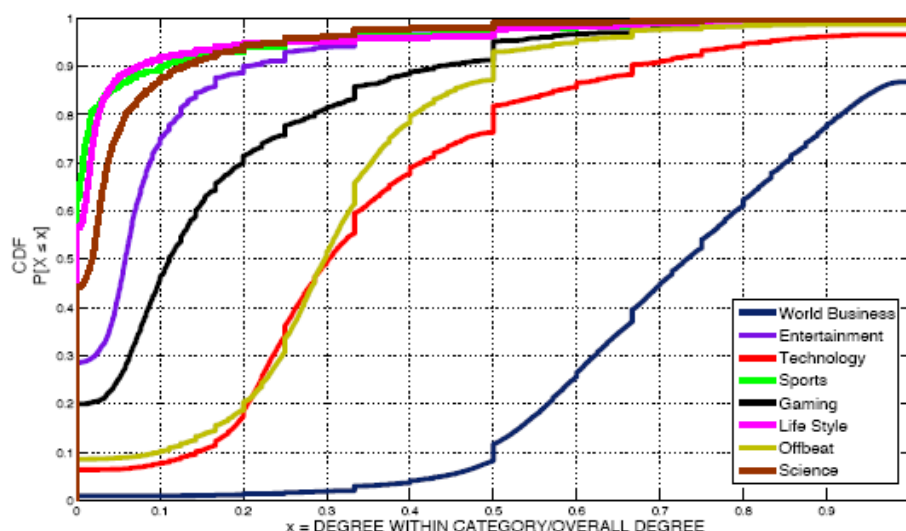


Figure 3.3 - Distribution of the ratio of within-category degree to the overall degree

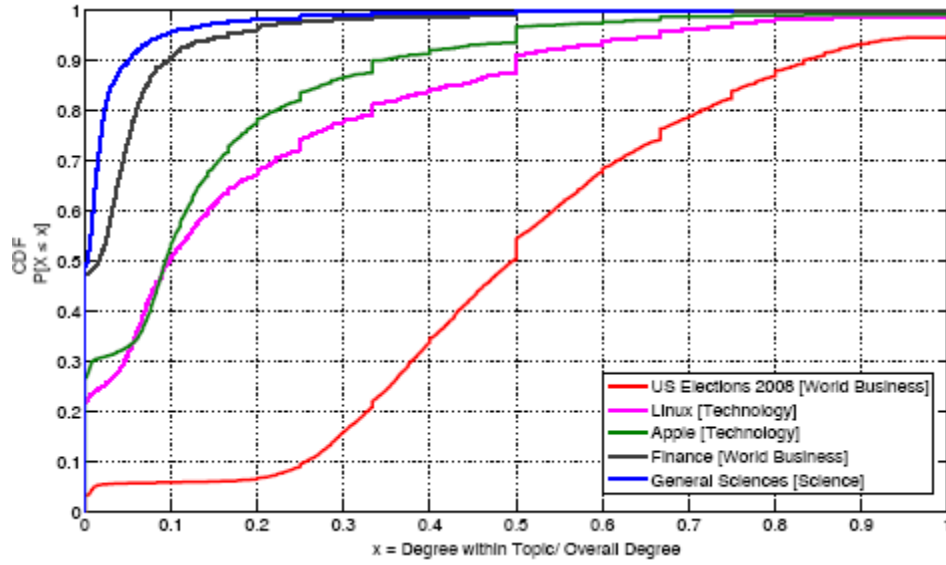


Figure 3.4 - Distribution of the ratio of within-topic to the overall degree

In Figure 3.3 we show the cumulative distribution function for that ratio of in-category degree (or within-category) to the overall degree. The in-category degree for a node is the number of one-hop neighbors who have the same membership as the user in consideration. The Figure 3.3 provides complementary results to the ones observed for the hundred most active users seen in the egonet analysis by allowing us to see the percentage of users below a specific value of the in-category ratio.

In Figure 3.4 we show the cumulative distribution for the ratio of the in-topic degree to the overall degree corresponding to five selected topics within the eight categories. It is interesting to see that the users who comment within “2008 US Elections” topic are highly topic-focused in comparison to the “Finance” topics, both within the “World Business” category. The “Technology” topics “Apple” and “Linux”

also have a high degree of in-topic focus. In both the Figures 3.3 and 3.4 we neglect users having an overall degree of zero. This does not have an effect on the trends observed and allows us to focus on the users with at least a single neighbor.

3.4. User Membership Analysis

As discussed in the previous section, users within the Digg community have overlapping interests and as such participate and comment across multiple areas of interest.

As done previously [1], we computed an entropy measure to capture the focus of the user. Users commenting within a large number of categories in comparison to a user commenting across a fewer number of categories would have higher entropy and less focus. As such, we can define the entropy for the user with respect to the categories as

$$H_1 = - \sum_i p_i \log(p_i)$$

where i iterates over the eight categories and p_i denotes the probability for the user to belong to category i . Similarly, we can compute an entropy measure for the user with respect to the 51 topics given by

$$H_2 = - \sum_j p_j \log(p_j)$$

The sum of the H_1 and H_2 represents the total hierarchical entropy for a user. Using such a two level hierarchical entropy definition allows us to differentiate between users who would comment on a diverse set of subcategories within a single category (less entropy because H_1 will be low) and users who would comment on a diverse set of subcategories spread across multiple categories (higher entropy because H_1 will be high).

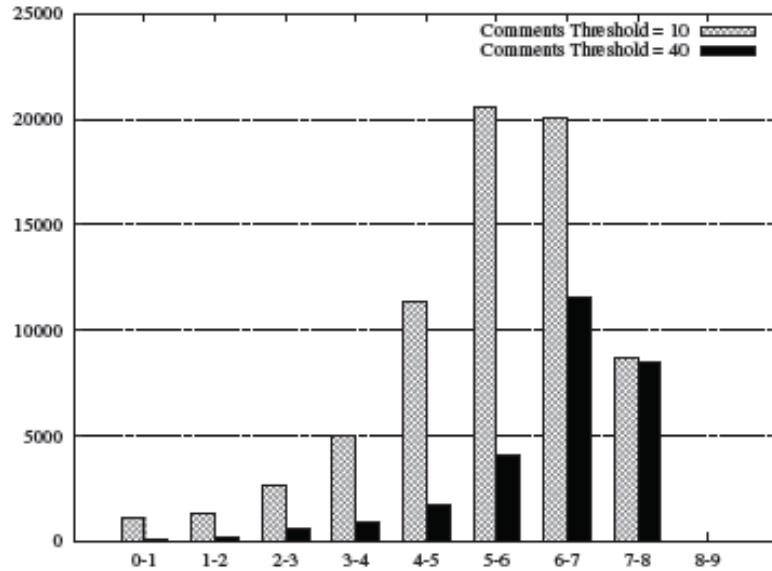


Figure 3.5 - Hierarchical Entropy Distribution: User Entropy (X) versus Number of Users (Y)

In Figure 3.5 we show the hierarchical entropy distribution for the 70,753 users who commented at least ten times and 27,645 users who commented at least forty times. Computing the entropies for users who comment very few times would bias the analysis (entropy for a user who comments once will be zero). A high percentage of users have a higher entropy i.e., in between 5.0-7.0 and 6.0-7.0 for users that commented at least 10 times and 40 times, respectively. It can be inferred that users have a tendency to participate and comment across multiple discussion topics. This suggests that the user community in Digg is not very focused but this could be due to loosely defined categories and subcategories (called topics). We also use the entropy measures of a comment author as features for predicting the popularity of a story.

CHAPTER 4: COMPARATIVE ANALYSIS OF NETWORK FORMATIONS

4.1. Introduction

In all the experiments that we conducted by far, the underlying social network formation was based on the principle of co-participation. As we discussed earlier, in our Digg co-participation network, two users are connected if they both comment on X number of similar stories. The density of such a network is indirectly proportional to this number X :

$$X \text{ (Co – participation link threshold)} \propto \frac{1}{\text{network density}}$$

Such a formation assumes that it's not a coincidence for any two random users to co-comment on a relevant number of stories, that is, such a behavior is indicator of existence of a common interest.

Essentially, researchers try to capture links that are truly representative of common interests and opinions. [2] suggests that there are other ways to model a social network out of comment threads. One important formation relies on replies received by comments from other commenters. If we model a link against each reply, the resulting network can be called *Reply-Network*.

In Figure 4.1 and 4.2, we show a miniature example to illustrate the generation of a reply-answer network from a sample comment thread.

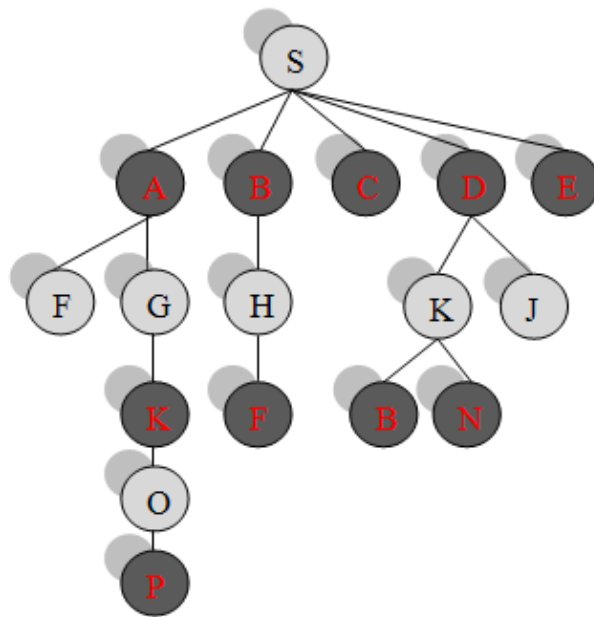


Figure 4.1 - A typical comment thread.

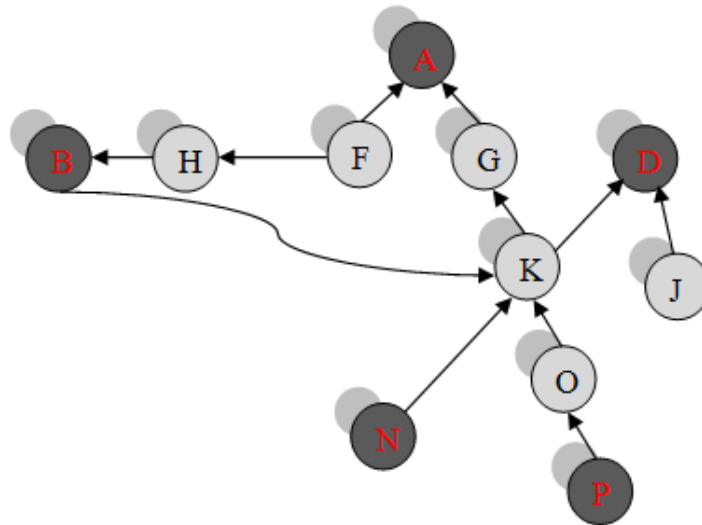


Figure 4.2 - Resultant Reply Network from the above comment thread

Figure 4.1 is an arbitrary comment thread representative of typical scenarios of commentaries on social bookmarks. The root node is the story itself, and each of the other nodes represents a comment. Each comment is tagged by an alphabet that's the username of the commenter. Figure 4.2 is the corresponding general reply network for this thread. Note here that this reply network is unconditional - one simple reply by X to Y results in one link whose weight is ignored here for the sake of simplicity. There can be a number of variations of a reply network based on how we determine the weight of the links. The more restrictive it is, the less dense the network becomes.

The corresponding co-participation network, being unconditional with co-participation threshold equal to 1, isn't shown here. It will be no different than a complete graph where there is a link b/w every pair of commenters because they all commented on story S. So, the very first obvious difference between a reply network and a co-participation network emerges out to be the density of the network.

We performed an exhaustive comparative analysis of the reply networks in [2], Digg's co-participation network, and a Digg reply network. Indicators of this analysis are essentially certain network properties that are critically important in any sort of social network analysis.

The first step to exploring a social network is to identify the information that's of interest to the researcher. As we discussed earlier, different formations tend to reveal varying insights from the same social network data. Comment threads, if merged together, result in a collaboration network where we can claim that the purpose of collaboration is to democratically decide the popularity of the news item.

In this section, we have done a rough comparative analysis to see if there are other ways in which our Digg network could have been modeled. Real networks tend to differ from Random Networks. This was studied in depth by Rapport [55, 56], and it indicates that we can manipulate some network properties in to improve our real networks. Further, this exploitation guides the formation of networks. Contemporary research on networks focus on a limited set of network properties, which are universally computable and applicable; for e.g., “*small world effect*” [14, 57], clustering or network transitivity [14], community structure [59, 60, 61], degree distribution [62, 63], and spectral properties [64, 65, 66].

