OPEN LARGE-SCALE ONLINE SOCIAL NETWORK DYNAMICS

A Dissertation

by

DANIEL JAMES CORLETTE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2011

Major Subject: Computer Science

Open Large-Scale Online Social Network Dynamics

Copyright 2011 Daniel James Corlette

OPEN LARGE-SCALE ONLINE SOCIAL NETWORK DYNAMICS

A Dissertation

by

DANIEL JAMES CORLETTE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Frank Shipman, III
Committee Members,	Heidi Campbell
	Yoonsuck Choe
	Ricardo Gutierrez-Osuna
Head of Department,	Valerie E. Taylor

May 2011

Major Subject: Computer Science

ABSTRACT

Open Large-Scale Online Social Network Dynamics.

(May 2011)

Daniel James Corlette, B.S.; M.S., California State University Sacramento

Chair of Advisory Committee: Dr. Frank M. Shipman, III

Online social networks have quickly become the most popular destination on the World Wide Web. These networks are still a fairly new form of online human interaction and have gained wide popularity only recently within the past three to four years. Few models or descriptions of the dynamics of these systems exist. This is largely due to the difficulty in gaining access to the data from these networks which is often viewed as very valuable. In these networks, members maintain list of friends with which they share content with by first uploading it to the social network service provider. The content is then distributed to members by the service provider who generates a feed for each member containing the content shared by all of the member's friends aggregated together. Direct access to dynamic linkage data for these large networks is especially difficult without a special relationship with the service provider. This makes it difficult for researchers to explore and better understand how humans interface with these systems. This dissertation examines an event driven sampling approach to acquire both dynamics link event data and blog content from the site known as LiveJournal. LiveJournal is one of the oldest online social networking sites whose features are very

similar to sites such as Facebook and Myspace yet smaller in scale as to be practical for a research setting. The event driven sampling methodology and analysis of the resulting network model provide insights for other researchers interested in acquiring social network dynamics from LiveJournal or insight into what might be expected if an event driven sampling approach was applied to other online social networks. A detailed analysis of both the static structure and network dynamics of the resulting network model was performed. The analysis helped motivated work on a model of link prediction using both topological and content-based metrics. The relationship between topological and content-based metrics was explored. Factored into the link prediction analysis is the open nature of the social network data where new members are constantly joining and current members are leaving. The data used for the analysis spanned approximately two years.

DEDICATION

To my family

ACKNOWLEDGEMENTS

I want to thank my wife Christina for supporting me throughout the process as I attempted to work fulltime outside the department and at the same time on my research. Without her help and support I would not have been able to complete my dissertation. I want to thank my mother, father, and brother for being understanding and supportive throughout the process especially when work and school was feeling like too much.

I want to thank my advisor Dr. Frank Shipman who provided a home within the department for me to pursue my research interests and guidance throughout the process.

I want to thank my committee for their support, encouragement and time. This is especially felt for Dr. Yoonsuck Choe who made himself available early on when I first entered the department for thought provoking discussions.

TABLE OF CONTENTS

CHAPTER Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

Social networking and blog sites are now the dominate destination for users of the internet in the U.S. According to Nielsen, in June 2010, 22.7% of the time U.S. users spent online was with social networking and blog related sites (Martin, 2010). This reflected a 43% increase in time spent online using online social networking platforms from June 2009. The next largest category taking up user's time online was online games coming in at 10.2%. The dominate destinations within the social networking and blog category was Facebook and Twitter. While the many features offered to users of social networking sites differ there are two general features that seem to define the social networking experience: the ability for users to manage a list of friends or subscribers and the ability to share or exchange content between these friends or subscribers. Each user is provided a feed where all of the shared content from their friends or subscribers is aggregated and viewable in near real-time.

The Nielsen report also highlighted that email and instant messaging usage had declined by 28% and 15% respectively providing insight that the social networking platforms maybe replacing these older forms of online communication. One interesting observation is that while the exchange of content through email and instant messaging has a more distributed architecture where the exchange takes place point to point between users, the online social networking architecture is more centralized.

This dissertation follows the style of Social Networks.

In the online social networking architecture users upload content to the social networking service provider which then distributes the content to its users. This architecture allows users to access their content at anytime from any location through the use of a web browser. Additionally, the service provider provides tools to its users to ease the process users go through when posting content such as text, pictures and video. With all of the friend lists and all of the users posted content being hosted centrally the opportunity to analyze this data exists and is made easier. Data of this nature provides a window into human interaction on a scale unknown until very recently. These new centralized social networking data sources provide the largest and most detailed record of human social interaction ever recorded. This fact has not gone unnoticed by either the service providers or those interested in marketing to the members of online social networks. The valuations placed on these networks range from the hundreds of millions to billions of dollars. The future impact of online social networking platforms is still unknown as they are new and undergoing a fast evolution. Currently there is no complete model of the dynamics of these systems. This can partly be attributed to online social networking systems being new. Additionally, access to complete datasets for use in research is both hard to obtain and computationally hard to process.

It is the view of this dissertation that avenues should be opened which allow academic study of these new social systems. A majority of the past research efforts which have focused their analysis on online social networks have examined research citation networks such as DBLP and arXiv. While it is true that citation graphs are essentially social networks, they are not the kind of social network represented by

Facebook, Twitter, or LiveJournal. A major reason for the use of DBLP and arXiv in past research efforts is that they are easy to access and they are both small and complete graphs. However there are many differences that set them apart from the type of online social network that is of interest in this dissertation. For example, the space of time that separates topological change within the citation networks is much larger than the space of time that separates change in LiveJournal and sites like it. Interactions between users of sites such as LiveJournal happen continuously and can affect topology immediately. The main reason more work has not been done analyzing sites such as LiveJournal is because access to such data is challenging even for small social networks. In cases where analysis has been applied to some of the larger online social networks often the researchers have obtained direct access to the social network data from the service provider. Approximately one third of this dissertation deals with the problem of accessing online social network data from the LiveJournal site. This dataset and the sampling methodology used to acquire it are discussed in detail in Chapter III. LiveJournal is one of the oldest online social networking sites on the internet. It was started in 1999 by Brad Fitzpatrick. Many of the features of LiveJournal have ended up as features in the newer and more well-known online social networks such as Facebook, Myspace, and Twitter. The architecture and size of the LiveJournal site made it practical and friendly for academic research.

The term "online social network" usually refers to sites that are architected for and focused on providing services to individuals rather than groups. These services usually provide for the hosting of personal profile pages, hyperlinks to other user profile

pages within the network and the ability to post media content of some form that is shared between members that are linked together within the same network domain. These three features are generally required for a website to meet the definition of an "online social network" or what is sometimes referred to as a social network service. A central purpose of an online social network such as LiveJournal is to facilitate the ability of its members to locate and track content from other members within the network domain through user profile pages and hyperlinks often referred to as "friend" links.

Part of what makes an online social network unique is that it facilitates interactions between members within a single domain space. In LiveJournal, the hyperlinks connecting users within the network domain, which are listed on each user's profile page, are referred to as "friend" links. The term "friend" is the label adopted by LiveJournal and many similar online social networks. Although these hyperlinks are referred to as "friend" links, the true semantics underlying these links can often differ from the commonly accepted meaning of two people sharing friendship offline. The term "friend", as it is commonly used within the context of an online social network and within this dissertation, only reflects the acceptance of members of the network to allow a hyperlink to exist between their profile pages and thus allowing members who have befriended a certain member to view that member's content that is posted to the network domain space. Users define the semantics of the term "friend" applied to hyperlinks within LiveJournal through their use of the linking facility that connects users within the network domain. Throughout this dissertation the terms "links", "friends" or

"friendships" all refer to the same hyperlinks that connect users within the LiveJournal network domain space.

Online social networks differ from both general blogging focused sites and general online communities in that the site architecture for the online social network is focused around individual users and their relationships between other users inside of the hosting network domain. Users of online social networks are generally not concerned with propelling their personal blogs or journals so that they can be recognized by large numbers of unknown users of the World Wide Web. This is often the situation with general blogging sites that are focused primarily on facilitating a user's ability to manage a blogging presence online. The topics that are written about by users of an online social network are more likely to be of a personal nature than what is commonly found on topic oriented blogs. LiveJournal's choice of the word "journal" over the word "blog" to represent the textual content posted by users is a reflection of the site being geared more toward personal expression rather than topic oriented blogging.

Sites often referred to as "blogging" sites or services have generally focused on supporting and providing tools to individuals interested in publishing content to wide audiences of unknown people and gaining their recognition. Many times the blogs of users of such services are hosted outside of the domain of the blogging service and the existence of links to other blogs, if they exist, generally use labels such as "follower" or "team member" and can point to pages outside of the network domain. Supporting links between users that point outside of the network domain space is a major difference from the type of links in an online social network such as LiveJournal. In such general

“blogging” focused sites, the notion of a community boundary is much less defined and the focus on achieving recognition for publish content is much higher than for an online social network such as LiveJournal. Users of general blogging sites such as Blogger.com are likely to be interested in monitoring their rank on an internet search engine such as technorati.com which provides measures of authority for the blogs they index in contrast to the users of an online social network such as LiveJournal who are much less likely to be interested in such a measure for their journal posts.

Online social networks differ from general online communities in that their main focus is not on facilitating community based discussions of topics as one would expect to see in an online community such as a bulletin board where the organizing element of such a network are topics rather than hyperlinks between users. The driving force for members of the online social network is to connect to other members within the network domain through friendship links facilitating the exchange of content. In an online community such as a bulletin board users generally navigate topics first and then examine posts by other members of the community as they relate to the topic. In an online social network the main hyperlinks used for navigation are a user’s friendship links and the main source of viewable content is the content posted by a user's friends that is aggregated into a single source for viewing.

A large amount of the research related to online social networks has focused primarily on topology, content generated by users, or the demographics of users. Work by (Kumar et al., 2004) studied the demographics of LiveJournal users in relation to friendships between users providing data that sought to explain the existence of

friendship between users in terms of their demographics. Work by (Balog et. al., 2006) focused strictly on the content of users blogs and provided a study temporal mood patterns exhibited within LiveJournal blogs. The methods used for collecting data from online social networks vary depending on the intent of the analysis used by researchers with some relying on the use questionnaires, interviews, their own personal observations, or the use of automated computational approaches. Oftentimes researchers that make use of questionnaires or personal interviews are interested in measuring very specific aspects of the humans that are using the on-line social networks such as (Ellison, Steinfeld, et al., 2007) which used a survey to measure what they termed "dimensions of social capital" in relation to 800 Michigan State University students who were Facebook users. The authors were interested in understanding how offline characteristics and behavior of users connected to their on-line usage. Sometimes when automated approaches to capturing data are used they are used in a limited fashion (i.e. snowball sampling) to capture a sample of specific subgroups within the network that are of interest to the researchers. An example of this approach is work by (Herring et al., 2007) which examined the static topology patterns between samples of users who posted to their LiveJournal journals using specific languages. The analysis of the LiveJournal social network presented in this dissertation uses computer monitoring to collect data from the LiveJournal site over a period of 2 years. Analysis is applied at a high abstract-level which views individual users and their friendships as nodes and edges in a mathematical graph representation. The choice of representation for users and friendships is dues largely because of the size of the data used for many portions of the

analysis (i.e. millions of users) and the methods used to acquire data. The data was collected using computer monitoring software which is only capable of capturing the online activity of the users of the social network as it is represented in a digital format. The goal of analysis is to better understand the link dynamics exhibited between users of the network over long periods of time and how both topology and the textual content of users blogs influences the networks evolution.

In addition to establishing a path for acquiring online social network data this dissertation also sought to contribute to the construction of a model of the network dynamics for these types of systems. A more complete model of the network dynamics would benefit both the designers and users of these systems. Understanding the forces which govern an online social network's topological and content dynamics would be valuable for the design of useful navigation tools such as friend recommendation. A general examination of some of the LiveJournal network dynamics was performed during work on this dissertation. As part of the exploration of the LiveJournal network dynamics a link prediction classifier was constructed and applied to the dataset. A study group of approximately 10,000 users was identified and followed over a 10 month period of time starting from the point at which they first became members of the LiveJournal network. Focusing effort on the construction of a link prediction classifier allowed for the testing of ideas related to the network dynamics. Out of this experimentation a new topological metric emerged that was capable of doubling, in some cases tripling the recall of link predictions without sacrificing precision. In addition to the exploration of purely topological metrics, content-based metrics were

also explored. Results from the exploration of content-based metrics showed that they performed poorly when applied without the use of the topological metrics. Topological metrics provided the best link prediction performance.

The dissertation is organized in the following manner. Chapter II provides a summary of other research efforts focused on contributing to models of online social networks and similar datasets. The chapter helps to understand how the research in this dissertation fits in with those past efforts. Chapter III provides a detailed examination of the application of a sampling technique capable of capturing most of the data from active users of the LiveJournal site. A large portion of the work in Chapter III is reproduced from (Corlette and Shipman, 2009). This sampling approach is referred to as event-driven sampling because it makes use of the event of a user's post appearing in the LiveJournal Atom feed to determine when to sample that user's friend list for changes. From the sample data that is acquired, a network model of the LiveJournal network is constructed which is capable of supporting the analysis of LiveJournal's network dynamics. In addition to capturing the network dynamics, the blog posts for all users are captured allowing for the analysis of the content exchanged between users. Chapter IV provides a detailed examination of the linking dynamics exhibited between users. Some of these results are used to support the creation of a predictive model of linking dynamics. Significant portions of Chapter IV are taken from (Corlette and Shipman, 2010). This chapter examines the practicality of applying a link prediction classifier to an open large-scale online social network such as LiveJournal. A practical application of the classifier would be for friend recommendation within online social networks. Both

precision and recall are examined within the context of varying training and test set data sizes. The metrics explored in Chapter IV are purely topological and do not incorporate the content of user blog posts. Chapter V examines augmenting the topological metrics examined in Chapter IV with metrics derived from the content of user blogs. All metrics in this chapter are based on the notion of document similarity and are computed using a document vector representation for user blog data and cosine similarity for computing document similarity. The metrics explored in Chapter V differ in terms of what constitutes a term in the document vector representation. The range of terms used in document vector representations span from the very simple, such as single space delimited tokens, to the complex, noun phrases with attached sentiment polarity. The range of terms was chosen to explore the effect of terms with varying degrees of attached semantics on document similarity metrics calculated between users. This was explored through the effect of the different document similarity metrics to augment the topological metrics from Chapter IV in link prediction. While metrics composed of the more semantically rich terms provided some increase in performance over the more semantically poor terms, the increase in performance was small. Content-based metrics in general provided a significant increase in link prediction precision over the topological metrics when used in conjunction with the topological metrics. However, content based metrics when applied by themselves to link prediction, without the topological metrics, performed poorly. Topological metrics performed much better than content-based metrics when applied alone. Chapter VI ends the dissertation with conclusions taken from the research and a discussion of potential future work.

CHAPTER II

ON-LINE SOCIAL NETWORK DYNAMICS RESEARCH

II.1 Introduction

The study of on-line social network dynamics is a very new area of research with many challenges. Its primary root is in the study of the static nature of large networks in general. Networks from a wide range of fields of study are often included in the general study of large networks. Some of these large networks include: protein regulator networks, research collaborations, sexual relationships, the internet and the World Wide Web (Barabasi, 2003). Research studying the static nature of large networks was energized with Barabasi's paper which showed the scale-free nature of the World Wide Web (Barabasi et al., 2000). Additionally, in 2001 Barabasi published the book "Linked" which opened up the field of large network analysis to the general public (Barabasi, 2002). Eventually studies were published examining the static nature of large social networks represented as hyperlinks between web pages and blogs (Adamic and Adar, 2005; Adamic and Glance, 2005; Mislove et al., 2007). Over time the notion of an on-line social network and the social hyperlink became stronger as hosting sites began forming that provided the ability for users to post their own content and to manage explicit links to others users within the same network. The focus of this dissertation is on the dynamics exhibited by these types of online social networks.