4.2. Network Properties

For many networks, following are few important properties that are important for their characterization.

Number of Nodes

For the same network data, this number,

$$N = |\text{number of nodes}|$$

varies based on network formation and the filters that are applied to the data. Whether the network constitutes of directed or undirected edges, relies on edge weight threshold or not, considers anonymous nodes (users) or not; all such aspects are primary factors that affect this number.

Number of Edges

Denoted by

$$M = |\text{number of edges}|$$

this number is controlled and contributes to density of the network. Typically, an undirected network has a larger number of edges as compared to a directed network; whereas, a network with no edge weight threshold will be denser as compared to a network where all edges below a certain edge weight are ignored.

Maximum Cluster Size

The compactness of the community inherent in the network can be derived by this number. The more compact a network is, the more typical it is as a social network. In such networks, the rate of flow of information is directly proportional to the maximum cluster size. It's an end goal of any social networking service providers to ensure that people have more means to create more connections resulting in a more connected (or compact) big network. It has been observed that undirected network tend to result in a bigger "giant component" as compared to directed network [2]

Average Degree

The degree of a node is the number of edges incident on it. The average degree of real-world networks is mostly found to be very unlike the random graph in their degree distribution. They are highly right skewed [8] meaning that the distributions have a long right tail of values that are far above the mean, and hence, they follow the power law. For more information on the degree distribution, please refer to the section 3.2.

Assortativity / Correlation Coefficient

[67] describe assortative mixing or assortativity as "*the correlations between properties of adjacent network nodes*". People tend to associate with others based on the similarities between them - friends are related by their assortative mixing behavior.

Consequently, in a social network, we expect to see this sort of assortative matching. The opposite of this behavior is called disassortative mixing. Assortative mixing is indeed present in many networks, it can be measured and it does affect a network's structure and behavior [67].

[67] defines an assortativity coefficient that's is similar to the correlation coefficient defined by Pearson. In any given network, if r is the assortativity coefficient, then it can have a value from +1 to -1. Any value closer to +1 show that there is a correlation amongst nodes of similar degree. For a random network, the value of r will be around 0. Lastly, values closer to -1 show complete disassortativity because such a value is representative of relationships amongst nodes of different degree.

People conducting research on networks prefer to measure assortative mixing by a scalar vertex property namely the vertex degree. We have assumed the same notion of correlation. Two observations can be made in terms of degree correlation:

- a) *high/low-degree nodes preferring other high/low-degree nodes, or*
- b) *high/low-degree nodes preferring other low/high-degree nodes.*

It has been observed that both situations are present in some networks [8]. To quantify degree correlation r , we used [67]:

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_q^2}$$

Weighted Clustering Coefficient

Previously, we showed some results pertaining to local egonets of a certain set of commenters. In real world networks, there are certain things that can be inferred based on

the high density of ties in a local neighborhood of nodes [14]. Such neighborhoods are examples of implicit formation of tightly knit clusters of nodes. Clustering Coefficient is a measure that assesses the degree to which nodes tend to cluster together.

The presence of social characteristics in a network can be witnessed if the probability of high-density ties is relatively higher than a random network of similar magnitude. Such a measure can be used to obtain the local cluster cohesiveness and it is defined for any node n as the fraction of its connected neighbors [14]. The average clustering coefficient, which is

$$CC = N^{-1} \sum_i C_i$$

thus expresses the statistical level of cohesiveness measuring the global density of interconnected vertices' triples in the network [14]. A problem with overestimation of clustering properties was also pointed out. A new metric, which combined the topological information with the weight distribution of the network, namely weighted clustering coefficient, was defined as,

$$C_i^w = \frac{1}{s_i (k_i - 1)} \sum_{j,h} \frac{(w_{ij} + w_{ih})}{2} a_{ij} a_{ih} a_{jh}$$

By using the weighted local clustering coefficient measure, we are not just considering the number of closed triangles in the neighborhood of a vertex but also their total relative weight with respect to the vertex' strength [14].

Average Path Length

A smaller value of average path length indicates that the network conforms to the idea of *small-world effect*. It is also concluded that most pairs of vertices in most networks seem to be connected by a short path through the network [2].

Maximum Distance Between Two Users

This value tells us about how many edges it would take to travel from one node to any other node within a particular cluster. A smaller value of this measure is an indicative of similarity with a traditional social network and presence of Small World Effect.

4.3. Experimental Results

Table 4.3 - Social network statistics for co-participation and reply-answer network

Indicator	Slashdot A	Slashdot B	Digg Net A	Digg Net B	Digg Net C
N	80962	80962	188494	253846	253846
M	1052395	905003	3084333	14519792	3397267
N/M	0.077	0.089	0.055	0.0046	0.0084
MC	73%	98%	76%	99.9%	99.8%
AC	-0.016	-0.039	0.000103	-0.45	-0.39
D	13(50.1, 49.4)	22.36(79.3)	16.363	114.398	26.766
WCC	0.026 (0.074)	0.047(0.12)	0.157 (0.259)	n/a	0.78 (0.357)
L	3.62 (0.7)	3.48(0.7)	3.79(0.72)	2.4(0.56)	2.29(0.48)
MD	10	9	10	10	6

In Table 4.3,

- *N* is the Number of nodes
- *M* is the Number of edges

- N/M is the Ratio of nodes to edges
- MC is the Maximum Cluster Size
- AC is the Assortativity Coefficient or Correlation Coefficient
- D is the Average Degree
- WCC is the Weighted Clustering Coefficient
- L is the Average Path Length
- MD is the Maximum Distance between two users

In this analysis, we have compared relevant network properties of five different network formations, which are derived from datasets collected from two different social bookmarking services namely, Digg and Slashdot. What follows is the description of the networks for which we have characterized the structural properties:

Slashdot A – Directed Reply Network of Slashdot

A directed edge exists from V_i to V_j if V_i replies to V_j at least once. For more on this network formation, please refer to [2].

Slashdot B – Undirected Reply network of Slashdot

An undirected edge exists between V_i and V_j if either of the users replies to the other at least once. For more on this network formation, please refer to [2].

Digg Net A – Directed Reply Network of Digg

This formation takes the co-participation to a next level, which is essentially more restrictive resulting in fewer edges. Recall that a network where edges are a result of one-on-one interaction patterns within the context of a comment thread is a reply network [2].

In this network, a directed edge exists between V_i and V_j if user I replied to V_j in any of the comment threads of the stories in dataset.

Digg Net B & C – Undirected Co-participation Digg Networks with threshold 4 & 8

This network is relatively gigantic because of the nature of its formation. The underlying principle is very simple – all commenters of a particular story are, to a certain extent, sharing interests. Then, if this behavior repeats itself, we capture it as something that strengthens the weight of individual edges b/w pairs of commenters.

To proceed, we picked up 4 and 8 as the threshold values for edge weights, generating two co-participation networks respectively. So, if two users co-comment on 4 different stories, there is an edge between them in the 4-threshold network, but not in the 8-threshold network. On the other hand, if two users co-comment at least eight times, there exist an edge b/w them in both of these networks.

The time complexity of the algorithm to compute this list of edges is directly proportional to the number of users in the network.

4.4. Discussion

The results of the Slashdot networks were taken from the published study by Gomez et. al [2]. Amongst the network formations belonging to one of the two services (Digg or Slashdot), it should be noted that even the smallest Digg network is at least 3 times the size of any Slashdot network in terms on number of edges. Basically, we report several network statistics to characterize the derived graphs [8] for the co-participation networks defined across the Digg dataset for edge threshold values of 4 and 8. We also

report these statistics for the reply network defined in the Slashdot data study [2]. We compute the directed reply network for our Digg dataset for comparative purposes.

The Table 4.3 also lists the total number of nodes N and edges M for the different network representations. As expected the co-participation networks has more edges or interactions between the different users in comparison to the directed reply networks. Looking at the ratios of nodes to edges, we realize that co-participation networks are relatively denser as compared to all other networks, in essence the reply networks. This is easy to justify because the easier the criteria for edge selection is, the denser the graph becomes. The ratios for the two co-participation networks are also supporting this idea as the ratio of Digg-B network (threshold 4) is almost half of the Digg-C network (threshold 8).

We also report the maximum cluster size (MC) or the giant component size [8]. In case of the co-participation networks defined for the Digg data we see that 99% of the users are within a single giant cluster and are connected to each other. In comparison, 73% and 76% of the users form the giant component for the reply-answer networks defined for the Slashdot and Digg dataset, respectively. From this conclusion, we can claim that Takahashi's (2000, 2004) postulate concerning higher density networks is applicable on our networks too. The postulate states that higher density may lead to a higher level of generalized exchange by enabling faster, more complex flow of information about past behaviors of beneficiaries, allowing for better sanctioning. In short, the density of the giant component of these social book-marking services is directly proportional to the rate of information flow.

A higher standard deviation in average degree (k) (i.e., number of edges per node) represents a higher level of heterogeneity within the community. For the reply networks, the average degrees are of the same magnitude with slightly high standard deviations. However, in case of co-participation networks, the standard deviations are significantly higher than those of the reply networks that show that the community captured by a co-participation formation is very heterogeneous. We have discussed the degree distributions earlier in section 3.1. We also report the average path length L and the maximal distance D in the largest cluster across the Digg and Slashdot networks. The co-participation network with edge threshold weight set to 8 has a smaller maximal distance.

We also computed the degree correlation r for all Digg networks and we noted that none of the networks shows any assortative mixing by degree. For a network whose nodes demonstrate assortative mixing, this value should be closer to +1. Perhaps, based on our values, which are closer to 0 and -1, the Digg networks can be characterized by disassortative mixing.

Interestingly, the values for Slashdot networks demonstrate a similar behavior. [2] states that Slashdot is characterized by neither assortative nor disassortative mixing. Such a behavior is comprehensible because people post comments on social bookmarking services without being concerned about the degree connectivity of the other commenters.

CHAPTER 5: POPULARITY PREDICTION OF ONLINE CONTENT

5.1. Introduction

Using the comment information associated with posted stories at Digg we predicted the popularity of a story, the Digg-score (See Section 2.5). We wanted to develop a predictive model that would be able to accurately infer the Digg-score using associated comments made by the user community but restricted to the first few hours of the initial posting of the story. As such we trained predictive models using comments only from the first ten hours and the first fifteen hours. We also trained models using all the available comment information for comparison purposes.

5.2. Prediction Methods

The prediction was performed by setting up three independent classification problems: (i) a 2-class, (ii) a 6-class, and (iii) a 14-class prediction problem. The bins for the 6-class and 14-class prediction problems were set in Digg-score intervals of 1000 and 500, respectively. For the 2-class prediction problem we split the instances into the first class having all stories with a Digg-score of less than 1000, and the second class with Digg-score greater than 1000. This allowed for a uniform size distribution split. We used the decision tree classifier [31], the nearest neighbor classifier [32], and support vector machines [33] for performing the classification. In this work we present the classification

results for the pruned C4.5 decision algorithm denoted by DT, the nearest neighbor classifier using nine neighbors denoted by 9-NN, and the support vector machine classifier using the linear and radial basis kernel function denoted by SVM (L) and SVM (R), respectively. For the K-class classification using SVMs, we trained K binary one-versus-rest classifiers for each of the K classes. We also estimated the Digg-score using - SVM regression method [6] (denoted by K=1).

5.3. Feature Description

We used several comment based and user focused features for predicting the Digg-score. These features capture different aspects of the data and are described in detail below:

Comment Statistics

We use the number of comments for a posted story, and the average word length of all the comments as features. A large number of posted comments are directly correlated to high level of user interest and hence, the popularity of stories. Both these features have shown success in predicting the best rated answer for a question within the context of Yahoo's QA Dataset [1].

When users comment they may chose to reply to a specific comment or make a new comment. This posting of comments induces a hierarchical tree structure where we can associate a level with every comment. A comment directly made to the story is considered as the first-level comment and can be thought of as initiating a thread. Analyzing our dataset we observed that a large number of levels are indicative of controversial stories. Controversial stories also have a tendency to be popular as seen in

the analysis done Slashdot [2]. Motivated by this we used four features that simply count the number of comments at the first, second, third, and fourth levels for each story.

Digg User Interest Peak

We captured the peak in user interest by determining the increase in user level activity within a fixed time span. We denote this feature as $burst(X)$, and is computed as the highest comment activity seen when sliding a window of “X” hours across all the posted comments for a story. Specifically, we can represent $burst(X)$ as

$$burst(X) = \max_{T=0 \dots T'} \frac{C(T \dots T + X)}{C(0 \dots T')}$$

where X is the burst window span, T' is the total time since the story was submitted to the community, and $C(x)$ is the number of comments within the x hours. We compute $burst(3)$, $burst(4)$, and $burst(5)$ as three features for predicting the Digg score. The use of different windows captures different types of user interest for a story. A higher burst weight with a shorter time span carries more information towards the popularity prediction.

Digg User Feedback

Digg users also have the option of rating the comments. As such, comments that are irrelevant with respect to the posted story or are spam get negative feedback. Comments that are relevant and even cause of controversy are seen to get positive feedback from the user community. For each comment we obtain their up score (positive feedback) and the down score (negative feedback). We derive two features that sum the “up” scores for all the comments and sum the “down” scores for all the comments

associated with the story. Generally, for popular posts it was seen that the sums of up scores were higher than the sum of down scores. A similar comment feedback measure was shown to be positively correlated with the popular stories posted on Slashdot [2].

User Community Structure and Membership

We also compute features that are focused on the egonets (described in Section 3.3) of the users with highly rated comments. We define top five comments per story as those comments having the highest comment score given by the difference of ups and downs. For each of these highly rated comment authors we use the degree or number of local neighbors (defined in Section 3.1) as a feature. This results in five features that use local information from the social network.

We also use the two entropies (H_1 and H_2) computed for the categories and topics (Section 3.4) as a measure of knowledge associated with commenter. We use the average entropies for all the comment authors as features, and believe that this captures the knowledge-base and involvement of a user which would be important for predicting the Digg-score.

Overall we use eighteen features to train our prediction models. For training models using the first ten hours, and the first fifteen hours after the story posting we recomputed the features. We standardize the feature values by centering on the mean.

5.4. Results and Discussion

Table 5.1 - Performance for Digg-Score Prediction (Ten Hours Data only)

	Ten Hours Data					
	K = 2			K = 6		
Method	ROC	F1	Q_2	ROC	F1	Q_6
DT	0.83	0.8	0.8	0.72	0.63	0.62
9-NN	0.81	0.75	0.75	0.76	0.59	0.63
SVM (L)	0.88	0.81	0.8	0.74	0.63	0.63
SVM (R)	0.84	0.79	0.78	0.79	0.66	0.64
	K = 14			K = α		
	ROC	F1	Q_14	CC		
DT	0.64	0.41	0.41	-		
9-NN	0.66	0.37	0.42	-		
SVM (L)	0.63	0.44	0.42	0.73		
SVM (R)	0.7	0.46	0.45	0.6		

Table 5.2 - Performance for Digg-Score Prediction (Fifteen Hours Data only)

	Fifteen Hours Data					
	K = 2			K = 6		
Method	ROC	F1	Q_2	ROC	F1	Q_6
DT	0.83	0.8	0.8	0.72	0.64	0.63
9-NN	0.81	0.75	0.76	0.76	0.59	0.64
SVM (L)	0.89	0.82	0.8	0.75	0.63	0.64
SVM (R)	0.84	0.79	0.78	0.8	0.66	0.64
	K = 14			K = α		
	ROC	F1	Q_14	CC		
DT	0.64	0.41	0.41	-		
9-NN	0.66	0.37	0.42	-		
SVM (L)	0.64	0.44	0.42	0.75		
SVM (R)	0.7	0.45	0.44	0.61		

Table 5.3 - Performance for Digg-Score Prediction (All Data)

	All Data					
	K = 2			K = 6		
Method	ROC	F1	Q_2	ROC	F1	Q_6
DT	0.87	0.82	0.82	0.76	0.66	0.67
9-NN	0.85	0.79	0.79	0.8	0.63	0.66
SVM (L)	0.91	0.84	0.83	0.79	0.65	0.67
SVM (R)	0.86	0.81	0.8	0.82	0.69	0.68
	K = 14			K = α		
Method	ROC	F1	Q_14	CC		
DT	0.65	0.43	0.44	-		
9-NN	0.69	0.38	0.43	-		
SVM (L)	0.67	0.46	0.45	0.8		
SVM (R)	0.74	0.48	0.45	0.64		

In Table 5.1, 5.2, and 5.3, we have listed the results of various regression and classification algorithms. DT, 9-NN, SVM (L), and SVM (R) denote the decision tree [31], 9 nearest neighbor classifier [32], SVM [33] with linear kernel, and SVM with radial basis kernel, respectively. ROC, F1, Q_K denote the average area under the ROC curve, F1 score, and K-way classification accuracy, respectively. CC denotes correlation coefficient. We highlight in bold the methods that perform the best classification or regression. The density estimation was performed using the μ -SVR method. The results shows the classification and estimation results for the different class definitions, and using the comment features extracted for the first ten hours, first fifteen hours, and the complete data. We report a small sampling of the experiments we performed. In particular, we used the default parameters (regularization, width) for the SVM based

methods, and report results only for the nine nearest neighbor classifier that showed the best prediction results.

To evaluate the performance of the classification and regression methods we performed 5-fold cross validation. The classification performance was evaluated using the K-way classification accuracy (Q/K), the area under the receiver operating characteristics curve [34] (ROC), and the F-score (F1). The ROC measures the area under the plot of true positive rate versus the false positive rate, whereas the F1 provides a weighted average between precision and recall. We report the correlation coefficient (CC) between the actual and predicted Digg-score for evaluating the regression results. We used the Weka Toolkit [35] and LibSVM [36] for the popularity prediction.

Firstly, we noticed that the two most discriminative features were the number of comments per story, and the sums of the ups (not shown here). These results are similar to the two similar works related to retrieving the popular posts in Slashdot [2], and predicting the best answer in the Yahoo QA dataset [1].

Analyzing these results, we observe that there is a slight improvement in the use of SVM based methods in comparison to the nearest neighbor and decision tree methods as the number of classes are increased. Solving the multi-class classification with higher number of classes is a challenging problem. The prediction performance of the classifiers and estimators when using the ten hours of data as well as the fifteen hours of data are comparable. We observe a 3.5%, 3.7%, and 1.32% decrease in the $Q/2$, $Q/6$, $Q/14$ accuracy when comparing the performance of prediction restricted to ten hours of data in comparison to the complete data, respectively. A similar trend is seen for the fifteen

hours of data. The low loss in accuracy suggests a merit in our predictive models for identifying the popularity of posted stories. Our results also show a strong CC for the predicted and original Digg-score using the -SVM regression method. The linear kernel is more effective in comparison to the radial basis kernel for the regression problem.

CHAPTER 6: OPINION MINING

6.1. Introduction

In this thesis we incorporated new features pertaining to the semantics of user comments to predict the popularity of posted Digg stories. We used SentiWordNet [37], a lexical resource, which can be used to quantify the semantics of the opinionated text using various scores that are defined by it for a dictionary. We wanted to show that these new features are essential factors for

(a) Evaluating the polarity of individual comments and

(b) For using them in identifying the clusters of users with similar opinions / sentiments.

There are two main types of textual information on the web as identified by [38]: *Facts and Opinions (or Sentiments).*

Marketers, researchers and numerous web analysis companies have shown significant interest in extracting the opinions and sentiments implied by the commenters in their blog posts, comments, and discussions. There is a momentous amount of opinions on the internet spread across hundreds of discussion forums, and social bookmarking websites.

Extracting a sentiment hidden in a user's product review is very important. This study of extraction opinions, called Opinion Mining or interchangeably Sentiment Analysis [39], *"is concerned not with the topic a text is about, but with the opinion it expresses"* [37]. Among other things, it helps independent online retailers to rank the products based on the exact sentiments expressed in the reviews. Similarly, summarizing the opinions that follow a newsbreak in the first few hours is also important. Summarization of these opinions also plays a vital role if we can notice a trend in the evolution of polarity of opinions over a time period. We can identify the bursts that triggered the shift in opinions, or we can identify the users who influenced such a drift.

Here, we are concentrating on identifying the most relevant feature set of Digg stories that contribute to its popularity. Previously, we achieved promising results by building classifiers over a number of such features. In this research extension, we have added semantic features that give us a summarized insight of the opinions embedded in the commentary that follows a news story.

6.2. Related Work

Analyzing the opinions expressed in comments requires us to know the common sense polarity of individual words used in the comment. Researchers have been using gold standards - manual tagging of commonly used words to come up with prior polarities [40, 41, and 42]. It has also been observed that it is *"harder to apply opinion bearing words collected from one domain to an application for another domain."* [42]

Opinions have also been categorized as either *Judgments* or *Predictive Opinions* [41]. Opinions where people express their likeness about something are called

Judgments, whereas, the expression of forecasting something based on one's beliefs is called Predictive Opinion.

A lexical resource built upon WORDNET [43], which associates three numerical scores called *Objectivity*, *Positivity*, and *Negativity* (derived by combining the results produced by a committee of eight ternary classifiers), is known as SentiWordNet (SWN) [37]. A number of research works are based on SentiWordNet including [44, 45, 46, 47, 48, 49, 50]. In this work, we have also used SentiWordNet to identify opinionated words in individual comments, following by a summarization of all comments in a particular story. This work is an extension of our previous endeavor to build classifiers to predict popularity of Digg Stories and to characterize user behaviors across communities.

6.3. SentiWordNet

SentiWordNet is officially promoted as a publicly available lexical resource that is ideally used for Opinion Mining. SWN extends WordNet synsets by associating with each of them three numerical scores: Objectivity(s), Positivity(s) and Negativity(s), which are a rough notation of how objective, positive, or negative the synset terms are.

To sum it all, SWN lists for each word the details of all of its senses where a sense is nothing but one of the common use of the word. For example, the word “casual” has 9 different senses listed in SWN. Few negative senses (where negativity-score > positivity-score) of this word are:

- *"an ability to interest casual students";*
- *"a casual remark";*
- *"a casual (or cursory) inspection failed to reveal the house's structural flaws";*

A positive sense (where *positivity-score* > *negativity-score*) can be:

- "*casual clothes*";

Moving on, for each sense, we have the following details:

1. Part of Speech,
2. Sense Rank,
3. Positivity Score,
4. Negativity Score, and
5. Objectivity Score ($1 - (\text{Positivity Score} + \text{Negativity Score})$).

Before the compilation of SentiWordNet, researchers used varying techniques to quantify the positivity or negativity of commonly used opinionated words [41]. However, promising results achieved in works [51] show that SentiWordNet is a valuable resource that can facilitate extraction of opinions and sentiments from varying texts of different natures, including commentaries as we will discuss later.

A direct parsing of the latest version of SentiWordNet was computationally infeasible because of the large number of lookups involved in the process as well as the unsorted nature of the data. We created a sorted and simplified list of all the words in SWN that enabled us to use binary search and achieve $O(\log N)$ time complexity. This version can be downloaded from our lab's website [83].

6.4. Method

Earlier, we described the community driven measure of popularity of a story posted at Digg, called Digg-Score. Stories can be roughly ranked based on these Digg Scores and a higher score represents a more popular story. In our research, we wanted to

train different classifiers to predict with maximum accuracy, the matured Digg-Scores of stories using information associated with the commentary done on those stories.

We trained our predictive models using comments from the first ten hours and the first fifteen hours. We also trained models using all the available comment information for comparison purposes. Here, we have tried to explain the effect of introduced semantic features of the comments on our prediction models.

In Table 2.1 (digg dataset statistics), we listed the breakup of our digg stories corpus into different categories. For each of these categories, we tried to build classifiers for the purpose of Digg-Score prediction by formulating three independent classification problems:

- (i) *a 2-class prediction problem,*
- (ii) *a 6-class prediction problem, and*
- (iii) *a 14-class prediction problem.*

The bins for the 6-class and 14-class prediction problems were set in Digg-score intervals of 1000 and 500, respectively. For the 2-class prediction problem we split the instances into the first class having all stories with a Digg-score of less than 1000, and the second class with Digg-score greater than 1000. This allowed for a uniform size distribution split. We used the decision tree classifier [31], the nearest neighbor classifier [52], and support vector machines [53] for performing the classification. We hypothesized that features of commentaries on stories differ in their influence, and hence, their affect on the performance of the classifier would also give varying results. Informally, it makes sense to correlate the popularity of a story with for example, the

number of opinionated words used in comments that followed the story. Or, it can be assumed that a story gets more hits if it becomes more controversial as more comments are posted on it.

6.5. Feature Description

As described earlier, for each word in SentiWordNet, there are different [positivity, negativity] sets of scores. One of these sets represents the most popular usage (called sense) of the word, and the rest follows it with decreasing popularity. While parsing the content of a particular comment, it wasn't easy to exactly identify the sense of some word used in the comment. So, we considered two sets of features and included them both:

- a. When considering weighted average of scores of all senses
- b. When considering the most popular score.

Number of Sentimental words

- a. Given a story, we counted all the words for which we have the sentiment scores from any of the senses of its occurrence in SentiWordNet.
- b. Given a story, we counted all the words for which we have the sentiment scores from only its most popular sense.