Fewer studies have examined the dynamic nature of large networks. Dynamics, as the term is applied to large networks often refers to how some aspect of the network,

most often topology, changes over time. The term network dynamics is also a reference to the underlying mechanisms that govern the change. Most studies that have examined network dynamics of online social network have done so at a very coarse level of detail focusing on global or macro-level metrics that characterize the networks as a whole with only minor mention of the underlying micro-level dynamics exhibited between individual users (Backstrom et al., 2006; Kumar et al., 2006). Examples of some of these macro-level metrics for topology include degree distributions, clustering coefficients and average path length between users. Fewer studies examining the dynamics of large online social networks exist largely because obtaining access to large complete datasets is challenging. Online social network data is often seen as possessing a large monetary value. Additionally, working with very large dynamic graph structures (tens of millions to billions of edges) can be very challenging in terms of computational resources. In this chapter an overview of the past and ongoing research targeted at both static and dynamic large network datasets, especially social networks, will be examined providing an understanding of where the research in this dissertation fits in conjunction with other research efforts.

The type of network represented by the LiveJournal dataset examined in this dissertation could be characterized as an example of a homogenous single mode social network where people are represented by blogs and connected explicitly by friendship links. The survey of link mining provided by (Getoor and Diehl, 2005) is an excellent introduction to some of the type of research problems associated with these types of datasets. Some of these research problems are: data representation, link-based object

ranking, link-based object classification, group detection, and link prediction. In this dissertation the problem of link prediction received special attention. While there was a lot of interest in examining the dynamics of the LiveJournal network in general there was also a goal of generating a predictive model of the network. The link prediction problem received a strong focus since it allows the scoring of a model's goodness based on how well it can predict future friendships between users. Developing a model of the dynamics of online social networks could provide many potential benefits. Understanding the dynamics of online social networks would aid designers and developers of online social network technology in the development of new social networking systems. Additionally, models of online social network dynamics could serve as support for existing theories of human social interaction and as insight that could lead to new theories. A very practical application of link prediction could be applied to friend recommender systems within existing or future online social networks. While many of the current online social networks provide a similar capability it is not well documented how successful these approaches are when applied to an open large-scale online social network such as LiveJournal.

The remaining sections of this chapter will discuss challenges associated with gaining access to data sources for study and will examine data sources that have been previously used for examining network static and dynamic properties. Additionally, prior attempts to apply link prediction to some of those datasets will be examined.

II.2 Access to Dynamic Network Data

The term "social network" has been used broadly in large networks research to represent the set of network structures that have either been digitized or have been gleaned directly from online sources and have a similarity to the kind of single mode homogenous network mentioned in Section II.1. Initial research on the static nature of large network structures revealed that most large networks, including different types of social networks, share many common global traits such as degree distributions that exhibit power laws, a high degree of clustering, and small world properties (Barabasi et al., 2000; Watts and Strogatz, 1998). Due to this initial recognition of common global network metrics among the many disparate large network datasets, the term "social network" has been loosely applied to any collaborative network of individuals. For example, DBLP and other citation networks are very popular datasets for large network research and are often referred to as "social networks". They have been the dominant source of network data to date for studies on both static and dynamic network properties of large networks (Hasan et al., 2006; Kunegis and Lommatzsch, 2009; Liben-Nowell and Kleinberg, 2003; O'Modadhain et al., 2006; Scripts, Tan, Chen et al., 2009; Scripts, Tan and Esfahanian, 2009; Tylenda et al., 2009; Wang et al., 2007). It is understandable for researchers to make use of the academic paper citation networks in their research on large social network dynamics because the citation networks are relatively stable, reasonable in size, and freely accessible. It is possible to obtain their entire dynamic network structure with time stamps on link formation. Unfortunately, these types of datasets vary greatly from common online social networks in use today (e.g. Facebook,

Twitter) both in terms of their size and the local dynamics exhibited between users (i.e. how users interact with the network and each other). The interval of time separating link changes in the citation networks is much longer (i.e. months or years) than conventional online social networks (i.e. minutes or days) such as Facebook or LiveJournal and this often goes unmentioned as to how this impacts research results.

LiveJournal was selected as the main source of social network data because such analyses are acceptable in its terms of use and it is close to sites such as Twitter and Facebook in terms of the type and rate of interactions that take place between users. In all of the online social networking systems such as LiveJournal users share information through their blogs, publish comments and respond to comments. Users of these types of networks span across all segments of society and are not limited to a narrow band such as research scientists as represented by the DBLP and other citation networks. Additionally, the size of the LiveJournal network leads to a dataset size that is orders of magnitude larger than DBLP and the other citation networks yet still manageable given the available resources for the research in this dissertation.

An important goal of this dissertation was to also look at the content of the communication between users and what effect this might have on a model of the network dynamics. For this reason both link dynamics between users and the content of user communication are required. In the case of LiveJournal this meant that an accurate timestamp of when a user's blog post was generated was needed. LiveJournal's unique architecture allowed for the collection of both which is discussed in detail in Chapter III.

There are many challenges to construct a dataset for study containing friendship link dynamics in conjunction with user posted content from an open large-scale online social network. These include hit rate limits imposed by the social network service provider, the large amount of resources required to repeatedly sample the network at some interval of interest, and determining which users should be sampled and how often. Hit rate limits refer to the number of times per second researchers can access the social network in an automated fashion without being banned. Most online social networking service providers limit automated access to their sites. For the major and most popular online social networks such as MySpace and Facebook, the access rate limits are either set too low or the size of the network is too large to allow repeated sampling on a reasonable timescale to capture link dynamics using conventional web crawling methods. Additionally, it was not practical to target MySpace or Facebook given their size and the available resources for research. LiveJournal did not have these issues and for this reason was chosen as the focus of the research in this dissertation.

The most attractive approach of acquiring dynamic network data is to get it directly from the social network service provider. This is the approach taken in (Kumar et al., 2006) which at the time it was published was the most comprehensive examination of on-line social network dynamics that had been seen. This approach often requires a special relationship with the service provider and is not available for all researchers interested in acquiring large complete network datasets containing link dynamics. In some cases service providers, such as LiveJournal, do not store the dynamic link data. Conventional approaches to capturing static linkage data from on-line social networks

are difficult to apply toward the capture of dynamic network link data. These approaches often require an initial seeding of a relation-based sampling algorithm (sometimes referred to as Snowball Sampling) with either random users or some predefined set of users chosen as points in the network to start a crawl. Sampling of users of the network takes place as friendship links are expanded outward using a Breadth-First strategy to a specified network depth (Caverlee and Webb, 2008). To capture link changes between users, this process needs to be repeated within some interval of time. The approach becomes prohibitive as the time interval between network snapshots decreases drawing heavily on the resources of both the researcher trying to acquire data and the social network service provider that is interested in providing service to its users. Acquiring just a single complete snap shot of a large on-line social network, even for one of the smaller on-line social networks like LiveJournal, is very challenging in terms of resources (Mislove et al., 2007). Choosing to crawl a subnetwork of a large social network has the potential for introducing bias in the resulting topology and therefore could potentially impact the analysis of the linking dynamics. There is also the potential that the resulting sampled subnetwork topology may not exhibit or maintain the scale-free and small-world properties of the whole network (Stumpf et al., 2005).

Conventional crawler-based approaches also suffer the inefficiency of sampling users based on their relationship to other users instead of the degree to which users contribute to the network's link dynamics. All users do not contribute to the networks link dynamics equally and therefore sampling all users at the same frequency is inefficient. A modification to the crawler-based approaches that would address the issue

of active and inactive users (who and when to sample) involves the possibility of modeling each individual user in the network and estimating a specific sampling rate for users based on their activity pattern in the network. Approaches that seek to maintain the freshness of search indexes follow a similar approach when deciding on when to examine a web page for changes that should be added to the search index (Grimes et al., 2008). These approaches share a similarity with Event-Driven Sampling (EDS), discussed in detail in Chapter III, which is defined in (Corlette and Shipman, 2009) in that both attempt to sample in proportion to the rate of change of the target. Approaches seeking to maintain search index freshness derive their model used to determine sampling rates from analyzing the updating behavior of target web pages. A model of the target website is built up over a period of time and is then used to predict when future sampling should occur. This differs from EDS which is directly informed on when to sample based on the frequency of user appearances in the Atom feed provided by LiveJournal. The EDS process along with its impact on the final dataset is described in detail in Chapter III. The following are some examples of the different approaches used by researchers to obtain access to data sets containing link dynamics that could be classified as on-line social networks.

Work in (Kumar et al., 2006) cites their work as “the first detailed evaluation of the growth processes that control online social networks in the large.” At the time of writing the authors resided at Yahoo Research with direct access to the two datasets used in their analysis, Flicker (www.flicker.com) and Yahoo! 360 (360.yahoo.com). Their work is an example of acquiring data directly from the backend storage of the online

social network service provider where they worked. Their dataset consisted of approximately 5 million members and 10 million friendship links with information regarding all event activity in the network. Their analysis focused on developing a model capable of explaining the evolutionary path of their networks at a global perspective.

Work in (Murata et al., 2007) presents an approach to link prediction based on "Weighted Proximity Measures". This study used a complete one month sample of the service of Yahoo! Chiebukuro (http://chiebukuro.yahoo.co.jp/) using 15 days of training and 15 days of test data. This dataset was provided to the authors by Yahoo. Results presented showed an increase in accuracy of predictions over the techniques of Newman's common neighbor metric, Adamic & Adar metric (Adamic and Adar, 2005), and preferential attachment (Barabasi, 2002).

Work in (Mislove et al., 2007) examined multiple online social networks at scale in some cases crawling 95% of the networks. Using automated scripts on a cluster of 58 machines they acquired the social networks of flicker.com, liverjournal.com, orkut.com, and youtube.com. Crawling runs spanned 1, 3, 39, and 1 days respectively for the fore mentioned networks. The focus of their network analysis was on static network structure. In the case of LiveJournal they captured the complete 15 million plus users (the total users at time of their writing) which required a 3 day period using their cluster of 58 machines. Their study especially highlights the challenges of acquiring social network data.

II.3 Models of Network Dynamics and Link Prediction

The science of large networks is still in its infancy. Duncan Watts stated that the science of networks will soon grow to a point as to be beyond the ability of any one person to learn in a lifetime (Watts, 2003). Due to computational limitations scientists often are forced to assume independence between the elements or objects of the systems they are studying. Increases in computational power has increased the ability for researchers to view the systems they are studying as a network of interconnected elements or objects where the connection between elements or objects explicitly model their relationships thus importing the science of large networks into their fields of study (Barabasi, 2003). The tendency in all fields is to reduce the system under study into a representation that is tractable in terms of computation and yet does not remove the essential elements that are necessary for accurate modeling (Getoor and Diehl, 2005). Therefore the domains that are aided by taking a network view of the domain will also be challenged both by the expanding knowledge of large scale networks and the challenges of determining a proper data representation within the domain of interest.

One of the first detailed studies of large scale social network dynamics was a study of Yahoo 360 and Flicker.com (Kumar et al., 2006). The study was conducted by researchers at Yahoo Research who had access to the backend storage of the network data because they were employed by the social network service provider. Their published research provided coarse grained details identifying high level user groups that they termed singletons, isolated communities, and the giant component. They provided a model of network growth that was capable of closely reproducing the values of the

global metrics of the analyzed networks. An interesting aspect of this research was the attempt to construct an agent-based model driven by local dynamics designed to replicate the global network metrics. In contrast to (Kumar et al., 2006) is the work (Gaston and desJardins, 2005) which attempted to learn an optimal network topology for autonomous agents to support team formation based on a set of fitness criteria. The two works contrast with each other in that part of the first work (Kumar et al., 2006) sought to identify lower level dynamics that were capable of creating the global network metric values thus reproducing the network structure to some degree. The second work (Gaston and desJardins, 2005) sought to identify the optimal network topology based on a set of predefined lower level dynamics and an overall system level fitness function. While the two studies differed in terms of their domains of interest and their research goals, together they provide a view into the struggle to understand the relationship between lower level dynamics and network topology. This is the goal of research in all domains that attempt to understand their systems through a network view. The goal is to connect the network topological structure with the individual local dynamics exhibited by the atomic elements of the network model for the domain. This is essentially the goal this dissertation is working towards within the domain of open large-scale online social networks.

Presently the challenges of bringing to light the dynamics of open large-scale online social networks are perhaps greater than they are for other domains due to the desire of social network service providers to maintain control of the data within their networks due to the large monetary value this data currently holds. A quick search on-

line will return a wide range of valuations for Facebook, the second most popular destination on the internet (Alexa, 2010), a social network with approximately half a billion users (Wortham, 2010). Valuations for the Facebook website today currently range from 14 up to 100 billion dollars. For a company with approximately 1,500 employees this is a staggering number. While many websites have a similar backend storage and access technology capability as Facebook, what separates Facebook from other websites is the amount and quality of social network data provided by its users. Because of the perceived financial value of social network data, access to large social network data has been limited and therefore open research targeted at these types of networks has also been limited.

The view taken in this dissertation is that a good model of a networks dynamics should have a predictive capability. A predictive capability refers to the model's ability to forecast what the network will look like topologically in the future given a present starting state. The view is that the stronger the predictive capability of the model, the more accurately the model is capable of capturing the network dynamics of the system. For this reason a strong emphasis is placed on what is termed the link prediction problem. A good portion of this dissertation examines the link prediction problem applied to the large-scale online social network known as LiveJournal. The first exhaustive examination of different topological metrics to support link prediction within a large network was (Liben-Nowell and Kleinberg, 2003). This study examined link prediction applied to the arXiv.org citation network using two static snapshots. One metric which showed the best promise and required the minimum amount of

computation was the Adamic/Adar metric first identified in (Adamic and Adar, 2005). This metric was used in the analysis of this dissertation applied to the LiveJournal network as a baseline metric to compare against attempts to improve link prediction. The use of two static snapshots in the (Liben-Nowell and Kleinberg, 2003) study did not take into account when the individual users joined the network (i.e. first published); this was something examined directly within the LiveJournal dataset collected and is discussed in detail in Chapter IV. Additionally, the size of the network used in (Liben-Nowell and Kleinberg, 2003) was three orders of magnitude smaller than the LiveJournal dataset examined in this dissertation.

II.4 Link Prediction Challenges

Presently, researchers who are interested in examining link prediction for open large-scale online social networks such as LiveJournal face many additional challenges beyond just the problem of gaining access to data. The link prediction problem for social network topologies, when viewed as a binary classification problem, has been shown to be a difficult problem due to the large amount of class skew that exists in training and test sets between the positive and negative edges (i.e. many more negative examples than positive examples) (Rattigan and Jensen, 2005). Two additional challenges to the link prediction problem, which are discussed below, are the ability to compare results across datasets and the ability to compare results when different evaluation metrics are used.