Example: Consider the word fidelity. In SentiWordNet, it has two senses. The most popular sense is neutral because its positivity as well as negativity scores are both zero. Hence, its occurrence in any comment would not count under the popular-sense strategy, i.e. (b). However, its 2nd most popular sense entails positivity and hence under

the weighted-average counting strategy, i.e. (a), an occurrence of the word “fidelity” in any story’s comment would contribute to the overall count of sentimental words.

Sum of Positivity

- a. Given a story, we sum up the positivity scores of all the words that were counted as sentimental words. As defined earlier, for each word, we calculated the weighted average of the positivity scores of all its senses:

$$\text{Weighted Positivity} = \frac{(\text{sense rank} * \text{positivity score})}{\text{sum of sense ranks}}$$

- b. In this case, for each word, we just picked its positivity score of the most popular sense.

Sum of Negativity

- a. Given a story, we sum up the negativity scores of all the words that were counted as sentimental words. As defined earlier, for each word, we calculated the weighted average of the negativity scores of all its senses:

$$\text{Weighted Negativity} = \frac{(\text{sense rank} * \text{negativity score})}{\text{sum of sense ranks}}$$

- b. In this case, for each word, we just picked its negativity score of the most popular sense.

In sum, we have twenty six features that we used to train our prediction models.

All the feature values were standardized using standard deviation:

$$\text{Normalized Value} = \frac{(\text{actual value} - \text{mean})}{\text{standard deviation}}$$

6.6. Classification Results

Table 6.1 - Prediction Results with Sentiment Features

# classes	Algorithm	Weighted ROC	F measure	% of correctly classified instances
2	Decision Tree	0.85	0.82	82
6	Decision Tree	0.75	0.66	67.4
14	Decision Tree	0.65	0.43	43.9
2	K Nearest Neighbor	0.85	0.79	79.12
6	K Nearest Neighbor	0.8	0.63	66.12
14	K Nearest Neighbor	0.7	0.4	43.31

Table 6.2 - Prediction Results without Sentiment Features

# classes	Algorithm	Weighted ROC	F measure	% of correctly classified instances
2	Decision Tree	0.87	0.82	82
6	Decision Tree	0.76	0.66	67
14	Decision Tree	0.65	0.43	44
2	K Nearest Neighbor	0.85	0.79	79
6	K Nearest Neighbor	0.8	0.63	66
14	K Nearest Neighbor	0.69	0.38	43

The above results are from a subset of the experiments we performed, and they show no clear sign of significant improvement in the prediction of Digg-score after we incorporate the sentiment-related features described previously. So, we conclude that a mere statistical analysis of semantic features isn't helpful enough in predicting the popularity of online content. Such a phenomenon exhibits a lack of a simplistic relation

between popularity of stories and naïve sentiments expressed in the follow up commentaries.

6.7. Discussion

A further exploration in this direction should focus on extracting the intention of a commenter when he writes a comment. It is possible to tag domain dependent words being used by people in their comments. Such a dictionary can be very useful when we combine it with the information about a comment's level. For e.g. when a person A replies to person B, we already know that A wants to address something in B's comment. What we don't know, however, is the nature of A's feedback – (a) it can be a refutation, (b) a statement of support, or (c) a mere expression that appends to whatever was stated by B. Recall that in a reply-answer network, there will be a directed edge from A to B. However, the edge itself doesn't tell us anything about nature of conversation between A and B. So, in a sense, in our reply-answer network and that of Slashdot's [2], we are making an assumption that an edge represents a weak friendship relationship. Using semantic analysis, we can get over this assumption and actually tag each of the links with tags like "agreement", "disagreement", or "neutral". Figure 6.3 demonstrates a typical scenario.

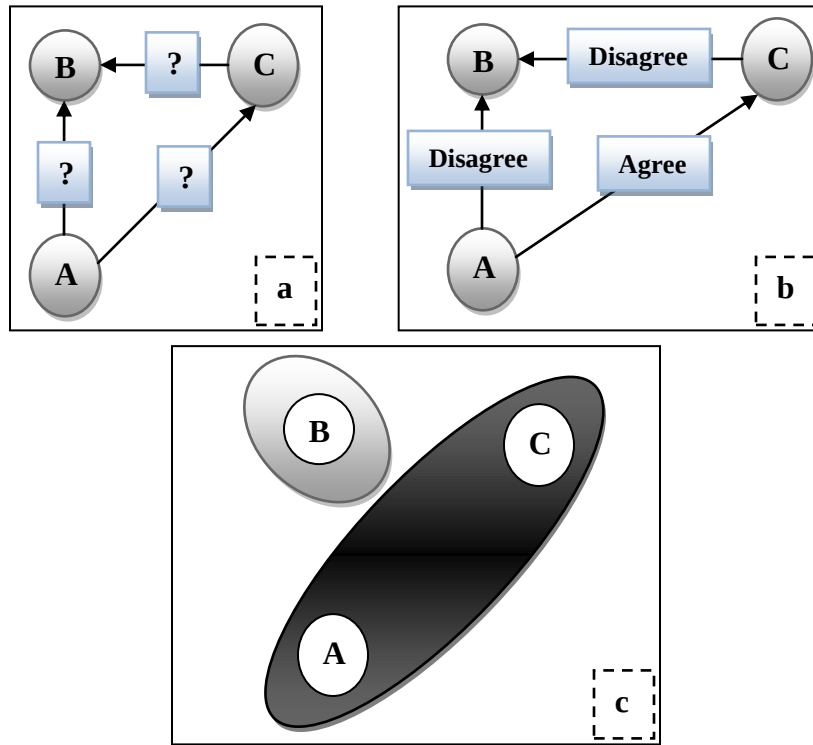


Figure 6.3 – Reply-network's edge between two users

Figure 6.3 shows how we can use the edge tags to identify the communities of users based on how they think about a certain topic of interest. It's showing a fragment of a comment thread where A and C replied to B and A also replied to C. Figure 6.3 (a) shows that in a typical reply-network with no tagging for edges, it's not possible to characterize user communities based on sentiments. Figure 6.3 (b) shows that for the same scenario, we can tag edges if we get to know the sentiment expressed in the respective replies. Figure 6.3 (c) shows that A and C should belong to a community that disagrees with the opinion of B.

Although, our existing state-of-the-art research methods aren't good enough to extract such sentiment information; but in the development of one such framework, we believe that the very first step would remain the same – i.e. the extraction of opinionated words and their intensity with respect to the norms of the underlying story's domain, but this alone does not suffice.

In the future, we plan to extract these opinions by plotting each comment against a Problem-Reduction model [54]. Such a technique is used in Expert System studies, where a problem is divided into sub problems, and until we reach the atomic problems that can be solved by mere knowledge of facts. Then the solution of these sub problems are combined and finally the root problem gets a weighted solution, which is nothing but a statistical solution suggested by its direct sub-problems.

The same analogy is applicable in the case of Opinion mining. Remember that our goal ideally is to tag each comment with one sentiment, although this can be tricky too because not all commenters are decisive in their suggestions about a particular topic they are discussing. However, by knowing the opinionated words used in the comment we already know the factual answers to the leaf nodes of our Problem-Reduction tree. Going up by one level, we should combine this information to devise the opinions expressed in individual sentences, and so on. Finally, we could combine these suggestive low-level opinions and tag the overall comment by a sentiment.

CHAPTER 7: TIMED EGONETS – TRACING PERIODICITY

7.1. Introduction

In Section 3.3, we demonstrated our ability to characterize users based on egonets of most active commenters across different Digg categories. This observation of communities was isolated for each of the eight categories, within which, we observed varying patterns of interactions.

Firstly, the egonets were generated for the most active few users of each of the eight categories. Also, it was a possibility for a particular user to show up in 2 or more list of most active users, i.e. a commenter who is amongst the most active people in multiple categories. To capture the interactions of these commenters and their neighbors, we assumed full knowledge of the dataset. We classified the interactions into two categories – interaction between users who belong to the same category (black edges), and interactions between users who belong to different categories (green edges). This categorization helped us to relate observations of egonets to heterogeneity of user interests. The dataset is analyzed a collection of stories spread across a span of approximately 500 days.

There are certain implications of this idea of full utilization of dataset – most importantly, the egonets were independent of any time factor. Is, the activity measure

taking into consideration the presence (or absence) of a constant user activity? What if the user was so active in only a few days that he turned out to be amongst the most active users who were constant in their activity throughout the 500 days? Should we distinguish between users who are sporadically active, versus constantly active? Does time play any role in the formation and comparison of egonets?

Here, we tried to answer these questions and others by introducing a relatively newer concept, which we call *Timed Egonets*. Previously, each of the active users was being represented by a single egonet. We extended this idea by generating multiple egonets for one user. The number of egonets per user is the number of bins or partitions that we created for the 500 days dataset. So, if we assume that there will be 5 bins, each of 100 days, then for each of the most active users, we will have 5 egonets. Each of these egonets would be generated by assuming the factual knowledge of only those days, which are covered by the particular bin representing the egonet. We hoped to capture insightful results by applying the concept to the same Digg dataset.

7.2. Process

The process of generating these egonets is as follows:

- First, we sort out in descending order the complete list of unique commenters by their cardinality, i.e. the number of comments they have posted over the duration of entire dataset. So, the user on top of this list is by all means the overall most active user.
- Then, we pick the top 15 users from this list.

- In the next step, we divided the entire dataset into 8 bins as reported in Table 7.1:

Table 7.1 - Classification of Digg stories into constant sized bins

2-month bins	# Stories	From Date	To Date
bin 0	4482	16 Nov 2007	15 Jan 2008
bin 1	3922	15 Jan 2008	15 Mar 2008
bin 2	4635	15 Mar 2008	14 May 2008
bin 3	4002	14 May 2008	13 Jul 2008
bin 4	4502	13 Jul 2008	11 Sep 2008
bin 5	5096	11 Sep 2008	10 Nov 2008
bin 6	4986	10 Nov 2008	09 Jan 2009
bin 7	5560	09 Jan 2009	10 Mar 2009
	37185	16 Nov 2007	10 Mar 2009

- Against each of the top user, an egonet was generated for each of the above-mentioned bins. So essentially, We had

$$15 \text{ commenters} * 8 \text{ story bins} = 120 \text{ Egonets}$$

7.3. Results and Discussion

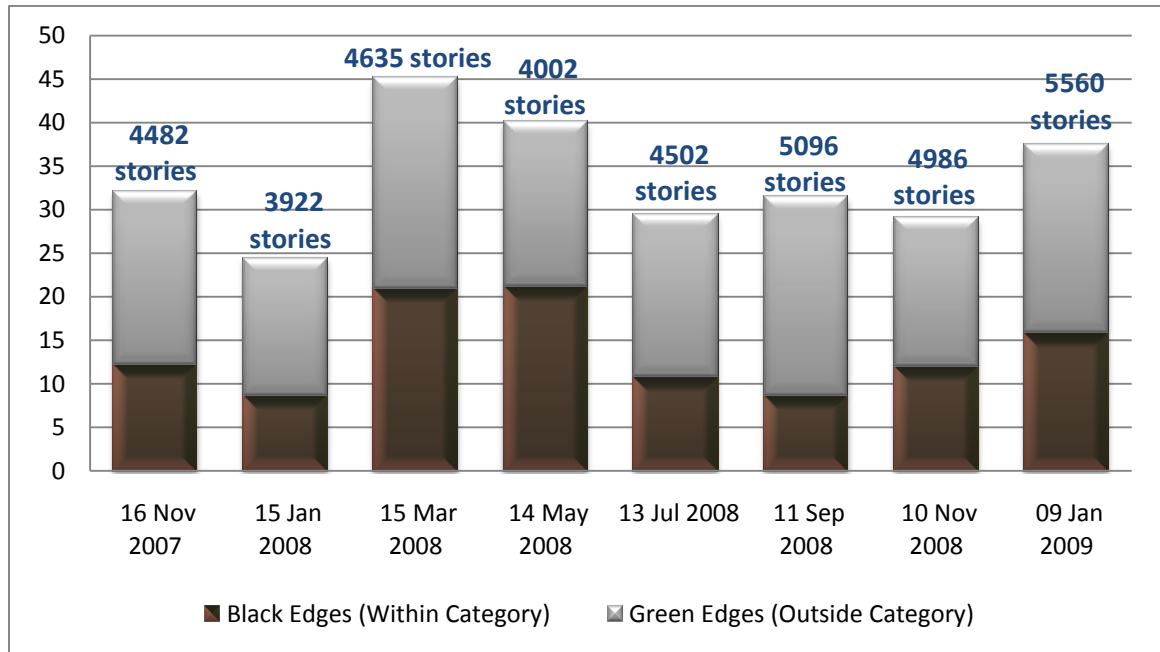


Figure 7.2 - Average Degree for the 16 most active users in two month periods.