It has been mentioned previously that many known and easily accessible datasets share common global topological metrics such as average path length, degree

distribution and average clustering coefficient (Barabasi et al., 2000; Li et al., 2006; Watts and Strogatz, 1998) yet these datasets differ greatly in terms of the type of local interactions taking place between the members of the networks. For example, often citation networks such as DBLP, CiteSeer, and PubMed, email datasets such as the Enron Email dataset, and various biological networks such as BIOBASE are analyzed within the context of link prediction with references made to those datasets as being "social networks" (Hasan et al., 2006; Kashima and Abe, 2006; Liben-Nowell and Kleinberg, 2003; Rattigan and Jensen, 2005; Scripts, Tan, Chen et al., 2009; Scripts, Tan and Esfahanian, 2009; Tylenda et al., 2009; Wang et al., 2007). As (O'Modadhain et al., 2006) points out there are two main problems in link prediction: (i) prediction of new links, (ii) prediction of both new and repeated links. Citation networks such as DBLP, CiteSeer and PubMed, and the Enron Email dataset contain event-based data (i.e. repeated links, in the form of multiple paper collaborations or repeated emails between individuals) and do not directly represent the type of topological information represented by the explicit persistent social hyperlinks (friendship links) in networks such as LiveJournal, Twitter or Facebook. While it may be possible to infer explicit friendship links between users with the type of event-based data referenced above, in this dissertation the focus is strictly on the type of explicit user-defined friendship links in networks such as LiveJournal, Twitter and Facebook that are established one time only and persist until the friendship is terminated by one of the users. This type of friendship link exists as an explicit list of "friends" connecting the users of the social network.

These social hyperlinks form the main hyperlink structure that users of the social network use to navigate the system.

The type of metrics chosen by researchers to evaluate their approaches to link prediction varies widely and is largely dependent on the intent of analysis. Some researchers are focused primarily on how their link prediction approach compares to past approaches and are less focused on how their approach generalizes to new data (Hasan et al., 2006; Kashima and Abe, 2006; Kunegis and Lommatzsch, 2009; Scripts, Tan, Chen et al., 2009). In some cases the training and test sets are drawn from the same distributions (Hasan et al., 2006; Kashima and Abe, 2006; Scripts, Tan, Chen et al., 2009) which speaks little on how well the classifiers and models created will have a predictive power on future friendships. In some cases the inclusion of training and test set members is controlled by applying a threshold to nodes that determines inclusion in the training and test sets. For the (Liben-Nowell and Kleinberg , 2003) study a value of 3 was chosen which required collaborations to have occurred at least three times before a link was instantiated in their training and test set data. In the cases where the link prediction algorithms apply a ranking to nodes, some approaches only include the top k nodes in the ranking which does not reflect the true performance of the algorithm across all potential users in the network (Li and Chen, 2009; O'Modadhain et al., 2006; Wang et al., 2007). Complicating comparisons between approaches are cases where the test set is modified to adjust for the class skew problem where negative examples (i.e. links that will not occur) are down sampled (Scripts, Tan, Chen et al., 2009; Wang et al., 2007). Additionally the notion of recall, which requires one to limit how far into the future the

link prediction algorithm predicts, has largely been ignored. This is something that research that is part of this dissertation began to address in (Corlette and Shipman, 2010). One study provided recall metric values which dealt with recommendations from an online bookstore, but this study only involved 2000 randomly sampled users (Li and Chen, 2009).

II.5 Conclusions

There have been far fewer studies dealing with the dynamics of large networks than there has been dealing with their static structure. The primary reason for this has been a lack of access to dynamic network data. Presently there is a very broad definition of "social network" in the literature which is driven by the area being relatively new. In this dissertation the main research efforts are focused on developing a means to study and understand the type of open large-scale online social network represented by LiveJournal. LiveJournal is very similar to sites such as Facebook, Twitter, and MySpace in terms of the features offered to users. Many of the core features of Facebook, Twitter, and MySpace were first seen in LiveJournal.

This dissertation is especially interested in modeling the link dynamics of such a network and has chosen to place a special emphasis on the link prediction problem to accomplish this goal. The contributions to a model of online social network dynamics are many including providing insight for designers of future online social network systems and informing new models of human social interaction.

Due to the large monetary value currently placed on online social network data it is challenging to gain access to a large enough volume of this type of data to support research. At the outset of the research for this dissertation an initial study was conducted which examined the feasibility of obtaining dynamic link data along with the content posted by users within the large-scale online social network known as LiveJournal. It was determined that obtaining the link dynamics and user content was possible for a large majority of the public users using an Event-Driven Sampling (EDS) approach which is discussed in great detail in Chapter III and appears in (Corlette and Shipman, 2009). Up to that point in time only select researchers had been granted the access required to study the network dynamics of other large-scale online social networks at the scale that would support link prediction research for these types of networks (Kumar et al., 2006; Leskovec et al., 2008). A central motivation of the work studying the EDS approach to capture the dynamics of the open large-scale online social network LiveJournal was to not only support the research in this dissertation but to also open the door to other researchers interested in studying both the link dynamics of and the link prediction problem applied to open large-scale online social networks. The open nature of these networks (i.e. new users joining and existing users leaving) is something that had not been addressed previously in the literature. Studies which did examine link prediction did so using only a few static snapshots of the network without regard to the length of time individual users were members of the network.

CHAPTER III

EVENT-DRIVEN SAMPLING AND THE LIVEJOURNAL DATA*

III.1 Introduction

This chapter examines the application of an Event-Driven Sampling (EDS) approach to the LiveJournal social network. The intention of the EDS approach is to capture the content of user blogs in conjunction with friendship link dynamics over time with a time step of a single day and with high accuracy. The EDS approach makes use of the "always on" Atom feed provided by LiveJournal that contains all public blog posts in near real-time to inform the sampling process of user friendship networks. This has the effect of targeting sampling towards the public active users of the network. The EDS approach is shown to be capable of maintaining 98% daily correctness across all user friendship link dynamics for the class of users that are both public and active. Additionally, the group of public active users represents approximately 85% of the active network link mass. Analysis shows that the network model maintains both small-world and scale-free properties common in large networks like social networks. Data used for the analysis of the EDS technique spans a period of seven months and involves the analysis of data from 4.8 million users and approximately 34 million friendship links. To the best of the author's knowledge, this study, which appears in (Corlette and Shipman, 2009), is the first to look at and analyze the use of an "always on" Atom feed

* © [2009] IEEE. Reprinted with permission, from Proceedings of the 12th IEEE International Conference. on Computational Science and Engineering: Symposium on Social Intelligence and Networking, Capturing on-line social network link dynamics using event-driven sampling. Corlette, D., Shipman, F., Vancouver, BC, Canada, 284-291. http://doi.ieeecomputersociety.org/10.1109/CSE.2009.287. For more information go to http://thesis.tamu.edu/forms/IEEE%20permission%20note.pdf/view

like the one provided by LiveJournal to inform an event-based sampling process targeted at capturing user blogs in conjunction with user link dynamics over time within the context of an on-line social network. While the procedure and analysis discussed in this chapter is targeted at the construction of a dataset to allow the study of an open large-scale online social network it also provides insights into possible methodologies that could be employed by competing social network service providers to capture a competitor's social network data. As recently as this month (10/2010) Google's CEO Eric Schmidt was quoted as saying that Google is interested in acquiring the social graph data Facebook currently possesses (Weintraub, 2010). While the methodology of using event data to capture link dynamics that is presented here is specific to the LiveJounal social network, the principles could be applied to other networks where user event-based data is available, such as data available from the user installed applications on Facebook. The computational cost for using an event driven approach to capture dynamics link data is shown to be low. The capturing of the LiveJournal social network data outlined in this chapter was accomplished using a single desktop PC with 2.4 GHz processor and 512MB RAM with additional hard-disk space as needed. The system used to analyze the LiveJournal data was larger consisting of a 5 PC cluster. The main requirement for the cluster stemmed from the need to keep entire LiveJournal graph in physical RAM to make access times to individual nodes of the graph reasonable.

On-line social networks have fast become one of the most popular destinations on the Web. According to Alexa.com, as of this writing, the top five websites are; Google, Facebook, Youtube, Yahoo, and Live. All five sites listed offer social

networking services facilitating the sharing of content using the social networking format with Facebook being a strictly social networking platform. While there have been many studies that have looked at the static structure of on-line social networks (Adamic and Adar, 2005; Adamic and Glance, 2005; Barabasi et al., 2000; Barabasi, 2002; Mislove et al., 2007) there have been noticeably fewer studies that have focused on their structural dynamics (Kumar et al., 2006; Leskovec et al., 2008; Murata and Moriyasu, 2007). There have not been studies that combine analysis of a network's structural link dynamics in conjunction with the content posted by users, something which is done in this dissertation. The lack of research focusing on the dynamics of large-scale on-line social networks seems to stem from the fact that dynamic network data is very difficult to obtain. Obtaining access to an as-complete-as-possible dataset for the construction of a corpus of blog posts in conjunction with the user friendship link dynamics connecting these blogs taken from an on-line social network was a central motivating factor for the research presented in this chapter. An additional desire was that the final corpus span a long period of time (eventually, greater than a year) with a small time Δ (approximately a single day) between dynamic link events allowing the examination of relationships between the natural language in user blogs and the structural network dynamics exhibited both locally and globally in the social network. It was a desire that the structural properties of the friendship network topology connecting users in the corpus maintain a small average distance between users, scale-free topology and small-world properties which are the expected properties in large networks like the LiveJournal social network.

The EDS approach makes use of the "always on" Atom feed provided by LiveJournal which contains all public posts by LiveJournal users in near real-time. The friendship networks for public users are sampled in proportion to their appearance in the Atom feed. When applied to the LiveJournal social network Atom feed, EDS captures all blog posts for public users in conjunction with all friendship link dynamics for public users with approximately 98% daily correctness across the entire network link mass of active public users. This value was computed by performing a complete verification crawl for all public active users in the network model generated through EDS. Correctness was calculated by taking the ratio of all link additions and removals in the network model and all link additions and removals seen in the verification crawl. Analysis showed that the public users for which both blog and link dynamics are captured comprise 85% of the total LiveJournal network linkage mass of active users. A user is determined to be active if they have posted during the prior 5 months. The analysis estimates that 15% of active LiveJournal users post privately or at least disallow their posts to be pushed to the Atom feed. Structural analysis of the network model generated through EDS shows that the network maintains small-world and scale-free properties. These results as very beneficial for researchers seeking a means to construct a corpus of both user blogs in conjunction with user friendship link dynamics within the context of an on-line social network. An added benefit of the EDS approach when applied to the LiveJournal site is that the daily hit rate limit for LiveJournal is not violated. EDS can run continuously allowing for the continuous monitoring of user link dynamics and user blog posts updates.

One possible criticism of the EDS approach is that it requires the LiveJournal Atom feed. Presently LiveJournal is the only on-line social network to provide such feed. However, the massively popular micro-blogging on-line social network, known as Twitter, had announced during the time research on EDS was taking place that they would be offering a similar feed to LiveJournal's Atom feed that they called the "fire hose" which would output in near real-time all public tweets (micro-blog posts) (Twitter, 2009). A down sampled version of the “fire hose” has been offered since that time with higher rates of access possible with permission from Twitter. Presently the business models of the most popular large-scale on-line social networks such as Facebook and Myspace restrict complete and open access to their networks. While researchers can still access these sites with web-based crawlers, the type and rate of access required for building a corpus of user posts in conjunction with friendship link dynamics with a daily interval like EDS does for the LiveJournal site is not currently possible.

III.2 Challenges Collecting Dynamic Network Data

There are many challenges to construct a dataset for study containing friendship link dynamics in conjunction with user posted content from a large-scale on-line social network. These include hit rate limits imposed by the social network service provider, the large amount of resources required to manage and analyze the data acquired from repeatedly sampling the network at some interval of interest, and determining which users should be sampled and how often. Hit rate limits refer to the amount of times per second researchers can access the social network in an automated fashion. Most social

networking service providers limit automated access to their sites. For the social networks examined, the combination of access rate limits being set to low and the size of the network being too large prevent the use of conventional web crawling methods to capture user friendship link dynamics.

The most attractive approach of acquiring dynamic network data is to get it directly from the social network service provider. This is the approach taken in (Kumar et al., 2006) which was the most comprehensive examination of macro-level on-line social network dynamics seen to date at the time research was progressing. This approach often requires a special relationship with the service provider and is not available for all researchers interested in acquiring large complete network datasets containing link dynamics. In some cases service providers, such as LiveJournal, do not store the dynamic link data. Conventional approaches to capturing static linkage data from on-line social networks are difficult to apply toward the capture of dynamic network link data. These approaches often require an initial seeding of a relation-based sampling algorithm (sometimes referred to as Snowball Sampling) with either random users or some predefined set of users chosen as points in the network to start a crawl. Sampling of users of the network takes place as friendship links are expanded outward using a Breadth-First strategy to a specified network depth (Caverlee and Webb, 2008). To capture link changes between users, this process needs to be repeated within some interval of time. The approach becomes prohibitive as the time interval between network snap shots decreases drawing heavily on the resources of both the researcher trying to acquire data and the social network service provider. Acquiring just a single complete

snap shot of a large on-line social network, even for one of the smaller on-line social networks like LiveJournal, is very challenging in terms of resources (Mislove et al., 2007). Choosing to crawl a subnetwork of a large social network has the potential for introducing bias in the resulting topology and therefore the link dynamics of users in the corpus. This means that the resulting sampled sub network topology may not exhibit or maintain the scale-free and small-world properties of the whole network (Stumpf et al., 2005).

Conventional crawler-based approaches also suffer the inefficiency of sampling users based on their relationship to other users instead of the degree to which users contribute to the network's link dynamics. All users do not contribute to the networks link dynamics equally and therefore sampling all users at the same frequency is inefficient. A modification to the crawler-based approaches that would address the issue of active and inactive users (who and when to sample) involves the possibility of modeling each individual user in the network and estimating a specific sampling rate for users based on their activity pattern in the network. Approaches that seek to maintain the freshness of search indexes follow a similar approach when deciding on when to examine a web page for changes that should be added to the search index (Grimes et al., 2008). Approaches to maintaining search index freshness share a similarity with Event-Driven Sampling in that both attempt to sample in proportion to the rate of change of the target. Approaches seeking to maintain search index freshness derive their model used to determine sampling rates from analyzing the updating behavior of target web pages. A model of the target website is built up over a period of time and is then used to predict

when future sampling should occur. This differs from EDS which is directly informed on when to sample based on the frequency of user blog post appearances in the Atom feed provided by LiveJournal.

III.3 Details of the LiveJournal Dataset

In this section the organization of the LiveJournal social network and the type of data from the site that was used to build the corpus and perform the analysis presented in this dissertation is discussed. LiveJournal is one of the oldest and first websites to offer the social networking platform for users. It has been active for more than 10 years. Many of its features are now features of the larger more well known online social networks. The LiveJournal site is focused primarily on the hosting of individual user journals or blogs. At the time of the investigation into applying EDS to the LiveJournal network the site had 18 million users with approximately 11 million users having ever having posted to their blogs. Approximately 1 million unique users on average are active during a one month period. This low ratio of active users to total user accounts is common within on-line social networks since most sites offer free membership. Many people create an account and check out the site, but far fewer individuals commit to frequent long term usage. In addition to allowing users the ability to maintain their own blog, LiveJournal also offers the ability for users to create and maintain communities. A community can be viewed as single blog shared by a community of members functioning in a similar fashion as a message board. Depending on the privacy settings all members of the community can read and post to the single community blog. In the research presented in

this dissertation the focus is strictly on the individual user blogs and the user friendship links that connect these blogs together. In addition to each user having a blog, each user also has a profile page that is used extensively in the analysis of the relationship between public and private users.

In summary, the three main data sources examined from the LiveJournal site include the "always on" Atom feed containing all public user posts in near real-time, FOAF (Friend of a Friend) data which provides for each user a list of their friends and all other users of the network connected to the user, and user profile data which contains, among many other items, a date when the blog was first created and the last time the blog was updated. The main elements used from the user profile data include the "Creation Date" value providing the day the user created the account, and the "Last Updated Date" which provides the last day the user's personal blog was updated.

Privacy settings can be applied in many ways within the LiveJournal site. The privacy setting which affects our analysis is whether or not a user has allowed their posts to be placed within the public Atom feed. The value of this setting determines whether a user is considered public or private. Regardless if a user is public or private, the profile page for each user is accessible. Accessibility of all users' profile pages was a key factor in our ability to perform analysis of the structural relationship between public and private users allowing us to determine whether or not those users identified as private were also active.

Friendship links in LiveJournal are directional. As long as a user's privacy settings allow others to view their blog it is possible for users to befriend another user

without reciprocation. The act of befriending another user can be viewed as subscribing to that user's blog. The blogs of all of a user's friends are continually aggregated on the user's "friends" page so that a user's friendship page will list in chronological order the posts from blogs the user has subscribed.

III.4 Event-Driven Sampling and the Network Model

In this section the Event-Driven Sampling approach is defined in more detail along with the other major data collection activities applied to the LiveJournal site. In this section an important distinction is drawn between what is referred to as the network sample, which is obtained directly through Event-Driven Sampling, and what is referred to as the network model which is constructed from the network sample data. An initial observation regarding the relationship between users' posting activity and the frequency of changes made to a user's friendship network is also discussed which initially motivated the investigation and support of the Event-Driven Sampling approach.

III.4.1 Event-Driven Sampling

A unique feature that separates LiveJournal from most other on-line social networks is the "always on" Atom feed that contains a stream of all public posts made by LiveJournal users to their blogs in near real-time. It is access to this Atom feed that makes the Event-Driven Sampling approach possible. An "event" is defined as the act of witnessing a user's post within the Atom feed during a 24 hour period. A user's friendship network is sampled if that user was seen in the LiveJournal Atom feed during a 24 hour period. This approach to sampling social networks differs significantly from

other approaches that require an initial set of "seed users" used to start a web crawler that incrementally builds its network sample by following the friendship links of users. In the web crawler-based approach the sampling of users is driven by the user friendship topology. In the Event-Driven Sampling approach the sampling is driven by the frequency of blog postings by public users. The Event-Driven Sampling approach directs sampling toward the active users of the network. On average approximately 80k to 120k users are seen in the LiveJournal Atom feed in a 24 hour period.

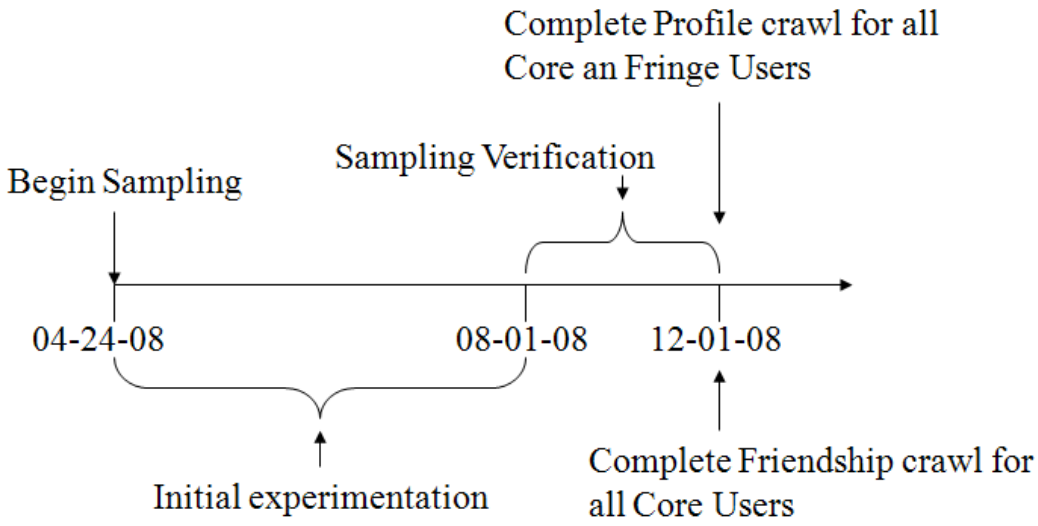

Fig. 1. Data Collection Activities

III.4.2 Data Collection Process

Data collection using the EDS approach began on 10-01-06 and ended 07-01-09. Three major data collection efforts were relevant to the study characterizing EDS. The data collection activities and their relevance to the EDS study are shown in Fig. 1. The portion of the LiveJournal data collected used in the EDS study in this chapter spanned

the period 04-24-08 to 12-01-08, approximately seven months. During this period of time all public blog posts for 1.4 million users were collected along with approximately 40 million friendship link dynamic events. A friendship link dynamic event includes the addition or removal of a friendship link between users. For each day EDS was applied to the LiveJournal network a folder was created containing all of the public posts for the given day along with all friendship linkage data for each user who publically posted to their blogs during the prior 24 hour period. The seven month period of sampling shown in Fig. 1 is divided into two sections. The first section spanning 04-24-08 to 08-01-08, referred to as "Initial Experimentation" in Fig. 1, represented a period of experimentation. The second section spanning 08-01-08 to 12-01-08, referenced as "Sampling Verification" in Fig. 1, is the period used to provide a detailed verification and structural analysis of the resulting network model constructed from data taken using EDS. These results are presented in Subsections III.5.3 and III.5.4.

Two more data collection efforts were carried out which were required for the portion of analysis focused on the verification of the Event-Driven Sampling approach. Both of these additional data collection efforts were carried out during the period of 12-01-08 to 12-06-08. The first collection effort starting on 12-01-08 involved a complete crawl of all user profiles for both public and private users seen during the 04-24-08 to 12-01-08 period. Profile data for 4.8 million users (1.4 million public users and 3.4 million private users) was captured. The second collection effort starting on 12-01-08 involved the collection of all friendship linkage data for the 1.4 million public users seen during the period 04-24-08 to 12-01-08. During this period friendship data was collected

for each of the 1.4 million public users regardless of posting activity. This complete crawl of the friendship data for all public users was used to verify the accuracy of the final network model generated from the network sampled data spanning the period 04-24-08 to 12-01-08.

III.4.3 Network Sample vs. Network Model

A distinction is drawn between the network sample obtained through Event-Driven Sampling taken from the LiveJournal site and what is referred to as the network model. The network sample is the collection of dated directories obtained through Event-Driven Sampling that contain all public user posts made during a 24 hour period along with all of the friendship link data for users who posted during this 24 hour period. The network model is the model of the LiveJournal network that is constructed from the network sample data. Construction of the LiveJournal network model can start from any point after sampling began. To build a network model a starting date is chosen equal to or after the date sampling began and the dated directories are advanced through in order. For each dated directory (i.e. folder 04-24-08) software reads in all of the friendship link data for users seen that particular day and makes modifications to the friendship links in the network model by adding new links when users have added new friends and removing links when friendships have been removed. When new users are seen in the sample they are added to the network model along with their friendship network. In order to accommodate the large number of users and links in the sample (1.4 million users and approximately 40 million links events, when crawling from 04/24 to 12/01) a giant hash table of hash tables (hash of hashes) data structure was constructed that

spanned multiple PCs. The first level of the (hash of hashes) table was indexed using a user's ID and the second level of the (hash of hashes) table was indexed using the ID of a user's friends. Software was written to allow the expansion of this data structure to whatever size was needed by simply adding a new PC. In order to maintain acceptable latency when accessing elements of the network model it was required that the entire network model reside in physical RAM within the PCs that held the network model. The entire LiveJournal network model constructed from approximately 1 years worth of network dynamics required the use of 4 PCs to store the network and a single PC to coordinate data access. The network model used for EDS verification and structural analysis was constructed starting with network sample data from 08-01-08 to 12-01-08.

III.4.4 Relationship between User Posting and User Friendship Link Changes

During a portion of the initial experimental phase shown in Figure 1 spanning from 05/01/08 to 06/01/08 the relationship between user posting rates and the corresponding rates of change experienced by a user's friendship network were examined. Two numbers received focus; the number of posts a user had made and the number of changes experienced in the user's friendship network. A Pearson correlation coefficient between these two numbers of 0.36 was calculated showing a weak but positive correlation. Users who post more often experience more change in their friendship network. The correlation coefficient was calculated initially to gain some insight into the possibility of there being a relationship between rate of posting and rate linkage changes. To further investigate this relationship a random sample of 8000 users was taken from the initial set of public users seen during the period 05-01-08 to 06-01-

08 with these users being tracked daily for network link changes over a two month period spanning 06-01-08 to 08-01-08 regardless of whether or not they made posts to their blog. Fig. 2 shows the percentage of the 8000 users sample who posted N days after they made a change to their friend network. Approximately 21% of the users posted the same day when they made a change to their friendship network. After 9 days approximately 50% of users had eventually made a post to their blog after they made a change to their network. It is important to note that when a user posts to their blog and is then sampled, any changes that have been made to their friend network between posts are captured. Further analysis performed during this period showed that the rate of change of the network as a whole, as it is represented in the sampled data, is low in comparison with the size of the network with approximately only 0.4% of the network links experiencing change daily.

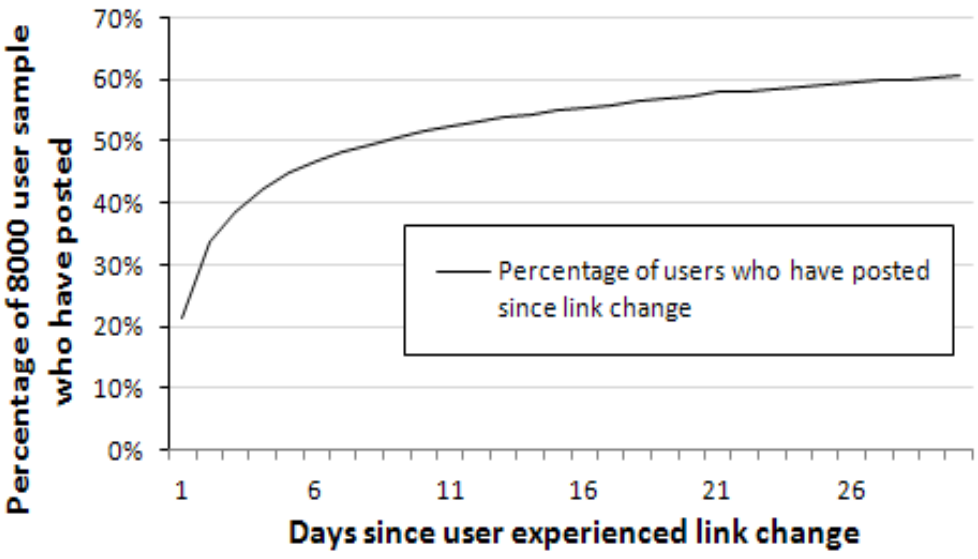

Fig. 2. Relationship between Posting and Link Changes

III.5 Construction and Verification of the Network Model

This section describes the process and details of constructing the network model from the network sample generated through Event-Driven Sampling. Constructing the model requires an initial ramping up phase as users are added to the model for the first time. This ramping up period could be overcome by an initial exhaustive crawl of the network; however an exhaustive crawl of this type is likely to violate the hit rate limits depending on the rate of crawl chosen. LiveJournal has an automated hit rate limit of 5 hits per second or 432000 hits per day for a single user. It would take approximately 40 days to capture the entire LiveJournal network of 18 million users from a single IP address, if hit rate limits are respected. For researchers wishing to capture and track user link dynamics over time a majority of the complete crawl referenced above would be wasted since approximately 1 million users, or 6% of all users in the network, are active during a 1 month period.

The next four subsections provide details related to the construction of the network model and details regarding the verification of this model in relation to LiveJournal network. In Subsection III.5.1 the process of constructing a network model using network samples from the period 08-01-08 to 12-01-08 is shown. This period of time is referenced in Fig. 1 as Sampling Verification. Subsection III.5.2 provides data on the relationship between public and private users where it is determined that 85% of the network topology of the active portion of the network is captured. Subsection III.5.3 discusses verification of the resulting network model. Subsection III.5.4 provides a

structural analysis showing that the resulting network model maintains the structural properties one would expect from a large-scale on-line social network.

III.5.1 Network Model Creation Process

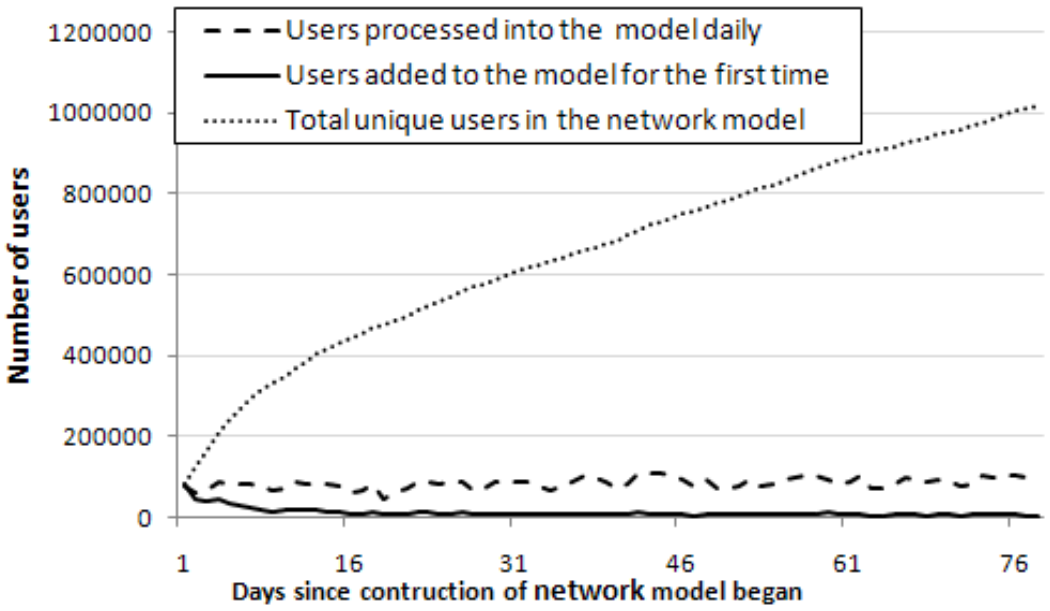

Fig. 3. Rate of Network Model Growth

Fig. 3 shows the initial transient phase as construction of the network model progresses. This ramping up period occurs as the first daily sampled user friendship link data obtained through Event-Driven Sampling is added to the network model. Initially when this process begins there is no information about any users in the network model and so every new user seen during this transient period provides many new links to the network model. Eventually the rate of growth for the network model in terms of new users and new links stabilizes.

The three parameters shown in Fig. 4 can be used to give an estimate of when the network model has stabilized. The model begins to stabilize when it has left the rapid transient period where it is initially flooded with new unseen users and the rate of both unseen users and the addition of new friendship links being added approach a constant.