In Figure 7.2 we present the average degree for the users (the average in-category degree and out-category degree as computed in Section 4.3). In essence, we derived the implicit social network between the users for the posts and comments posted within the two month period windows. We notice that there is an apparent increase in the average degrees for periods of four months from March 15, 2008 to July 14, 2008. Within the particular period there might be topics that have indulged users to increase in their usual pattern of activity. The period was known for the current president, Barack Obama

winning the Democratic candidate nomination and transitioning to campaign for the President’s office. We also notice that the average overall degree is correlated to the average in-category degree represented by the black colored edges in the egonets. This may suggest that as content becomes popular it tends to increase the homogeneity of user interests.

Recall that we use egonet analysis to understand the relationships amongst different users within the different categories. Similar analysis was done in [1] to differentiate between community of users that were discussion prone or not. A one level egonet for a user is defined as the user, the set of users who interact directly with the user (neighbors), and the relationships between those users. We can extend the definition of egonet to have neighbors who are N hops (links) away from the user in consideration.

As described previously, our 1-hop egonets have a two color coding scheme where a co-participation edge E_{ij} is colored black if both V_i and V_j have the same category membership, whereas edge V_i and V_j is colored green if users V_i and V_j belong to different categories.

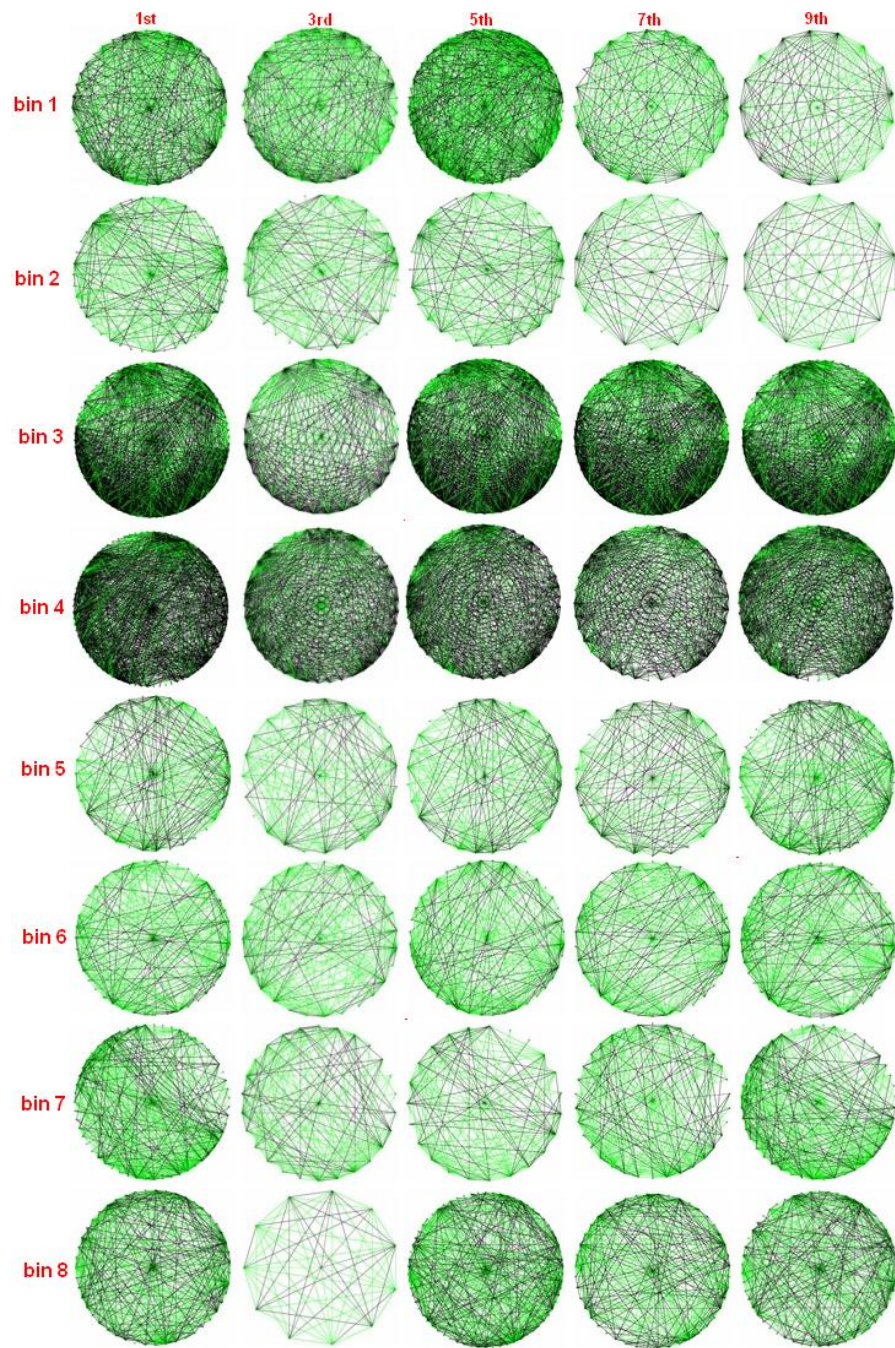


Figure 7.3 – Timed Egonets

Recall that we use egonet analysis to understand the relationships amongst different users within the different categories. Similar analysis was done in [1] to differentiate between community of users that were discussion prone or not. A one level egonet for a user is defined as the user, the set of users who interact directly with the user (neighbors), and the relationships between those users. We can extend the definition of egonet to have neighbors who are N hops (links) away from the user in consideration.

As described previously, our 1-hop egonets have a two color coding scheme where a co-participation edge E_{ij} is colored black if both V_i and V_j have the same category membership, whereas edge V_i and V_j is colored green if users V_i and V_j belong to different categories.

The idea of timed egonets is to trace the evolution of interaction behavior of active users, and somehow, characterize the evolution patterns of commenters in general, across different categories. Such an evolution could be correlated to world events that were incident at a certain point in time.

Earlier, we prescribed two criteria to analyze an egonet – Edge Densities, and Colored Densities. From the results, we could have expected a few outcomes:

- *Steady increase in edge density (and/or color densities) of a user's time egonets*
- *Steady decrease in edge density (and/or color densities) of a user's timed egonets*
- *Constant density (and/or color densities)*
- *Constant edge density but changing color densities.*
- *No visible pattern, i.e. abrupt changes in densities.*

Having mentioned that, let's discuss the timed-egonets shown above. Figure 7.3 shows the timed egonets for 1st, 3rd, 5th, 7th and 9th most active users for each of the 8 story bins. Each of the egonets for a particular commenter is mutually exclusive – meaning that it represents his interaction behavior for one specific chunk of stories contained in the respective story bin. The idea is to capture a trend of his activity over the time-sorted bins. A mutually inclusive set of egonets wouldn't tell us much except that each egonet would be to a certain extent, denser than the previous one.

Before analyzing the color variations, the very first thing that's apparent is the burst in density of egonets for bin 2 and bin 3, which is also evident in Figure 7.3. So, in a sense, the growth of social associations on social book-marking services is primarily dependent on the content that is being shared by the community. If the content pushes the community to comment, then indirectly, the strength of edges amongst the active users improves in due proportion. Next, we don't see any visible density evolution patterns. The densities of egonets, which are sorted by time, are rather presenting a random change. This behavior is unlike what can be observed in typical social networks. Apparently, one of the most important factors in improving the average edge strength in our implicit network is again the content that is being served by the Digg platform.

Finally, we see that for the same bins where we witnessed sudden burst in activities, the density of black color edges in particular, is higher as compared to green edges. There cannot be a hard and fast conclusion that can be made from this rough observation. However, to some extent, people tend to become more active in proportion to the people they know as being from the same interest group (the Digg category). Note

that, with each commenter we've an associated Digg category. This association is derived by classifying the comments a user posted into different Digg categories, and then, picking up the category that received highest number of comments. So, increase in the black-edge density means that the strength of a commenter's linkages with new people (and his existing neighbors) has increase.

CHAPTER 8: APPLICATIONS

8.1. “THE DIGG EFFECT”

To handle load, a website can be hosted on a number of parallel running web-servers to serve as much web requests as possible. This replication essentially increases the overall capacity of the web services, but, it incurs a cost, i.e. upfront investment in physical web servers, virtualization technologies, database replications, and load balancing switches, which is a risk in itself. Such a strategy might be effective and perhaps necessary for popular websites, but it’s unjustified when it comes to less popular websites, for which, the hit counts are at a low average, with some unpredictable bursts in hit counts.

When a popular website like Digg links to a smaller site, causing a massive increase in traffic, the smaller website may slow down or even temporarily shut-down; this phenomenon is called *The Slashdot Effect*, or *Slashdotting*, or *The Digg Effect*. The community of Slashdot and Digg also known as “flash crowds” [68], are responsible for this behavior because of the almost instantaneous millions of clicks that are generated for submissions promoted to the front page.

To understand better, let’s follow a simple example. Consider that we are hosting a personal video blog where we upload one popular video on a daily basis. Usually, our

webpage get an average of 200-300 hits per day and the capacity of your web server is good enough to handle this load. Figure 8.1 is a simulated graph that shows a typical visitors graph of your website for a random day.

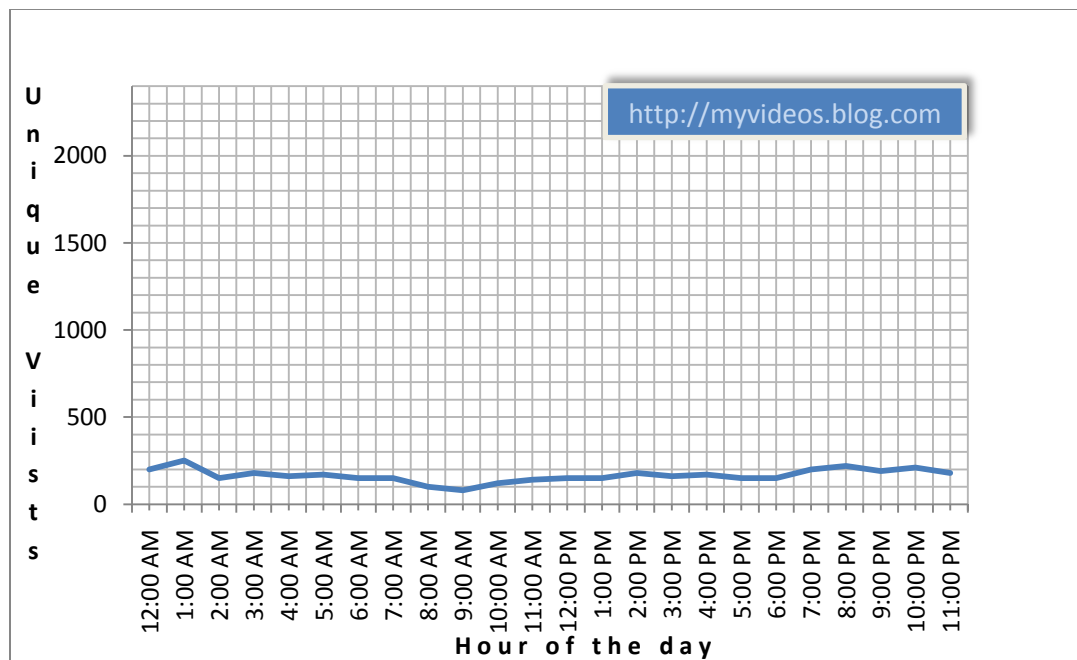


Figure 8.1 - Unique visits per hour - Normal behavior without Digg Effect

Now, suppose that someone in the internet community liked one of our video links and posted it on any of the popular social bookmarking services, like Slashdot, Reddit and Digg. Very soon, other people started liking your link, and eventually, the video link becomes so popular that it swiftly makes it way to the front page of the social bookmarking website.

If the website is vulnerable to the Digg Effect, then for the particular day, our visitor graph will look like Figure 8.2. At approximately 3:00 am, the website link makes it to the front page of Digg that result in flash crowds hitting the server at a much higher rate. For as long as the web server can sustain, the visit counts kept on increasing until the threshold is reached at 7:00 am, which results in thrashing and denial of services. Finally, at 9:00 am, the server is down, and no more requests are being served. With human intervention, the server is up again by 3:00 pm after 6 hours of denial of services. The visit counts start to grow again, and eventually, once the story gradually slides out of the front page, the visit count shrinks, and the trend would return to the normal behavior.