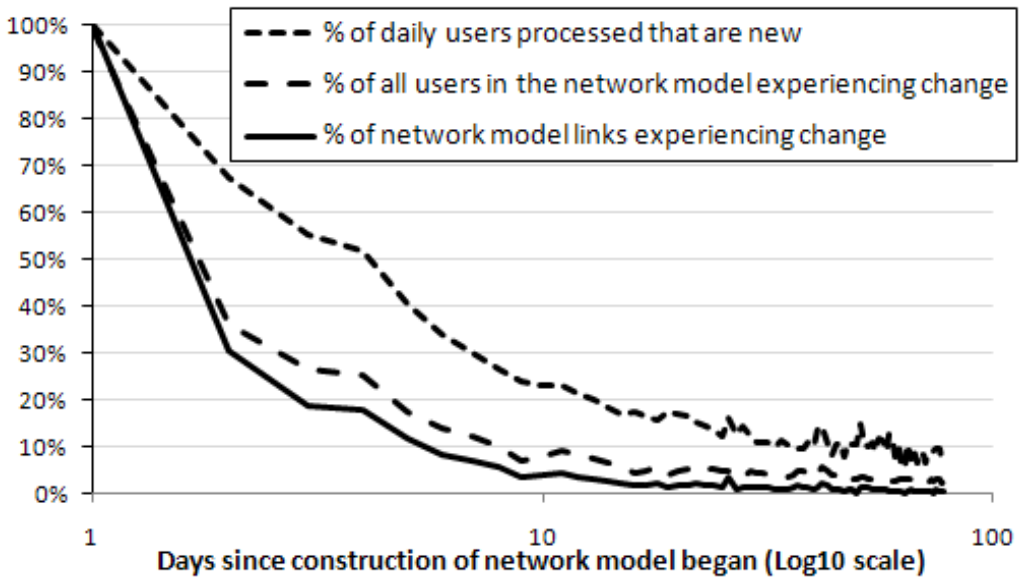

Fig. 4. Percentage of the Network Model Experiencing Change

The first parameter shown in Fig. 4 is the number of users seen in the daily sample that are new to the network model. Of the three parameters listed this parameter stabilizes last with a value close to 10% after 30 days. This means that after 30 days into the construction of the network model 10% of the users read into the model from the network sample are new to the model. The second parameter shown in Fig. 4 is the percentage of users added to the network model that experience a change in their friendship network. This parameter stabilizes with a value of approximately 4% after 15

days. The third parameter shows the percentage of network links in the network model that are experiencing change. This parameter stabilizes around the same time as the second parameter with a value of approximately 2% after 15 days.

III.5.2 Public, Private, Active and Inactive Users

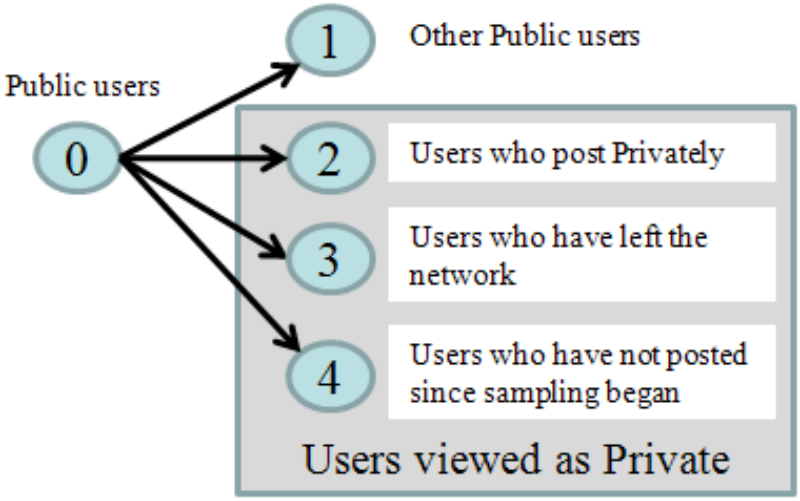

Fig. 5. Four Types of Friendship Links in Network Model

As the network model is constructed from the samples generated during Event-Driven Sampling only the outward directional links of users are considered. This leads to users in the network model having the four types of connections shown in Fig. 5. The first type of connection (0 → 1) is a connection between two users who post publically both having appeared in the Atom feed and sampled by Event-Driven Sampling. The second type of connection (0 → 2) is a connection between a public user seen in the Atom feed and another user of the LiveJournal site that has selected for their blog posts to be private. There are multiple levels of privacy in LiveJournal which adds some complication to the analysis. Users can select for their posts to be public to other users of

the LiveJournal service but not included in the LiveJournal Atom feed. Users can select for their posts to be public to their friends but private to all other LiveJournal users, including the Atom feed. A user is considered "private" if they have selected for their blogs to not be included in the Atom feed. The third type of connection (0 → 3) connects a public user seen in the Atom feed to another user who has left the network. The user who has left the network could have been either a public or private poster. Further distinctions are not drawn for the (0 → 3) connection since their impact on the network model is the same. The fourth type of connection (0 → 4) connects a public user seen in the Atom feed and another user who posts publically to their blog but has not posted during the period of sampling. For purposes of the analysis both (0 → 3) and (0 → 4) type of connections can be viewed as the same since they both represent inactive users.

To shed more light on the group of private users, the profile information gathered on 12-01-08 for each private user seen during sampling was examined. The LastUpdated entry on a user's profile provides the last time this user posted to their blog regardless of the user's privacy settings. LiveJournal is an open system in the sense that new users join and leave every day. Joining the network is a much more concrete event than when a user leaves the network, unless the user actually cancels their account. An examination of the LastUpdated profile entries show that a large percentage of the users originally classified as private, because a blog post was never witnessed in the Atom stream, are actually inactive users. Over 50% of the private users contained in the network model on 12-01-08 have not posted to their blogs in over 1 year.

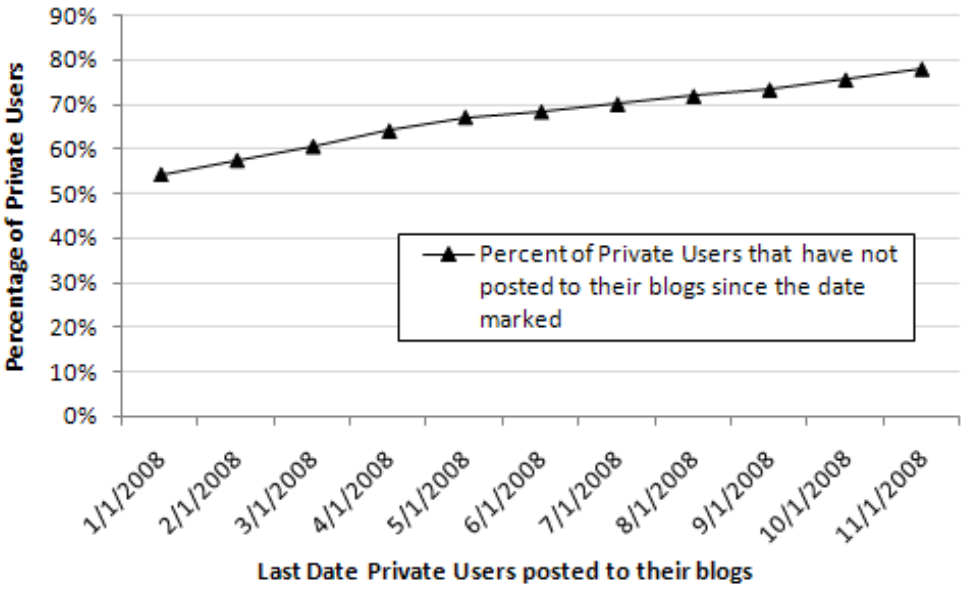

Fig. 6. LastUpdate Values for Private Users

Fig. 6 provides the results of analyzing the LastUpdated entry for profile data collected on 12-01-08. The removing of these inactive users from the network model is viewed as reasonable since they are not contributing to the activity of the network. The difficulty comes in how to draw a line to determine which users are kept and which are removed. Different cutoff points were explored when removing inactive users and the effect removal had on the structural relationship between both public and private users. Figure 7 provides the result of this analysis. As the cutoff date is increased (i.e. a private user must have made a more recent post to remain included in the final network model) the mass of network links shifts in favor of links connecting only the public users. The date 05-01-08 was chosen as a cutoff date in analysis meaning that users must have posted to their blog within a window of three months before construction of the network model began on 08-01-08. Fig. 7 shows the effect on the network link mass by choosing

different cutoff dates. If the goal is to maintain a network model that followed active users overtime then the cutoff window would be applied repeatedly at some interval as monitoring of the network advanced. This sliding window would have the effect of shaving off inactive users from the network model. Fig. 8 shows the effect of the three month cutoff window on our network model at 12-01-08 in terms of the network link mass connecting public and private users.

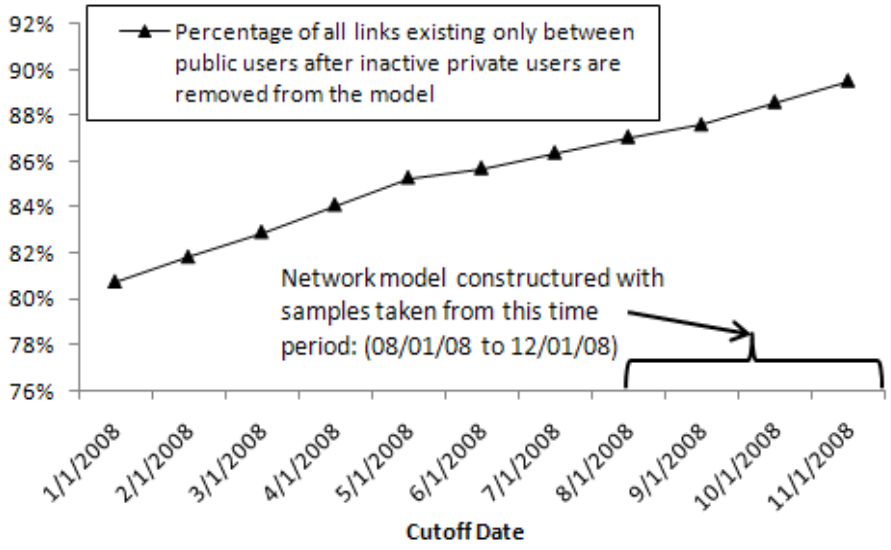

Fig. 7. Application of Cutoff Dates

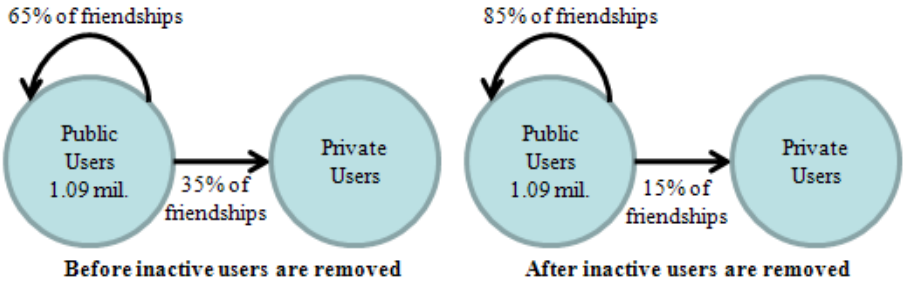

Fig. 8. Network Linkage Mass after Removing Inactive Users

III.5.3 Verification of the Network Model

To verify the correctness of public user friendship links in the final network model generated from sampling data taken from 08-01-08 to 12-01-08, a complete crawl of the LiveJournal network for all 1.09 million unique public users seen during this period was performed in a staggered fashion between 12-01-08 and 12-06-08. At 12-01-08 the network model contained four months of sampling data provided by EDS. Based on the removal of inactive private users discussed in the previous subsection, public users comprise 85% of the total network linkage mass. The verification crawl that was performed started on 12-01-08 and ended on 12-06-08. A staggered crawl was needed due to the large volume of friendship link data required for verification and the need to respect hit rate limits. During this staggered crawl updates to the network model were made from new samples generated through EDS. This was necessary in case a link change was detected through EDS before a user's friendship data could be gathered from LiveJournal through the verification crawl. Each public user's friendships in the network model were compared to the user's friendship link data obtained in the verification crawl. We take the ratio of the link events that match to the total number of link events examined. The correctness value computed was 98.05% with 31693329 link events matching out of 32320601 link events examined. Again, a link event is either an addition or a removal. Matching of the links in the final network model to those gathered in the verification crawl took into account both additions and removals.

III.5.4 Structural Analysis of the Network Model

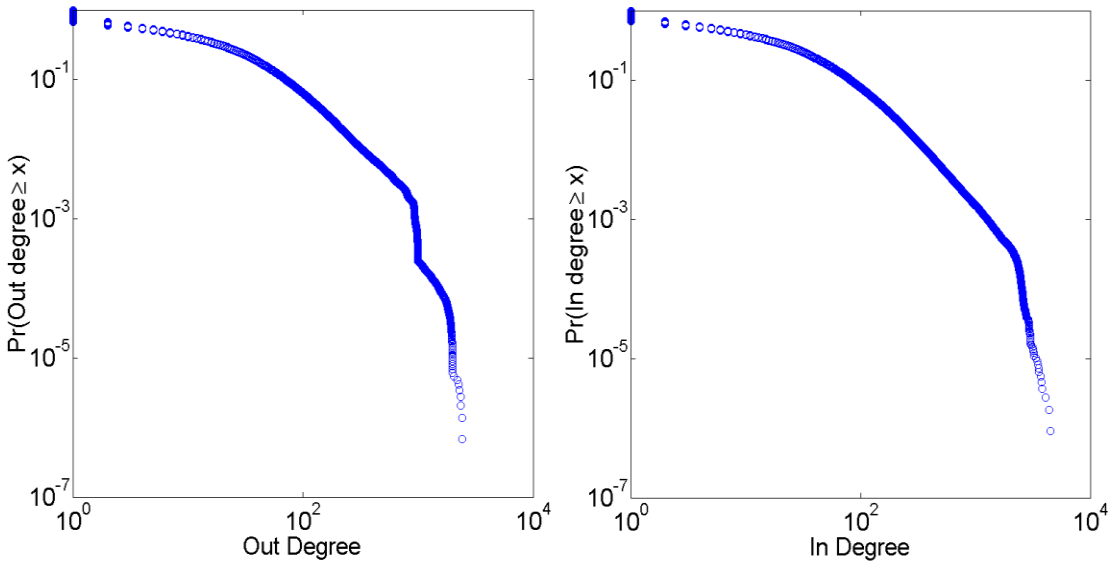

Fig. 9. Log-log Plot of In-Degree and Out-Degree

The structural analysis provided in this subsection applies to the network model at 12-01-08 that was generated from the four months of data captured from 08-01-08 to 12-01-08 using Event-Driven Sampling. Analysis is applied only to the 85% of the network mass of links connecting public users. The analysis shown in Figs. 9 through 12 show that at the point of verification, 12-01-08, that the network model has both the scale-free and small-world properties one would expect for a large-scale on-line social network such as LiveJournal (Barabasi et al., 2000; Li et al., 2006; Mislove et al., 2007; Watts and Strogatz, 1998). Fig. 9 provides the degree distributions for both outward and inward links of public users. The maximum likelihood method was used to calculate the best power-law fit for the degree distributions for both the In-Degree and Out-Degree public user friendship links. The Kolmogorov-Smirnov fitness test, which computes a D-

value for the estimated power-law fit to the data, was used to determine how well the two distributions fit the power-law degree distributions. Typically the scaling coefficient lies in the range $2 < \alpha < 3$ for real world data (Clauset et al., 2009). The value of 2.65 was obtained for a (scaling coefficient) with Kolmogorov-Smirnov fitness score of 0.0267 for Out-Degree and an α coefficient and a fitness score of 2.89 and 0.0168 respectively for In Degree.