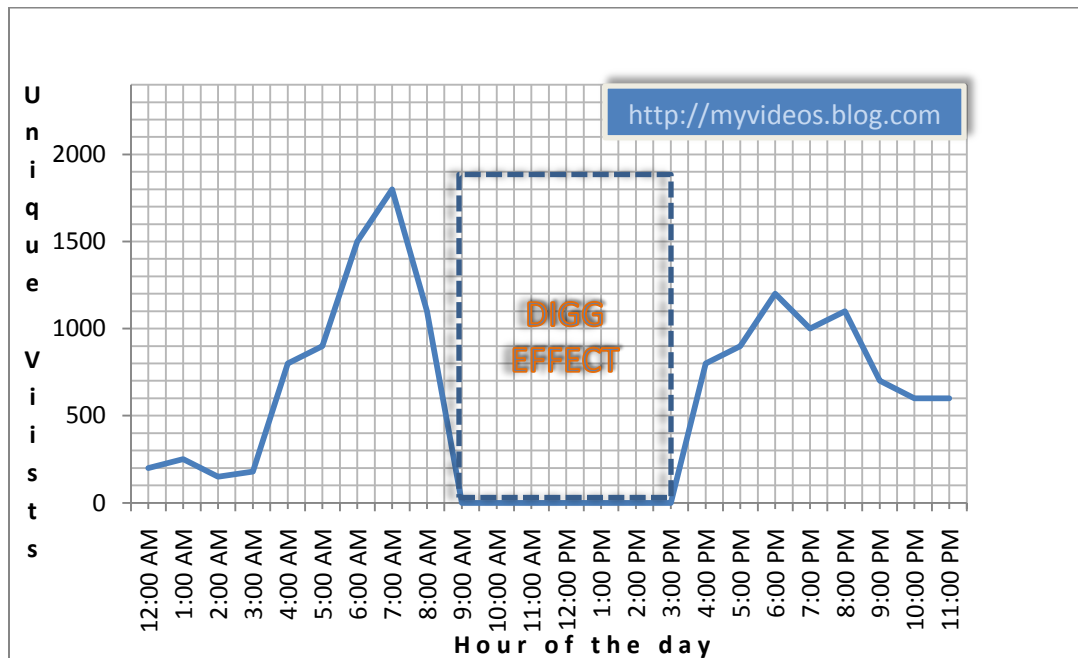


Figure 8.2 - Unique visits per hour – with Digg Effect [84]

Nowadays, it's not uncommon to observe such down-times for the websites that are digg-ed up as much as it takes for them to show up on the front page. There are various ways in which websites can remain prepared for such abrupt changes in visitors' traffic. Techniques like web site optimization [69], usage of accelerators [70] to reduce server load, and caching [71]. Such a constant readiness is expensive and perhaps, for many website owners it's impractical; rather, they'll prefer a down-time of a few hours over added expenditure. To discuss and analyze these preparation techniques is beyond the scope of this thesis, and can be assumed as a natural extension to our framework.

Using our digg-score prediction model, we can predict the chances of a social bookmark appearing on the front page of Digg. Apparently, none of the features that we used in our model were specific to Digg, and hence, the same model can be regenerated to cope with other social bookmarking websites. This information will enable the webmasters to consume their efforts only when it's feasible.

Let's revisit our video blog's example. Recall that we reported a 1.0-4.0% loss in multiclass classification accuracy while predicting the popularity score using the first few hours of comment data in comparison to all the available comment data. If we get to know beforehand with more than 60% accuracy that our blog's link is about to show up on the front page in the next couple of hours, then only, we can trigger the mechanisms that will expand the capacity of our web server. This can be a very effective technique in saving the costs that would have incurred if the anti-digg-effect measures were always in place, constantly.

Further work on this application is beyond the scope of this thesis. However, inclusion of this information while designing the load balancing architecture of a website can turn out to be an essential factor, and would be a future extension of this work.

8.2. Evolution of Opinions

Opinions play an important role in our lives. For most of things happening around us in the world today, we read opinions of authorities, governments, and independent critics. At the same time, we ourselves hold certain beliefs and opinions about these very same matters. There is a huge volume of opinionated feedbacks on the internet in various forms. These expressions exist on discussion forums, comment threads following news stories, first person video shots, and personal blogs, etc. The comment threads on social bookmarking services are ideally suited to a research, which focuses on summarization of these opinions. There are two types of influences that affect our opinions: *Internal Influence* and *External Influence*.

Almost always, for affairs that affect our personal lifestyles, our opinions are driven by the very nature of these affects: if we are not satisfied, then we are very articulate in expressing out our criticism and negative feedbacks – this is how the human beings operate, and such a biased opinion is a result of something that we call *Internal Influence*.

Marketers around the world are collectively one big lot of an influence that primarily operates on the principal of repetitive information reinforcement. On top of that, peer pressure existent in the communities where we live, be it online or real-world, forces us to reconsider our opinions, and often times, influence our stances so much so

that we flip sides in a matter of moments. Collectively, all of this can be called *External Influence*.

[72] created a computer simulation to observe the evolution of such opinions. The results were quite interesting - it was noted that if there is no external influence than a certain opinion will diminish overtime. If external influence is exerted, however, the same opinion strengthens, and achieves a 'balance'.

Each of the commenters within Digg is a denizen of a society where each one of them holds certain beliefs. Millions of people comment on thousands of social bookmarks which they themselves submit on a bunch of popular social bookmarking services. In these comments, they express positive and negative opinions, suggestions, and beliefs about current affairs, newly launched products, entertainment industry chronicles, and sports. Before posting these comments, they also take a sneak peak on what's already posted and more or less they choose their words accordingly.

When news story breaks, or when a new product is launched, it's almost certain that a related story will be posted on the social bookmarking websites. Such a posting becomes a medium for marketers and product developers to quantify the interest that people show in their products. In the same manner, if a person manually analyzes all the comments posted on a certain news topic, he/she will be able to give out subjective claims about the opinions of the people.

We hypothesize that these opinions drift over time and this drift can be traced out. Accordingly, the next natural step for sentiment analysis is to analyze and evaluate the drift in the sentiments over a period of time.

Imagine a situation where a new product, let's call it B, is launched recently and that comes in a direct competition to a similar popular product called A. These products improve over time, and with each improvement they take a share of customers from the other product's customer base. Before this new product was launched, people used to talk about the pros and cons of product A. Using our sentiment analysis techniques we can summarize the ratio of positive to negative feedbacks over different time slots, as depicted in Table 8.3.

Table 8.3 - Product popularity comparison based on summarized sentiments

Duration	Product	Positive (%)	Negative (%)	Neutral (%)	Event
<i>Jan - March</i>	A - Δ	80	10	10	A's market is mature, and B is creating a pre-launch hype.
	B - Δ	60	30	10	
<i>April - June</i>	A - $\downarrow$	70	20	10	Positive opinions about A decreased by 10%, and B has been launched.
	B - $\downarrow$	33	33	34	
<i>July - Sept</i>	A - $\downarrow$	55	35	10	A's positive opinions decreased further by 15%, and B is enjoying success.
	B - $\uparrow$	70	15	15	
<i>Oct - Dec</i>	A - $\uparrow$	65	25	10	The new stable set of opinions, where finally, B is more favorable than A.
	B - $\uparrow$	80	10	10	

The above table shows a timeline of sentiment drifts for two products, A and B, over the duration of one complete year. Against each time slot of 3 months, we have expressed sample numbers that represent percentage of summarized positive and negative opinions as well as neutral or irrelevant opinions about the products. Note that although

the above case study is about products; but, opinions can exist of any object - be it a product, person, event, organization, or a topic. Bing Liu [73] suggests that each of these objects can be represented as a hierarchy of components, and sub-components. Further, with each of the components, a set of attributes can be associated. Collectively, he called these components, sub-components, and attributes as features.

An extension of our framework, such that we could deduce these summarized opinions, was possible but beyond the scope. But the preliminary statistical tests and the problem settings we discussed in the chapter on sentiment analysis are ideal tools to progress in this direction. Apparently, the opportunities to infer inside information from the analysis as in the above Table 8.3 are endless. Forecasting the ideal launch dates of products can be easily done with such background knowledge. Halting or introducing further improvements, identifying potential future threats, discovering disliked features and changes, and many such questions can be answered when it comes to the consumer market. [74] concluded that simple text classification techniques would yield as impressive results as sentimental analysis, when it comes to the numerous informal political discourses that take place over internet forums.

CHAPTER 9: CONCLUSION AND FUTURE DIRECTIONS

This thesis presented a novel and unified approach for an in-depth analysis of highly active comment threads, which follow shared links on collaborative news aggregators, and which are collectively maintained by millions of commenters. All of the contributions of our work fall in a relatively newer domain, known as Comment Mining. In short, we (a) formulated a classification and regression based popularity prediction model, which enabled us to predict popularity index of digg stories using time-constrained comment data, (b) performed a comparative analysis of two variations of implicit social network formations: co-participation network and reply-answer network, (c) conducted insightful user characterization based on egonet analysis with and without time constraints, and (d) formed a basis for opinion mining by doing a preliminary analysis of comment content using SentiWordNet. All these experiments were performed against the Digg dataset.

The contributions of this thesis pertaining to (a) and (c) are document in our paper, titled “*Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis*” [79], which was published by IEEE for WISM’09-AICI’09.

Using Egonet analysis for projecting local neighborhoods, we identified the characteristics of highly active individual users with and without time constraints. The

time-based egonets effectively improved our ability to observe variations in activity patterns. Our framework to apply data mining techniques to these comments (and comment threads) helped us in predicting the popularity of news stories. We reported a very small loss of 1.0-4.0% in multi-class classification accuracy while predicting the popularity score using the first few hours of comment data in comparison to all the available comment data. This was achieved by exploiting the hidden trends and patterns as features of our classification framework. We also performed a comparative analysis of two network formations: co-participation and reply-answer. This helped us in relating these implicit networks that we formed with characteristic attributes of social networks. Further, we conducted preliminary experiments to improve the strength of a link in our co-participation network by analyzing the positive, negative or neutral sentiments expressed by users in their commentaries.

In future, we will enhance this framework to an extent where we are dependent on a minimal number of parameters. This automation should be done such that the parameters are always optimally selected based on a carefully selected set of network properties. We plan to explore opinion mining in hope of achieving a stronger prediction model. Our understanding of domain-specific sentiments expressed by millions of commenters can unleash a new set of ways in which we can link these commenters. The clustering of the resulting network formation would allow us to identify the niche of commenters having a certain sentiment about a certain topic. Following this, we want to further explore the opportunities to conduct dynamic network analysis of our networks.

There is a huge potential in our ability to understand the evolution of opinions if we could relate it to events that trigger people to flip their Judgement Opinions [41].

Another future direction to our thesis is our ability to demonstrate real-world scenarios where this information could be utilized. We want to focus on creating a durable and dependable set of tools that would enable advertisers to connect to the most accurate niche of audience. For example, the methods presented earlier to quantify the focus of commenters (using entropy measures) are highly effective in differentiating between people who are attached to a certain domain, versus people who are casual in their behavior on such communities.