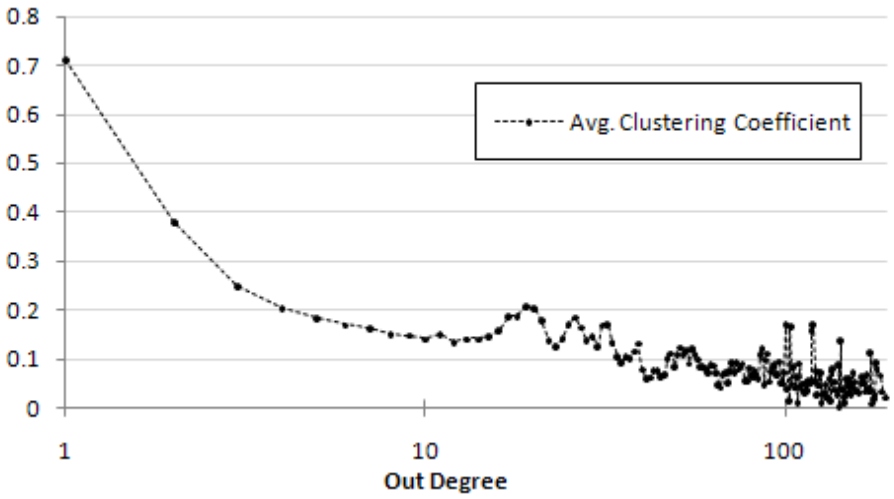

Fig. 10. Average Clustering Coefficient vs. Out-Degree

Fig. 10 shows a high average clustering coefficient exhibited by the network model that decreases as the degree of users increases. Fig. 11 shows the low average path distance between public users. Figs. 10 and 11 both indicate that the network model exhibits small-world structure having both high average clustering coefficients in conjunction with low average path length (Watts and Strogatz, 1998). Figs. 10 and 11 provide values that are similar to the values obtained in (Mislove et al., 2007) which performed a complete crawl of the LiveJournal network.

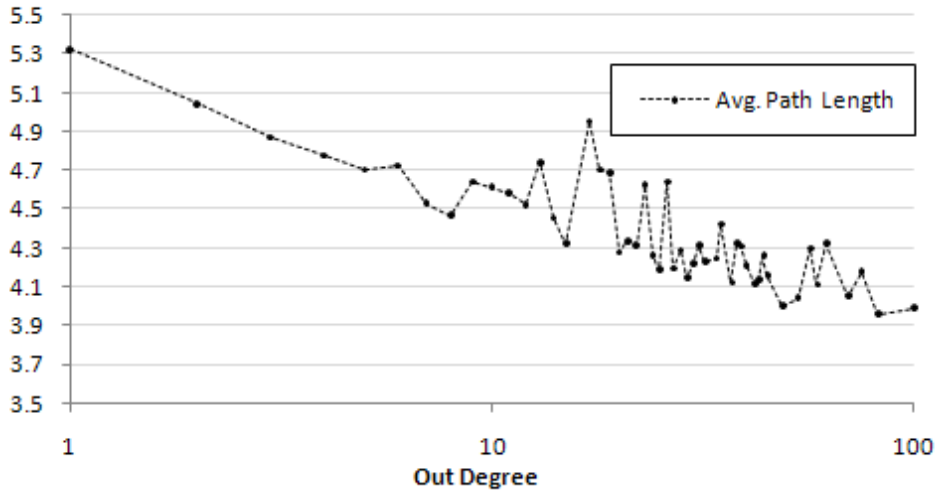

Fig. 11. Average Path Distance vs. Out-Degree

III.6 Conclusions

This chapter presented an empirical analysis of an Event-Driven Sampling approach applied to the Atom feed of the LiveJournal on-line social network. The EDS approach was shown capable of capturing all of the public user blogs posts while maintaining 98% daily correctness for the friendship link dynamics of these users. Also shown was that the group of public users comprised 85% of the network linkage mass that is active. The resulting network model constructed from the network sample data was shown to maintain small-world and scale-free properties. The EDS approach applied to the LiveJournal site allowed for the construction of a corpus of blog posts with friendship link dynamics that is the most complete seen to date. This provides a dataset capable of supporting analysis of relationships between the natural language within blogs and the link dynamics exhibited between users over long periods of time within

the context of an on-line social network. This analysis makes up a majority of the research presented in Chapters IV and V. The length of time of analysis is limited only by the duration of time EDS is applied to the LiveJournal network.

The EDS approach could possibly benefit from closer examination of the public users responsible for errors in the link dynamics that account for the daily 2% inaccuracy in dynamic link data. Presently all public users are sampled based on whether or not they appeared in the Atom feed for a given day. While this simple method of determining the sampling rate for each user provided good performance, other more complex methods are likely possible that could help reduce the 2% daily error rate. For researchers interested in the link dynamics of private users the EDS approach could be expanded to capture their dynamics by further studying the relationship between public and private users and potentially driving the sampling of private users based on public user activity, however this would not provide access to the blogs of private users. The network data for private users was not an immediate concern to the research in this dissertation since one of the primary areas of interest was in the construction of a corpus of blog posts with friendship link dynamics. Private users are not of interest since one cannot access the content of private user blogs.

One possible criticism of the EDS approach is that it can only be applied to the LiveJournal site. The primary goal in applying EDS to LiveJournal was to determine the possibility of constructing a corpus of blog posts with friendship link dynamics allowing the study of relationships between the natural language contained in user blog posts and the link dynamics experienced both locally by users and by the network as a whole. The

goal was not to address the general problem of sampling the link dynamics for all open large-scale online social networks. But, as mentioned previously, the new and widely popular micro-blogging social network site known as Twitter announced during the time this research was ongoing that they planned on providing a feed similar to LiveJournal's (Twitter, 2009). The work on EDS also demonstrated that an event-based approach to sampling a large-scale online social network can very efficiently capture the active portion of the network. This is something that has recently gained more relevance as social network service providers and users compete for customer data (Weintraub, 2010).

CHAPTER IV

EXPERIMENTS WITH TOPOLOGY-BASED LINK PREDICTION [*]

IV.1 Introduction

In this chapter experiments are described examining the practicality of applying link prediction using strictly topological metrics to an open large-scale online social network. The approach is practical in the sense that researchers or other interested parties outside of a social network service provider who do not have direct access to social network data could acquire data and study the link prediction problem. Only topological metrics are discussed making use of one previously identified metric as a baseline and introducing one new metric. The open nature of the network is directly addressed through a study of the linking dynamics over time between users and the effect the openness of the network (i.e. users entering and leaving the network) has on the prediction of new friendship links, something which had not yet received attention in the literature at the time of the ongoing research for this dissertation. The data collection and study of the LiveJournal social network in this chapter focuses on a period of approximately 2 years. Groups of users are followed from the time they first join the network out to approximately 10 months after joining. The effect of applying link prediction at different points in time during the users' temporal progression in the network is examined. Analysis shows that prediction results are best when the link

[*] Significant portions of the material in this chapter appear in Corlette, D., Shipman, F., 2010. Link prediction applied to an open large-scale online social network. Proceedings of the 21st ACM Conference on Hypertext and Hypermedia, Toronto, Ontario, Canada, 135-140. © 2010 ACM, Inc. http://dx.doi.org/10.1145/1810617.1810641 Reprinted by permission.

prediction algorithm is applied soon after users have entered the network and that link prediction results for precision and recall diminish the longer users have been members of the network. A larger training window leads to better recall of predicted links but potentially reduced precision. To date, to the best of the author's knowledge, the analysis presented here is the most comprehensive in terms of showing how practical the application of link prediction is in an open large-scale online social network such as LiveJournal.

A main concern for this chapter is to determine if it is possible to predict future friendship links between users of a large-scale social network and, if so, to what degree. Additionally there is an interest in understanding how the length of time a user has been a member of the network impacts the ability to predict future friendships. Understanding the ability of service providers and other interested parties to predict future friendship links for users is important as increasing numbers of the public continue to import their off-line social networks into online social networks. Applications of link prediction range from simply recommending future friends, as found in some social networks, to potentially forecasting large scale social trends.

During exploration of the link prediction problem a number of classifier approaches were examined. Ultimately it was determined that a Naïve Bayes classifier performed best for the considered metrics when applied to the LiveJournal data. A previously identified metric, sometimes referred to as the "Adamic-Adar" metric (Adamic and Adar, 2005), has been shown to be one of the strongest metrics for link prediction in (Liben-Nowell and Kleinberg, 2003) which examined the application of a

large set of topological metrics to the arXiv citation network. This metric was used as a baseline to compare new potential metrics. During the course of research a second metric was identified which is referred through the dissertation as the Restricted Clustering Coefficient, abbreviated as CC, which showed a strong capability of increasing recall scores.

The study of the link prediction problem presented here differs from past studies since the open nature of the network (i.e. users entering and exiting the network) is not neglected. Past approaches to the link prediction problem relied on multiple static snap shots over extended periods of time in a similar fashion to the approach taken in this chapter. However the distance between static snapshots often spanned months or years and the length of time users were members of the network was not taken into account when choosing a set of users from the network for link prediction. (Liben-Nowell and Kleinberg, 2003; Murata and Moriyasu, 2007; Scripts, Tan, Chen et al., 2009) In this chapter, friendship link dynamics for groups of new users as they first join the network are examined over a period of 10 months. The application of link prediction is applied at different points in time beginning very shortly after a user first joins. The effect of different training data sizes on precision and recall metrics are examined and issues related to the interpretation of recall in the context of link prediction in an open large-scale online social network are discussed.

Gaining access to the kind of topological and temporal data required for performing the type of comprehensive analysis needed to support the study of link prediction on these types of networks has been prohibitive in the past except for select

researchers (Kumar et al., 2006; Leskovec et al., 2008). Chapter III outlined a technique for easily accessing both the topological and temporal content data of the LiveJournal social network using an Event-Driven Sampling approach. This approach allowed for 85% of the network mass for active users to be obtained while maintaining 98.5% daily correctness across all of the friendship links. The EDS approach is the technique used to acquire the data used in the analysis presented in this chapter.

This chapter is organized in the following manner. In 4.2 a concise definition of the approach to link prediction is provided. Section IV.3 provides details of the LiveJournal dataset, the Event-Driven Sampling (EDS) approach used to capture the data and details regarding the groups of users chosen for link prediction analysis. Section IV.4 provides analysis of the network dynamics that motivated the link prediction approach. In Section IV.5, link prediction implementation details and results from the experiments are provided. Section IV.6 ends the chapter with providing learned insights related to the link prediction problem along with general conclusions.

IV.2 Link Prediction Problem Definition

The LiveJournal network is modeled abstractly as a graph structure $G = < V, E >$ with the set V of graph vertices representing individual users from the LiveJournal network. Users can be connected by an edge $e = < u, v > \in E$ representing an explicit directional friendship link pointing from user u to user v. The LiveJournal dataset is partitioned into daily snapshots where each daily snapshot represents the topology of friendship links between all users for that day. A single snapshot of the LiveJournal

network is represented as G_d with d corresponding to the day the snapshot of the network was generated (i.e. d = 03-01-2007). The delta between all snapshots is 24 hours. U is the set of all users in our dataset and u_d is the set of users who joined the network on day d where $u_d \subseteq U$. A training set is calculated for a set of users u_d using n days of graph snapshots. A training set is represented as $Tr_{d1,d1+n}(u_d)$ which is read as the training set for u_d for the period $d1$ to $d1 + n$. For each of the graph snapshots in the training set metrics are calculated between users. Metrics calculated between two users who are currently friends form positive training examples. Metrics calculated between users who are not currently friends form the negative training examples. The source of potential new friends for a user i is restricted to the neighborhood that is 2 hops away. The neighborhood 2 hops away for user i is represented as $k2_i = \{ v_k \mid e_{ij} \in E \wedge e_{jk} \in E \wedge e_{ik} \notin E \wedge i \neq j \wedge j \neq k \wedge i \neq k \} where\ v_k \in V$. This can be read as the "friends of a user's current friends that are not currently friends of that user ". Negative samples are drawn from the $k2$ neighborhood by sampling. The number of negative samples taken to form the negative training instances is equal to the number of positive training instances in the current training set for a given day. From this training set a classifier C is constructed which is applied to the metrics values calculated on the graph snapshot G_{d1+n+1} to predict the future friendships for the set of users in u_d. Effects related to varying the number of daily snapshots that comprise the training dataset were explored. Also examined were the effects of predicting future friendships for new users, m days after the users join the network. The study focused on examining the sets of users where d equals 03-01-07, 03-02-07, and 03-03-07 representing the days

these sets of users joined the LiveJournal network. The values of n explored range from 15 days to 3 months, meaning that the amount of training data used to train the classifier ranges from 15 days to 3 months. For values of m, different values were examined over a period of 10 months.

IV.2.1 Properties of the LJ Data Set

The dataset used in the study was collected from the LiveJournal social networking site using an Event-Driven Sampling (EDS) methodology described in Chapter III. Briefly, the main idea behind the EDS approach to sampling an on-line social network involves looking for user generated events that indicate the user is actively using the network and thus has potentially made a modification to their social hyperlinks (friendship links). A user's friendship network is sampled in proportion to the rate of the witnessed events. The event used in the LiveJournal network to determine when a user's list of friends should be sampled is the appearance of a blog post in the LiveJournal Atom feed generated by the user. This Atom feed was monitored continuously throughout the sampling process. The Atom feed is an always-on near real-time continuous stream of all recent public blog posts. As users update their blogs, the text of the new blog post is added to the Atom feed. Sampling of user friendship links is done daily with a user's friends list being sampled if that user was seen in the LiveJournal Atom feed during the prior 24 hour period. For each user, LiveJournal provides a friends list in the FOAF format which contains both inward and outward links for the user. Sampling the LiveJournal network with an EDS approach allows us to the track the network's topological dynamics over extended periods of time while at the

same time not violating the hit rate limits of LiveJournal (5 hits per sec). The EDS approach leads to a dataset containing 85% of the "active network link mass" (edges) of the giant component and maintained approximately 98.5% daily correctness across these links. These two values were obtained by performing a complete verification crawl after sampling the LiveJournal network using EDS for 4 months and using a threshold of 5 months of user post inactivity to estimate when users had left the network. The term "active network mass" refers to the friendship links of the active users. These are users who have posted to their blogs within the previous 5 months before sampling began. The 15% of the active network mass not captured during belonged to users who did not allow their blog posts to appear in the LiveJournal Atom feed. Work outlined in Chapter III examined the effect of different inactivity thresholds when estimating that a user had left the LiveJournal network. Determination of user activity was made retroactively using user profile information taken from the LiveJournal site in the form of a "Creation Date" and a "Last Updated Date" timestamp obtained for each user in our sample. It is possible for private users to appear in the sample if a public user has a private user as one of their friends. These two timestamps provide specific dates on when the user created their account and the last time a blog entry was posted to the account. These two timestamps are available for all users of the LiveJoural site, both public and private.

Link prediction analysis presented here spans a 10 month period of activity in the LiveJournal network for the period between 03-01-2007 to 12-31-2007. Capturing of link dynamics from the LiveJournal network using EDS began on 10-01-2006. Table 1 provides numbers regarding the size of our LiveJournal network sample at 03-01-07

where link prediction analysis began and Table 2 shows the size of the network at 12-31-07 where the analysis ended. The number of removes in Table 1 is the number of removes experienced since sampling began on 10-01-2006.