A.1 DIGG STORY CRAWLER

The following PHP script was used to crawl the Digg dataset.

```
<?php
require_once 'Services/Digg.php';
Services_Digg::$appKey='http://salmanjamali.blogspot.com';

//-----
function printComments($story){
    global $doc, $commentCounter;
    $commentsItem=$doc->createElement( "comments" );
    $maxDate=time();
    $maxDateCounter=-1;
    while (1){
        if ($maxDateCounter!=0){
            $comments=$story->comments(array('sort'=>'date-desc',
'count'=>100));
            $maxDateCounter = count($comments->comments);
        }else{
            $comments=$story->comments(array('sort'=>'date-desc','count'=>100,
'max_date'=>$maxDate));
            If ((count($comments->comments) == 1) || (count($comments-
>comments) == 2)) {
                return $commentsItem;
            } else if (count($comments->comments) == 0) {
                return $commentsItem;
            } else {
                $maxDateCounter = count($comments->comments);
            }
        }
        if (count($comments->comments)!=0)
            foreach ($comments->comments as $comment) {
                $maxDateCounter--;
                $resultsArray = parseComment($comment, $maxDateCounter,
$maxDate);
                $commentsItem->appendChild($resultsArray["0"]);
                $maxDate = $resultsArray["1"];
            }
    }
}

//-----
```

```

function parseComment($comment, $maxDateCounter, $maxDate) {
    global $doc, $commentCounter;
    $commentItem = $doc->createElement("comment");

    $commentID = $doc->createElement( "id" );
    $commentID->appendChild($doc->createTextNode( $comment->id ) );
    $commentItem->appendChild( $commentID);

    $commentUser = $doc->createElement( "user" );
    $commentUser->appendChild($doc->createTextNode( $comment->user ) );
    $commentItem->appendChild( $commentUser);

    $ups = $doc->createElement( "ups" );
    $ups->appendChild($doc->createTextNode( $comment->up ) );
    $commentItem->appendChild( $ups );

    $downs = $doc->createElement( "downs" );
    $downs->appendChild($doc->createTextNode( $comment->down ) );
    $commentItem->appendChild( $downs );

    $commentDate = $doc->createElement( "date" );
    $commentDate->appendChild($doc->createTextNode( $comment->date ) );
    $commentItem->appendChild( $commentDate);

    $replies = $doc->createElement( "replies" );
    $replies->appendChild($doc->createTextNode( $comment->replies ));
    $commentItem->appendChild( $replies );

    $level = $doc->createElement( "level" );
    $level->appendChild($doc->createTextNode( $comment->level ));
    $commentItem->appendChild( $level );

    $root = $doc->createElement( "root" );
    $root->appendChild($doc->createTextNode( $comment->root ));
    $commentItem->appendChild( $root );

    $commentNumber = $doc->createElement( "number" );
    $commentNumber ->appendChild($doc->createTextNode(++$commentCounter)
);
    $commentItem->appendChild( $commentNumber );

    $content = $doc->createElement( "content" );
    $content->appendChild($doc->createTextNode( $comment->content ) );
    $commentItem->appendChild( $content);

    if ($maxDateCounter==0) {

```

```

    $maxDate = $comment->date;
}

$repliesItem = parseReplies($comment);

if ( $repliesItem !=0 )
    $commentItem->appendChild($repliesItem);
return array("0" => $commentItem, "1" => $maxDate);
}

//-----
function parseReplies($comment){
    global $doc, $commentCounter;
    $repliesItem = $doc->createElement( "replies" );
    $params = array('sort' => 'date-desc', 'count' => 100);
    $replies = $comment->replies($params);
    $numberOfReplies = count($replies->comments);
    if ($numberOfReplies==0)
        return 0;
    else if ($numberOfReplies > 0) {
        foreach ($replies->comments as $reply){
            $commentItem = $doc->createElement("comment");
            $id = $doc->createElement( "id" );
            $id->appendChild($doc->createTextNode( $reply->id ));
            $commentItem->appendChild( $id );

            $replyUser = $doc->createElement( "user" );
            $replyUser->appendChild($doc->createTextNode( $reply->user));
            $commentItem->appendChild( $replyUser );

            $ups = $doc->createElement( "ups" );
            $ups->appendChild($doc->createTextNode( $reply->up) );
            $commentItem->appendChild( $ups );

            $downs = $doc->createElement( "downs" );
            $downs->appendChild($doc->createTextNode( $reply->down) );
            $commentItem->appendChild( $downs );

            $replyDate = $doc->createElement( "date" );
            $replyDate->appendChild($doc->createTextNode( $reply->date));
            $commentItem->appendChild( $replyDate );

            $replies = $doc->createElement( "replies" );
            $replies->appendChild($doc->createTextNode( $reply->replies ));
            $commentItem->appendChild( $replies );
        }
    }
}

```

```

        $level = $doc->createElement( "level" );
        $level->appendChild($doc->createTextNode( $reply->level ));
        $commentItem->appendChild( $level );

        $root = $doc->createElement( "root" );
        $root->appendChild($doc->createTextNode( $reply->root ));
        $commentItem->appendChild( $root );

        $replyNumber = $doc->createElement( "number" );
        $replyNumber->appendChild($doc->createTextNode( ++$commentCounter
    ));
        $commentItem->appendChild( $replyNumber );

        $content = $doc->createElement( "content" );
        $content->appendChild($doc->createTextNode( $reply->content ));
        $commentItem->appendChild( $content );

        $repliesToReply = parseReplies($reply);
        if ($repliesToReply != 0)
            $commentItem->appendChild( $repliesToReply );
        $repliesItem -> appendChild ( $commentItem);
    }
    return $repliesItem;
}
}

//-----
try {

    $api = Services_Digg::factory('Stories');
    $storyCounter = 0; //number of stories crawled
    $maxSubmitDate = 1195222968;
    while($storyCounter <= 50000) {
        $params=array('count'=>100,'status'=>'popular','max_submit_date'=>
$maxSubmitDate);
        $res = $api->getAll($params);

        foreach ($res as $story) {
            global $maxSubmitDate;
            $doc = new DOMDocument();
            $doc->formatOutput = true;
            $commentCounter = 0;
            $stories = $doc->createElement( "stories" );
            $doc->appendChild( $stories );

            $dataFile = "All/datafile".++$storyCounter.".xml";

```

```

$fileHandle = fopen($dataFile, 'w+') or die("can't open file");
$storyItem = $doc->createElement("story");

$storyID = $doc->createElement( "storyID" );
$storyID->appendChild($doc->createTextNode( $story->id ) );
$storyItem->appendChild( $storyID);

$storyTopic = $doc->createElement( "storyTopic" );
$storyTopic->appendChild($doc->createTextNode($story->topic-
>short_name) );
$storyItem->appendChild( $storyTopic);

$storyContainer = $doc->createElement( "storyContainer" );
$storyContainer->appendChild($doc->createTextNode($story-
>container->short_name));
$storyItem->appendChild( $storyContainer);

$storyUser = $doc->createElement( "storyUser" );
$storyUser->appendChild($doc->createTextNode($story->user->name)
);
$storyItem->appendChild( $storyUser);

$storyDiggs = $doc->createElement("storyDiggs");
$storyDiggs->appendChild($doc->createTextNode($story->diggs));
$storyItem->appendChild($storyDiggs);

$storyLink = $doc->createElement( "storyLink" );
$storyLink->appendChild($doc->createTextNode( $story->href) );
$storyItem->appendChild( $storyLink);

$title = $doc->createElement( "storyTitle" );
$title->appendChild($doc->createTextNode( $story->title ) );
$storyItem->appendChild( $title );

$description = $doc->createElement( "storyDescription" );
$description->appendChild($doc->createTextNode( $story->
>description ) );
$storyItem->appendChild( $description );

$date = $doc->createElement( "storyDate" );
$date->appendChild($doc->createTextNode( $story->submit_date) );
$storyItem->appendChild( $date );

$commentsItem = printComments($story);

$numberOfComments = $doc->createElement( "numberOfComments" );

```

```

        $numberOfComments->appendChild($doc->createTextNode(
$commentCounter ) );
        $storyItem->appendChild( $numberOfComments );

        $storyItem->appendChild($commentsItem);
        $stories->appendChild($storyItem);
        $doc->save("All/datafile".$storyCounter.".xml");
        $maxSubmitDate = $story->submit_date;
        sleep(1);
    }
}
} catch (Services_Digg_Exception $e) {
    echo $api->lastCall . "\n";
    echo $api->lastResponse . "\n";
}
?>

```

A.2 DIGG DATASET DESCRIPTION

The following XML represents one single Digg story that was generated by the crawler (for brevity, only few comment tags are shown):

```
<?xml version="1.0" ?>
<stories>
<story>
<storyID>9056657</storyID>
  <storyTopic>health</storyTopic>
  <storyContainer>lifestyle</storyContainer>
  <storyUser>hdar3415</storyUser>
  <storyDiggs>386</storyDiggs>
<storyLink>http://digg.com/health/Better_beer_college_team_creating_ant
icancer_brew_3</storyLink>
  <storyTitle>Better beer: college team creating anticancer
brew</storyTitle>
  <storyDescription>College students often spend their free time
thinking about beer, but a group of Rice University students are taking
it to the next level. They're using genetic engineering to create beer
that contains resveratrol, a chemical in wine that's been shown to
reduce cancer and heart disease in lab animals.</storyDescription>
  <storyDate>1224247364</storyDate>
  <numberOfComments>10</numberOfComments>
  <comments>
<comment>
  <id>19922225</id>
  <user>supaklaw</user>
  <ups>1</ups>
  <downs>0</downs>
  <date>1224522562</date>
  <replies>0</replies>
  <level>0</level>
  <root>19922225</root>
  <number>1</number>
  <content>To science! "glook glook glook"</content>
</comment>
<comment>
  <id>19875599</id>
  <user>Half-Fast</user>
  <ups>1</ups>
  <downs>0</downs>
  <date>1224364535</date>
```

```

    <replies>0</replies>
    <level>0</level>
    <root>19875599</root>
    <number>2</number>
    <content>So how does that work with the old "The worse it tastes,
the better it works" corollary that my Mom used to use on me as a
kid?</content>
  </comment>
  <comment>
    <id>19867153</id>
    <user>atact88</user>
    <ups>1</ups>
    <downs>0</downs>
    <date>1224338936</date>
    <replies>0</replies>
    <level>0</level>
    <root>19867153</root>
    <number>3</number>
    <content>I should have done my thesis on beer...</content>
  </comment>
  <comment>
    <id>19866552</id>
    <user>jaymzdean</user>
    <ups>1</ups>
    <downs>0</downs>
    <date>1224336452</date>
    <replies>0</replies>
    <level>0</level>
    <root>19866552</root>
    <number>4</number>
    <content>Beer has maltose in it. Which is a sugar that feeds cancer
(see PET scan) and causes a release of insulin, which is a growth
factor for cancer.</content>
  </comment>
  <comment>
    <id>19864787</id>
    <user>damian7</user>
    <ups>6</ups>
    <downs>2</downs>
    <date>1224325898</date>
    <replies>0</replies>
    <level>0</level>
    <root>19864787</root>
    <number>5</number>
    <content>Double front page dupe, Digg loves beer <a class="user"
href="http://digg.com/food_drink/BioBeer_Fights_Cancer_and_Gets_You_Dru

```

```

nk">http://digg.com/food_drink/BioBeer_Fights_Cancer_a
...</a></content>
</comment>
<comment>
  <id>19849661</id>
  <user>AsSubtleAsABrik</user>
  <ups>8</ups>
  <downs>1</downs>
  <date>1224274769</date>
  <replies>0</replies>
  <level>0</level>
  <root>19849661</root>
  <number>6</number>
  <content>These types of people are the heroes of our
generation.</content>
</comment>
<comment>
  <id>19847899</id>
  <user>Trekhawk</user>
  <ups>3</ups>
  <downs>8</downs>
  <date>1224271084</date>
  <replies>2</replies>
  <level>0</level>
  <root>19847899</root>
  <number>7</number>
  <content>If they can put resveratrol in beer, why not put it in
other foods that won't cause you to swerve into on-coming traffic?
Cancer becomes a moot point when your head meets the windshield at
60mph.</content>
  <replies>
    <comment>
      <id>19866630</id>
      <user>jfujita</user>
      <ups>1</ups>
      <downs>0</downs>
      <date>1224336789</date>
      <replies>0</replies>
      <level>1</level>
      <root>19847899</root>
      <number>8</number>
      <content>RTFA. Beer only works because it is fermented. If you
really want resveratrol in your diet get it in pill form or drink
wine.</content>
    </comment>
  </comment>

```

```
<id>19865579</id>
<user>xero8472</user>
<ups>2</ups>
<downs>1</downs>
<date>1224331251</date>
<replies>0</replies>
<level>1</level>
<root>19847899</root>
<number>9</number>
<content>Get cancer and meet a windshield yourself.</content>
</comment>
</replies>
</comment>
<comment>
  <id>19847645</id>
  <user>FlyingPhotog</user>
  <ups>9</ups>
  <downs>2</downs>
  <date>1224270552</date>
  <replies>0</replies>
  <level>0</level>
  <root>19847645</root>
  <number>10</number>
  <content>Can it also prevent liver damage?</content>
</comment>
</comments>
</story>
</stories>
```

REFERENCES

REFERENCES

- [1] Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In WWW '08: Proceeding of the 17th international conference on World Wide Web, pages 665-674, New York, NY, USA; ACM, 2008.
- [2] Vicenc G´omez, Andreas Kaltenbrunner, and Vicente L´opez. Statistical analysis of the social network and discussion threads in slashdot. In WWW '08: Proceeding of the 17th international conference on World Wide Web, pages 645-654, New York, NY, USA; ACM, 2008.
- [3] Gilad Mishne and Natalie Glance. Leave a reply: An analysis of weblog comments. In Third annual workshop on the Weblogging ecosystem, 2006.
- [4] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. Journal of the American Society for Information Science and Technology, pages 1019-1031, 2007.
- [5] X. Liu, J. Bollen, M. Nelson, and V. Sompel. Co-authorship networks in the digital library research community. Information Processing and Management, 41:1462-1480, 2005.
- [6] Howard T. Welser, Eric Gleave, Danyel Fisher, and Marc Smith. Visualizing the signatures of social roles in online discussion groups. The Journal of Social Structure, 8(2), 2007.
- [7] <http://www.ebizmba.com/articles/social-bookmarking-websites>, last accessed on 11/28/2009.
- [8] M. E. J. Newman. The Structure and Function of Complex Networks. In SIAM Review, volume 45, pages 167-256, 2003.
- [9] Adamic, Lada A. and Buyukkokten, Orkut and Adar, Eytan. A Social Network Caught in the Web. In First Monday Journal, volume 8, June 2003.

- [10] Kerem Tomak and Mu Xia. In Book Series: Integrated Series in Information Systems by Springer US. Volume 1, Chapter 6:355-371, 2002.
- [11] Mariolis, P., Interlocking directorates and control of corporations: The theory of bank control, *Social Science Quarterly* 56, 425-439, 1975.
- [12] Mizruchi, M. S., *The American Corporate Network, 1904-1974*, Sage, Beverley Hills, 1982.
- [13] Michael Strong and David Eisenberg. In Book Series: Progress in Drug Research. by Birkhäuser Basel. Volume 64, 191-215, 2007.
- [14] Watts, D. J. and Strogatz, S. H., Collective dynamics of 'small-world' networks, *Nature* 393, 440-442, 1998.
- [15] Wasserman, S. and Faust, K., *Social Network Analysis*, Cambridge University Press, Cambridge, 1994.
- [16] Scott, J., *Social Network Analysis: A Handbook*, Sage Publications, London, 2nd ed., 2000.
- [17] Marsden, P. V., Network data and measurement, *Annual Review of Sociology* 16, 435-463, 1990.
- [18] K. Zhongbao and Z. Changshui. Reply networks on a bulletin board system. *Phys. Rev. E*, 67(3):036117, 2003.
- [19] Kumar, Ravi and Novak, Jasmine and Tomkins, Andrew. Structure and evolution of online social networks. *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 611-617; ACM, 2006.
- [20] User Guide of AutoMap: A software tool by CASOS @ CMU for Network Text Analysis.
- [21] Jin, Ruoming and McCallen, Scott and Almaas, Eivind. Trend Motif: A Graph Mining Approach for Analysis of Dynamic Complex Networks. *ICDM '07: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*. Pages 541--546, published by IEEE Computer Society, 2007.
- [22] Tong, Hanghang and Papadimitriou, Spiros and Sun, Jimeng and Yu, Philip S. and Faloutsos, Christos. Colibri: fast mining of large static and dynamic graphs. *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. Pages 686-694, published by ACM, 2008.

- [23] Lin, Yu-Ru and Chi, Yun and Zhu, Shenghuo and Sundaram, Hari and Tseng, Belle L. Facetnet: a framework for analyzing communities and their evolutions in dynamic networks. WWW '08: Proceeding of the 17th international conference on World Wide Web Pages 685-694, published by ACM, 2008.
- [24] Asur, Sitaram and Parthasarathy, Srinivasan and Ucar, Duygu. An event-based framework for characterizing the evolutionary behavior of interaction graphs. KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. Pages 913--921, published by ACM, 2007.
- [25] G. Szabo and B. Huberman. Predicting the popularity of online content. Technical Report HP Labs, pages 1-6, 2008.
- [26] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. Journal of the American Society for Information Science and Technology, pages 1019-1031, 2007.
- [27] X. Liu, J. Bollen, M. Nelson, and V. Sompel. Co-authorship networks in the digital library research community. Information Processing and Management, 41:1462-1480, 2005.
- [28] Kamal Nigam and Matthew Hurst. Towards a robust metric of polarity. In Computing Attitude and Affect in Text: Theories and Applications, 2006.
- [29] Xiaowen Ding, Bing Liu, and Philip S. Yu. A holistic lexicon-based approach to opinion mining. In WSDM '08: Proceedings of the international conference on Web search and web data mining, pages 231–240, New York, NY, USA; 2008.
- [30] Howard T. Welser, Eric Gleave, Danyel Fisher, and Marc Smith. Visualizing the signatures of social roles in online discussion groups. The Journal of Social Structure, 8(2), 2007.
- [31] Geoffrey Webb. Decision tree grafting. In In IJCAI-97: Fifteenth International Joint Conference on Artificial Intelligence, pages 846-851. Morgan Kaufmann, 1997.
- [32] David W. Aha, Dennis Kibler, and Marc K. Albert. Instance-based learning algorithms. Machine Learning, 6(1):37-66, 1991.
- [33] Vladimir N. Vapnik. The Nature of Statistical Learning Theory. Springer Verlag, 1995.

- [34] Andrew P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145-1159, 1997.
- [35] I. Witten and E. Frank. *Data mining: Practical machine learning tools and techniques*. 2005.
- [36] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- [37] Esuli, Andrea and Sebastiani, Fabrizio. SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation. 2006.
- [38] Hu, Minqing and Liu, Bing. Mining and summarizing customer reviews. *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2004.
- [39] Bo Pang and Lillian Lee. In *Foundations and Trends in Information Retrieval*. Vol. 2, No 1-2 (2008) 1-135, 2008.
- [40] Kim, Soo-Min & Hovy, Eduard. Crystal: Analyzing Predictive Opinions on the Web Proceedings of the Joint Conference on Empirical Methods in Natural Language processing and Computational Natural Language Learning (EMNLP-CoNLL). 2007.
- [41] Kim, Soo-Min & Hovy, Eduard. Determining the Sentiment of Opinions. *Proceedings of the International Conference on Computational Linguistics (COLING)*. 2004.
- [42] Kim, Soo-Min & Hovy, Eduard. Identifying and Analyzing Judgment Opinions. *Proceedings of the Joint Human Language Technology/North American Chapter of the ACL Conference (HLT-NAACL)*. 2006.
- [43] Christiane Fellbaum. *Book: WordNet An Electronic Lexical Database*. Published by The MIT Press, 1998.
- [44] Angela Fahrni, Manfred Klenner. Old Wine and Warm Beer: Target-Specific Sentiment Analysis of Adjectives. *AISB* 2008.
- [45] Ann Devitt, Khurshid Ahmad. Sentiment Polarity Identification in Financial News: A Cohesion-based Approach. *ACL* 2007.
- [46] Andrea Esuli, Fabrizio Sebastiani. PageRanking WordNet Synsets: An Application to Opinion Mining. *ACL* 2007.

- [47] Francois-Regis Chaumartin. UPAR7: A knowledge-based system for headline sentiment tagging. SemEval (Affective Text task), 2007.
- [48] Andrea Esuli, Fabrizio Sebastiani. Random-Walk Models of Term Semantics: An Application to Opinion-Related Properties. LTC, 2007.
- [49] Ethan Zhang, Yi Zhang. UCSC on TREC 2006 Blog Opinion Mining. TREC 2006 Blog Track, Opinion Retrieval Task, 2006.
- [50] Giuseppe Attardi, Maria Simi. Blog Mining through Opinionated Words, TREC 2006 Blog Track, Opinion Retrieval Task, 2006.
- [51] G. Szabo and B. Huberman. Predicting the popularity of online content. Technical Report HP Labs, pages 1-6, 2008.
- [52] [2] David W. Aha, Dennis Kibler, and Marc K. Albert. Instance-based learning algorithms. Machine Learning, 6(1):37-66, January 1991.
- [53] Vladimir N. Vapnik. The Nature of Statistical Learning Theory. Springer Verlag, 1995.
- [54] Gheorghe Tecuci and Mihai Boicu. Learning Agent Center Report: LEARNING-BASED KNOWLEDGE REPRESENTATION. Version 4, 2008.
- [55] Rapoport, A., Contribution to the theory of random and biased nets, Bulletin of Mathematical Biophysics 19, 257-277, 1957.
- [56] Rapoport, A., Cycle distribution in random nets, Bulletin of Mathematical Biophysics 10, 145-157, 1968.
- [57] J. Travers and S. Milgram, An experimental study of the small world problem. Sociometry 32, 425-443, 1969.
- [58] Barrat, A., M. Barthélemy, M., Pastor-Satorras, R., and Vespignani, A. The architecture of complex weighted networks. In the Proceedings of the National Academy of Sciences of the United States of America. Pages 3747--3752, March, 2004.
- [59] M. Girvan and M. E. J. Newman, Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99, 8271-8276, 2002.
- [60] P. Holme, M. Huss, and H. Jeong, Subnetwork hierarchies of biochemical pathways. Preprint cond-mat/0206292, 2002.

- [61] R. Guimer`a, L. Danon, A. D'iaz-Guilera, F. Giralt, and A. Arenas, Self-similar community structure in organisations. Preprint cond-mat/0211498, 2002.
- [62] A.-L. Barab´asi and R. Albert, Emergence of scaling in random networks. Science 286, 509-512, 1999.
- [63] L. A. N. Amaral, A. Scala, M. Barth´el´emy, and H. E. Stanley, Classes of small-world networks. Proc. Natl. Acad. Sci. USA 97, 11149-11152, 2000.
- [64] R. Monasson, Diffusion, localization and dispersion relations on ‘small-world’ lattices. Eur. Phys. J. B 12, 555-567, 1999.
- [65] K.-I. Goh, B. Kahng, and D. Kim, Spectra and eigen-vectors of scale-free networks. Phys. Rev. E 64, 051903, 2001.
- [66] I. J. Farkas, I. Der´enyi, A.-L. Barab´asi, and T. Vicsek, Spectra of "real-world" graphs: Beyond the semicircle law. Phys. Rev. E 64, 026704, 2001.
- [67] M. E. J. Newman. Mixing patterns in networks. <http://arxiv.org/abs/cond-mat/0209450>, 2002.
- [68] Jaeyeon Jung, Balachander Krishnamurthy, and Michael Rabinovich. Flash Crowds and Denial of Service Attacks: Characterization and Implications for CDNs and Web Sites, WWW10, WWW2002, May 7-11, 2002.
- [69] Andrew B. King. Book: Website Optimization, Publisher: O'Reilly Media, 1st edition, 2008.
- [70] <http://www.eaccelerator.net/>
- [71] <http://wordpress.org/extend/plugins/wp-cache/>
- [72] Guo Long, Chang Yun-feng and Cai Xu. The evolution of opinions on scale-free networks. In journal: Frontiers of Pyhsics in China. Higher Education Press, co-published with Springer-Verlag GmbH, Volume 1, Number 4, December, 2006.
- [73] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval. Vol. 2, No 1-2 pages 1–135, 2008.
- [74] T. Mullen, R. Malouf. A Preliminary Investigation into Sentiment Analysis of Informal Political Discourse. In Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, 2006.

- [75] D. Fisher, Marc. Smith, and Howard T. Welser. You are who you talk to: Detecting roles in usenet newsgroups. Proceedings of the HICSS, Hawaii, 3:56-59, 2006.
- [76] Steve Whittaker, Loren Terveen, Will Hill, and Lynn Cherny. The dynamics of mass interaction. In CSCW '98: Proceedings of the 1998 ACM conference on Computer supported cooperative work, pages 257-264, New York, NY, USA. ACM, 1998.
- [77] Kou Zhongbao and Zhang Changshui. Reply networks on a bulletin board system. Phys. Rev. E, 67(3):036117, Mar 2003.
- [78] Noor Ali-Hasan and Lada A. Adamic. Expressing social relationships on the blog through links and comments. 2007.
- [79] Salman Jamali and Huzefa Rangwala. Digging Digg : Comment Mining, Popularity Prediction, and Social Network Analysis. In the proceedings of WISM'09-AICI'09, published by IEEE, 2009.
- [80] Andrés Corrada-Emmanuel, Andrew McCallum, Padhraic Smyth, Mark Steyvers, and Chaitanya Chemudugunta. Social network analysis and topic discovery for the enron email dataset. Submitted to the 'Workshop on Link Analysis, Counterterrorism and Security' to take place at the 2005 SIAM International Conference in Data Mining., January 2005.
- [81] Dragomir R. Radev, Pradeep Muthukrishnan, Vahed Qazvinian. The ACL Anthology Network Corpus. Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, ACL-IJCNLP, pages 54–61, 2009.
- [82] Bradley Malin, Edoardo Airoldi, Kathleen M. Carley. A Network Analysis Model for Disambiguation of Names in Lists. Computational & Mathematical Organization Theory, 11, 119–139, 2005.
- [83] <http://cs.gmu.edu/~hrangwal/>
- [84] <http://www.ndesign-studio.com/blog/updates/the-digg-effect/>

CURRICULUM VITAE

Salman Jamali received his Bachelor of Science in Computer Science in 2006 from FAST - National University of Computer and Emerging Science, Karachi, Pakistan. He worked at Creative Chaos (Pvt.) Limited, Karachi for 4-5 months before joining GMU. He worked for two years as a Graduate Research Assistant with Dr. Huzefa Rangwala at Data Mining Lab, and previously with Dr. Gheorghe Tecuci at Learning Agents Center, both at GMU. Nowadays, he works at Rivet Logic Corporation in Reston, VA as a Software Engineer.