Table 1
Size of Network Sample at 03-01-07

complete set of public users	
# of users	1071449
# of friends	33826402
# of active friends	31949231
# of removed friends	1877171

Table 2
Size of Network Sample at 12-31-07

complete set of public users	
# of users	2009517
# of friends	54696147
# of active friends	47673090
# of removed friends	7023057

Because link prediction is being applied to sets of new users as they join the network, how the EDS approach impacts the correctness of user friendship links for these new users as they enter the network needed to be examined. To address this issue the correctness of a set of user friendship links were examined n days after new users joined the network over a 30-day period without regard to their posting activity.

Approximately 3000 new users joined daily during the 30 day period of analysis. This provided approximately 90,000 unique users to examine during the 30 day period study of the impact of the EDS approach on new users entering the network. A single verification crawl of the 90,000 users was performed at the end of the 30 day period to produce the results shown in Figs. 12 and 13.

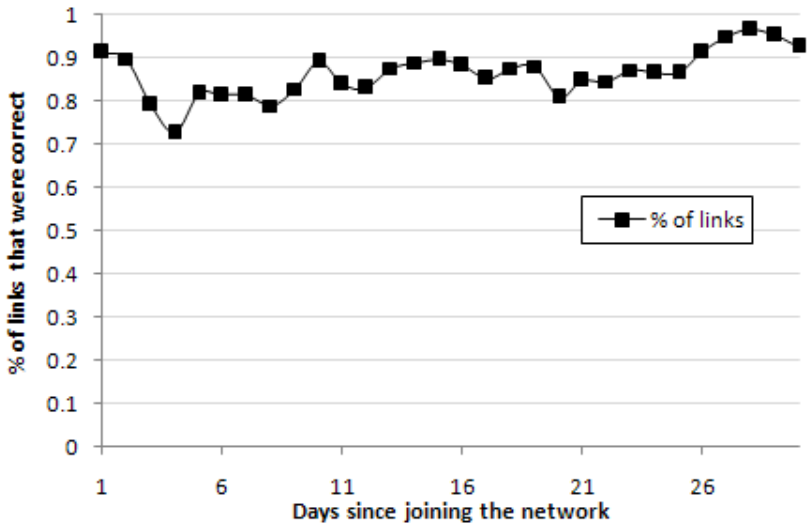

Fig. 12. % of Links Correct n-Days after Joining the Network

The x-axis of both Figs. 12 and 13 provide the number of days each unique group of approximately 3000 users have been members of the LiveJournal network. In Fig. 12, the y-axis shows the percentage of link events there were correct for each group, this applies to both addition and removal events. The average over the 30 days was 86% meaning that on average 86% of the linking events were correct for new user linkage data captured using EDS over the 30 day period as they first became active in the network. This number is lower than the 98.5% verification shown in Chapter III applied

to the entire network graph, but still very reasonable. Fig. 13 shows the percentage of outward bound links from these new users that point to other public active users. On average, 93% of the links for these new users point to other public active users for friendship link information is available. The remaining approximately seven percent of links point to private users not seen in the LiveJournal Atom feed.

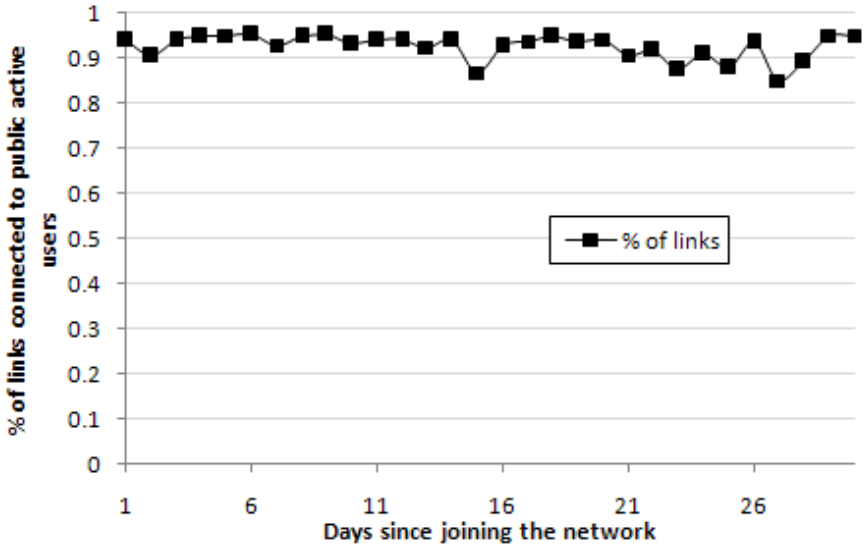

Fig. 13. % of New Links Pointing to Public Active Users

Not having dynamic link data for private users is a reality for researchers that do not have direct access to the backend storage of social network service providers. Additionally, researchers interested in studying aspects of on-line social network dynamics are limited by access rate limits imposed by the service providers when developing a crawling strategy (the number of times per second you can access the site). The application of the EDS approach to the LiveJournal site provided us with a dataset

that was close to complete for the active portion of the giant component and in a manner that did not violate access rate limits set by LiveJournal. It will be demonstrated in the sections that follow that the quality of the dataset is good enough to provide insight into the friendship link dynamics exhibited between users. In addition it will be shown that the dataset is capable of serving as basis for the study of the link prediction problem. Currently the author knows of no other way to access on-line social network link dynamics over extended periods of time for an open large-scale (millions of active uses) on-line social network such as LiveJournal.

IV.2.2 Properties of the Study Group

For the link prediction study that follows all of the users from the captured dataset that joined the LiveJounal network on the three separate days of 03-01-07, 03-02-07, and 03-03-07 were selected. The examination of the dynamics of these groups begins from the first day they join the network. No special significance was attached to these dates. Each group contains approximately 3000 users. The analysis examines these three groups over a 10 month period from 03-01-2007 to 12-31-2007. Throughout the remainder of this chapter these three groups will be referred to collectively as "the study group". A majority of the figures and analysis that follows combine results obtained from the three groups making up the study group mentioned above unless explicitly stated otherwise. Fig. 14 shows the blog posting activity pattern for the study group from the point the group enters the LiveJournal network out to approximately 2 years. These data points were derived from user profile information obtained on 12-01-08; approximately 2 years after the study group first joined the network. The "Created Date"

and the "Last Updated Date" values from user profiles were used in the creation of Fig. 14. Approximately 30% of the users of the study group never returned to use their accounts after they first create them and approximately 16% of the users were still active after 611 days. Table 3 provides the number of users in each group of users of our study group and the final count of friendship links for each of the three sets at the end of the analysis on 12-31-07.

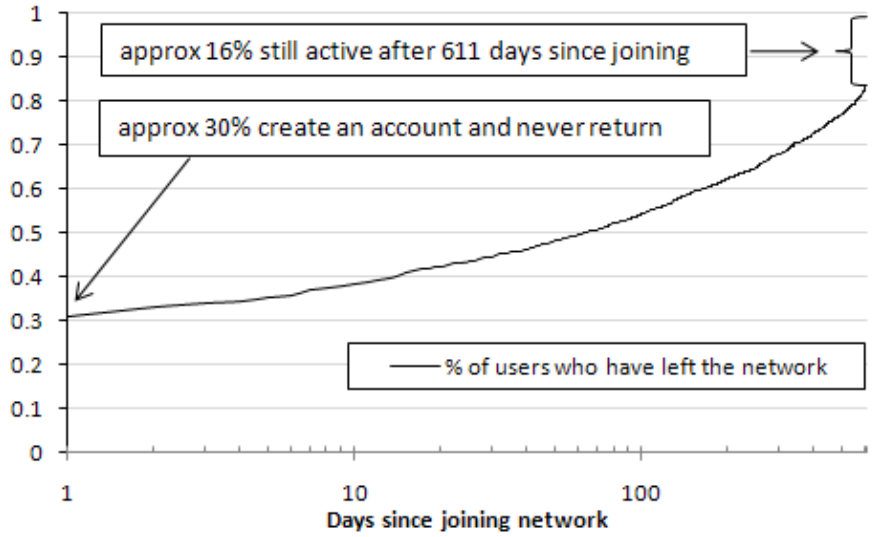

Fig. 14. Activity Pattern for Sampled Users

Table 3
Size of Groups Analyzed

group	# of users	# of links at 12-31-07
3/1/2007	3222	25011
3/2/2007	3251	23490
3/3/2007	3127	21851

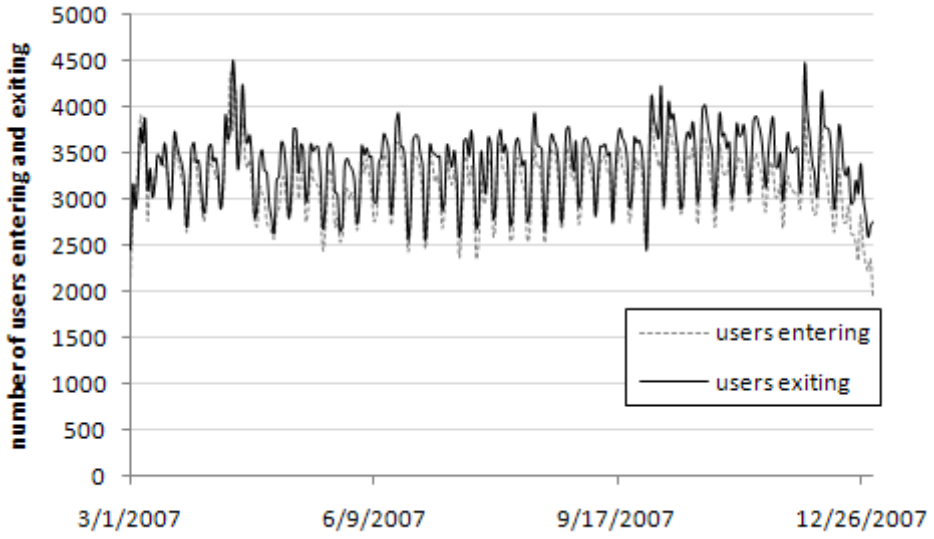

Fig. 15. Users Entering and Exiting the Network

Fig. 15 was generated using the user profile information collected on 12-01-08. The figure shows the open nature of the LiveJournal network showing users joining and providing and estimate on the number of users leaving. The dark solid line represents the users who have stopped posting to their blogs and thus potentially having left the network while the dashed line represents new users joining the network. The dark solid line of Fig. 15 was constructed from the LastUpdated value taken from the profile data. It provides an estimate of when a user might have left the network since they have stopped contributing new posts to their blogs. This value is available for all users both public and private. Since the profile data used to construct Fig. 15 was collected on 12-01-08, a point on the dark solid line at 12-01-07 in the figure represents a user who has not posted to their blog in one year. It is possible a user who does not continue to post may still log into the network and read the blogs of other users, however this is seen as

less likely with the longer the amount of time that has passed since the user's last post. Fig. 15 represents what is meant by an "open" network.

IV.3 Friendship Linking Dynamics

In this section the linking dynamics exhibited by the study group from the day they first join the network are examined out to approximately 10 months. This involves examining the source of new friends for each user in the study group over the 10 month period of analysis. It is shown that almost one third of all the new links ever established during the 10 month period are done so during the first 10 days of activity and that a majority of these new links during this period come from users that are 6 or more hops away or are disconnected. After the first 10 days this dynamic quickly changes with a majority of the new friendship links coming from users that are 2 hops away.

The link prediction efforts were focused on predicting new friendships between users in the study group and their $k2$ neighborhood. A user from the $k2$ neighborhood from the perspective of user A is simply another user in the network model for which A is not directly connected and resides a link distance of 2 from the user. The term "$k2$ neighborhood" was defined precisely in Section IV.2. For ease of reference in the remaining parts of this section this group is referred to as being 1 hop away from the current friends of the user under consideration for link prediction. Likewise, the other potential friends for the user under consideration will be referred to as being 2, 3, 4 or 5 hops away. These hop values represent the minimum distance between any member of the set of friends of the user in which link prediction is being performed and the

potential new friend. Potential friends that are more than 5 hops away are considered disconnected. Distances beyond 5 hops were not calculated. Therefore, potential new friends coming from more than 5 hops away are included in the 5 hop set. The purpose of this section is to examine the rational for the choice to focus on the $k2$ neighborhood as a source of potential new friends for the study group while ignoring the other groups. To better understand this choice examination of the source of new friends for the study group over the entire 10 month period of analysis is necessary.

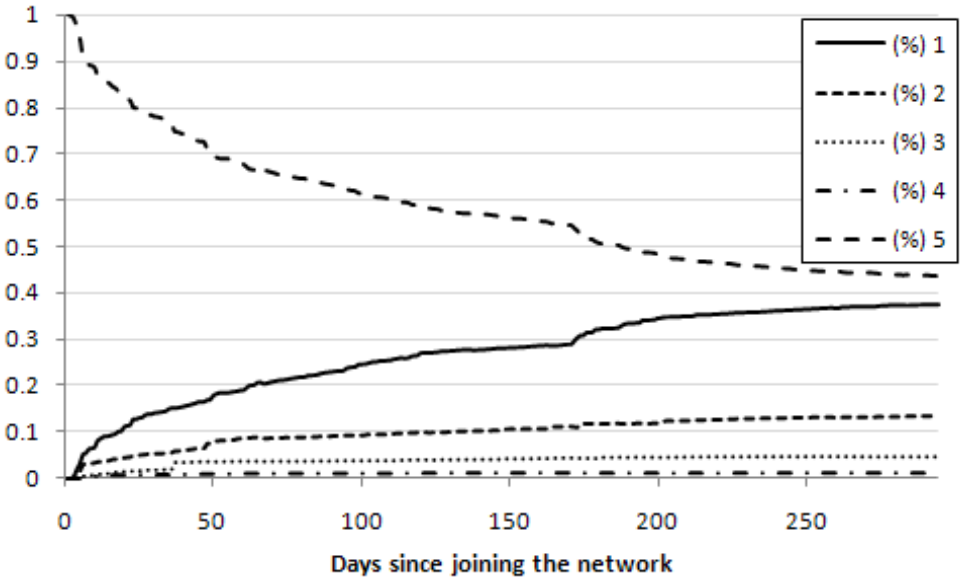

Fig. 16. Source of New Friends

Fig. 16 shows that the two main sources of new friends for the study group come from the 1 and 5 or more hops away from a user's current friend set. Each entry on the x-axis shows the number of days since the users in the study group have joined the network. The y-axis shows the percentage of new friends coming from each source up to

and including the day shown on the x-axis. For example, at point 100 on the x-axis (100 days after the users joined the network) approximately 23% of all new friendships made to this point have come from the 1 hop away and approximately 61% of all new friendships have come from the 5 or more hops away with the remaining 16% having come from the other hop distances. A majority of the new friends for the study group over the 10 month period of analysis come from users who are 5 or more hops away from users' current friends or potentially disconnected. Of course, on day 0, the moment immediately before a user enters the network, the user is completely disconnected from all other users of the network. Fig. 16 shows that as time progresses, the increase in the number of new friends coming from 1 hop away is proportional to the decrease in the number of friends coming from 5 hops away. The figure shows that the source of new friends changes over time. To further understand how the linking dynamics change over time, a more detailed examining of what happens during the first 10 days of after the users in study group join the network was performed.

Fig. 17 provides a breakdown of the percentage of all new friendships made over the 10 month period of analysis in relation to the number of days since the users joined the network. The x-axis is in log scale to emphasize what happens during the first 10 days. Each point on the y-axis corresponds to the percentage of all new links that occurred over the entire 10 month period of analysis that were made on that day. For example, at point 2 on the x-axis, two days after the study group joined the network, close to 15% of all new links made during the 10 month period by users in the study group occurred on the second day. Summing up the values from 1 to 10 along the x-axis

arrives at an approximation that 33% of all of the new friendship links ever made by the study group over the 10 months period were made during the first 10 days.

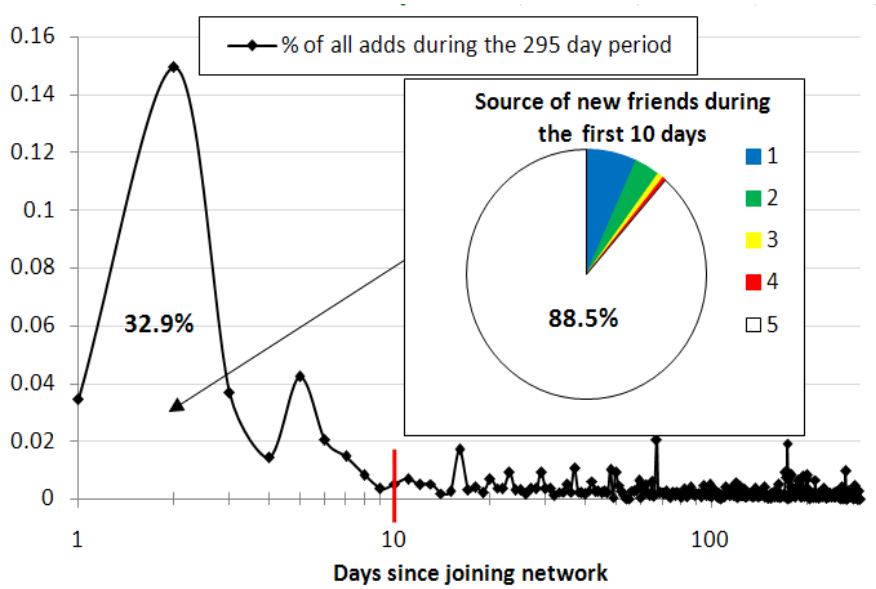

Fig. 17. Addition Link Dynamics during First 10 Days

The figure also provides a percentage brake down of the source of all of the new friendships (i.e. hop distance) made during the 10 day period in the chart labeled "Source of new friends during the first 10 days". Again, for clarity, the hop distance is the smallest distance between the potential friend and a user's current friends. The chart shows that approximately 89% of the new links made during the first 10 days are made to users residing 5 or more hops away, potentially disconnected. The first 10 days seem to reflect a structurally chaotic period where the new users are entering the network and forming their first friendships. Methods of link prediction that rely strictly on topological metrics would perform very poorly during the first 10 days or so with close to 90% of

new friends coming from the 5 hops away or potentially being disconnected. The space of possible new friends on average grows exponentially with each successive hop distance from the user under consideration. This observation is related to the low average path distance between all users.

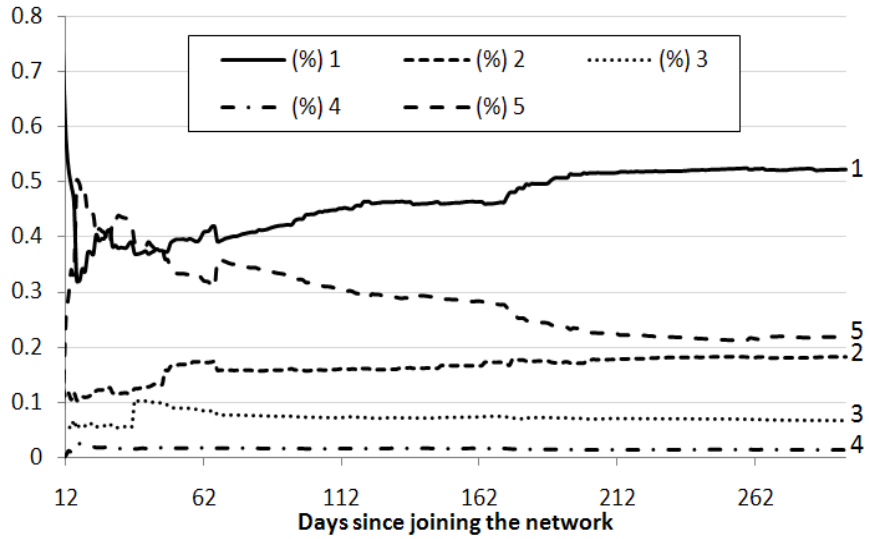

Fig. 18. Addition Link Dynamics after First 10 Days

Past approaches to link prediction have represented the network dynamics in the form of a few static snapshots of the network's topology (Liben-Nowell and Kleinberg, 2003; Scripts, Tan, Chen et al., 2009). The Liben-Nowell study made use of the citation network arXiv.org consisting of approximately 11000 nodes and 22,000 edges. The Scripts study made use of the DBLP citation network, the TakingItGlobal.org social network website, and the WebKB dataset. The DBLP citation network consisted of approximately 6000 nodes and 30000 edges. TakingItGlobal.org consisted of approximately 1200 nodes and 6000 edges. The WebKB dataset consisted of

approximately 1000 nodes and 1600 edges. The amount of time a user under consideration for link prediction has been in the network has not been considered as a factor in the link prediction process. The results in Fig. 17 lead to the choice of focusing efforts on predicting future friendship links for users of the study group at points in time after the first 10 days as it is clear that any successful topological metrics that are used would need to be designed to handle 5 or more hops away and are not likely to be successful.

To gain a better understanding of the link dynamics beyond the initial 10 day period after users first join the network, the remaining 64 % of the new friendship link additions were examined after removing information about the first 10 days of linking dynamics. Fig. 18 shows the percentage of new friends for each hop distance for the remaining 64% of new links made after the first 10 days. New friendships from a hop distance of 1 dominate the other sources of new friends, contributing approximately 52% for this period with about 21% from the 5 hop distance, 19% from the 2 hop distance, and 7% from the 3 hop distance. This result is supported by (Leskovec et al., 2008) which found that most edges were created locally between nodes that are close. The datasets used for that study involved, FLICKR (flickr.com), DELICIOUS (del.icio.us), YAHOO! ANSWERS (answers.yahoo.com) and LINKEDIN (linkedin.com). Since almost half of all new friendships come from distance of 1 hop from the current friends of the user under consideration for link prediction, a choice was made to focus efforts on predicting future friends coming from this distance.

A reasonable concern about the results in Figs. 16, 17 and 18 is whether the characterization of the source of new friends is affected by the lack of information about active private users. To explore the effect of this lack of information a random list of 15% of the public active users seen in the dataset on 12-31-07 was generated and all friendship link information was removed from the network model. The source of new friendships was then recalculated on this reduced network model to assess the impact of the missing information. Table 4 shows the user and friendship link counts for the LiveJournal network model after removing 15% of the users and their corresponding friendship links from the users randomly selected from the network model for the date 12-31-07. Table 5 shows the counts before the random removal. Remove of the 15% of users resulted in a removal of approximately 18% of all links connecting public active users.

Table 4
Node Removal Counts

dataset after random removal at 12-31-07	
# of users	1704659
# of friends	44728053
# of active friends	38976806
# of removed friends	5751247

Table 5
Size of Network Sample at 12-31-07

complete set of public users	
# of users	2009517
# of friends	54696147
# of active friends	47673090
# of removed friends	7023057

A network model was recalculated from 10-01-06 to 12-31-07 preventing the participation of the 15% of randomly chosen public active users. The linking dynamics for the study group were then recalculated over the period 03-01-07 to 12-31-07. The intended effect was to simulate the loss of data for approximately 15% of the links from public active users to assess the impact on the analysis of the study groups linking dynamics. The 18% reduction in link mass is approximately equal to the number of links from the identified private active users discussed in Chapter III. If the missing active private user links impact the source of new friends calculations shown in Figs. 16, 17 and 18 it would be expected to see some evidence of this impact through the removal of the 18% of links from public active users. The root mean squared error (RMSE) was calculated between the two "source of new friends" calculations derived from the two network models. For new friends coming from 1 hop away the RMSE was 0.0138. For new friends coming from 5 hops away the RMSE was 0.01 with the RMSE values for the remaining distances being below 0.003.

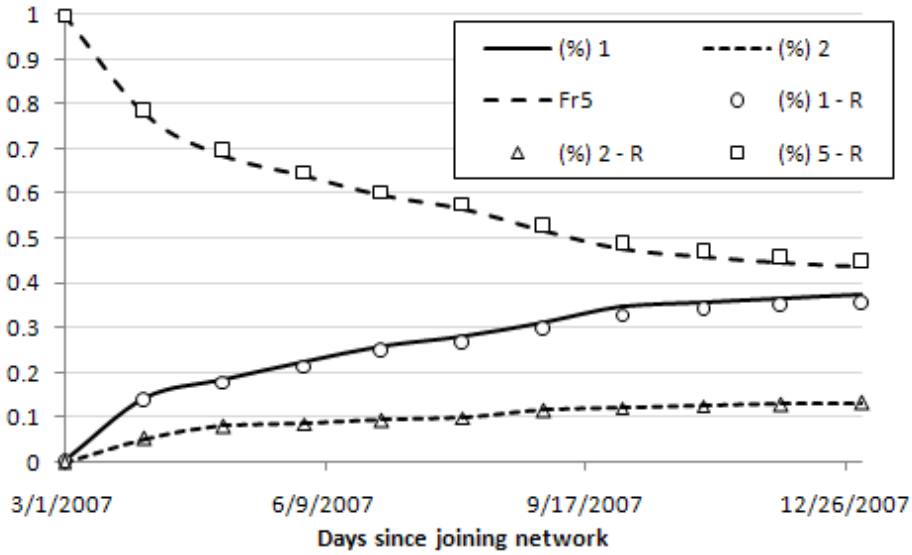

Fig. 19. Source of New Friends after Random Removal

Fig. 19 shows these results graphically by plotting the source of new friends for the study group over the 10 month period using both the complete network model and the network model after randomly removing 15% of the users and their corresponding friendship links. In the figure's legend, 1-R corresponds to the link source calculation on the dataset with randomly removed users for the hop distance one with 2-R and 5-R corresponding to hop distances two and five respectively. Fig. 19 shows a slight increase in the percentage of links coming from hop distance 5 and a slight decrease from links coming from hop distance 1 but this is negligible and even less noticeable at the earlier dates in Fig. 19. These results are not surprising as it has been pointed out before that scale-free networks are robust to random deletion of nodes (Barabasi et al., 2000). One possible criticism is that the random removal of nodes might not reflect the true topological relationship between the public and private users. But this is countered by

the evidence that public users link more often to other public users than private users. This was shown in Fig. 14 which showed over 90% on average of the new links for each set of users in the study group connecting to other public active users even though the active private users comprise approximately 15% of the giant component of our network sample. This hints at some isolation between public and private users.

IV.4 Link Prediction Experiments

In this section the approach to link prediction applied to the network model derived from EDS to the LiveJournal social network is described. The approach is motivated by the link dynamics observed for the study group and presented in Section IV.3. The two primary findings that motivated the approach is that a majority of the new links during the first 10 days come from a distance of 5 hops away from a user's current friends and that after this period a majority (over 50%) of new links come from a distance of 1 hop from a user's current friends. For this reason it was seen as natural to focus efforts on predicting new friendship links coming from the 1 hop distance. This section begins by describing the two metrics used in the link prediction classifier. One of the metrics is well known and has shown to be one of the best metrics for link prediction applied to social networking related datasets (Liben-Nowell and Kleinberg, 2003). The second metric is considered new. How these metrics support the construction of a classifier for predicting new friendship links coming from a distance of 1 hop from current friends of the study group is discussed. The construction of the training and test sets is discussed along with how the construction of these sets can impact the classifiers

performance. As a part of that discussion an approach to handling the notion of recall when predicting future friendships is examined.

IV.4.1 Link Prediction Metrics

During the course of the link prediction investigation many different novel metrics were examined in a search for new previously unidentified metrics capable of supporting link prediction. Present here are the two metrics found to provide the best performance. These two metrics when used together provided the best link prediction performance for the study group. The first and most powerful metric used as part of the classifier design is the familiar Adamic/Adar metric first defined in (Adamic and Adar, 2005) and shown below as $Eq1$ for two users x and y where $\Gamma(x)$ corresponds to the list of friends of x. The Adamic/Adar metric, or AA metric, is an extension of the common neighbor metric (Murata and Moriyasu, 2007) $cn_{x,y} = |\Gamma(x) \cap \Gamma(y)|$ that gives more weight to the common neighbors with fewer friends.

$$(Eq1) \quad m1_{xy} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{log|\Gamma(z)|}$$

Fig. 20 provides a graphical representation of the AA metric to help clarify the calculation. Shown in the figure are two users x and y with mutual friends labeld $Z1$ to Zn. These mutual friends are referred to as common neighbors and referenced in $Eq1$ as $\boldsymbol{z} \in \Gamma(x) \cap \Gamma(y)$. The AA metric sums the number of friends for each common neighbor z of x and y. A weighting is applied by taking one over the log of the number

of friends of for each common neighbor z in the sum which gives more weight to common neighbors with few friends.

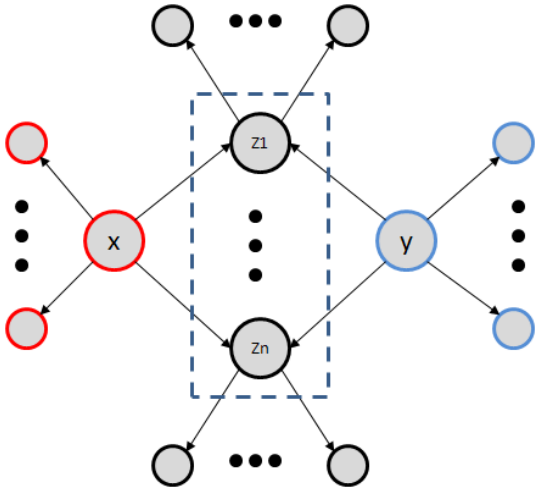

Fig. 20. Graphical View of AA Metric

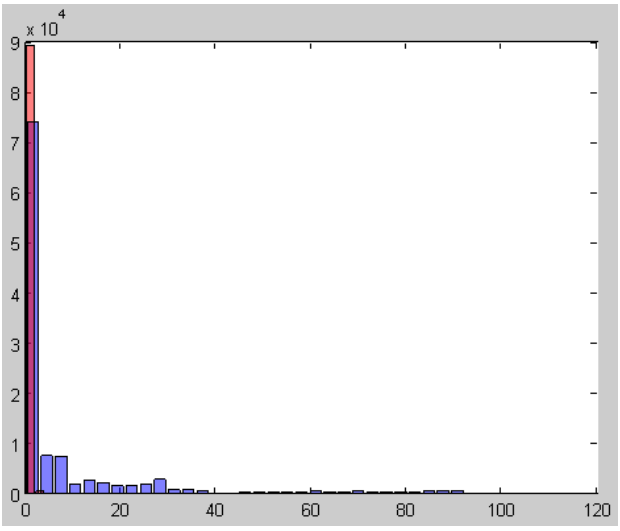

Fig. 21. Examples Values for AA Metric

Fig. 21 provides examples values of the AA metric taken from training data for the study group for the period of 04-15-07 to 04-31-07. The figure is a transparent histogram chart showing values for the AA metric for both positive (blue) and negative (red) examples of friendships. A one to one ratio between positive and negative examples was used to create the chart so the negative examples did not over power the positive examples in the histogram. Positive and negative values that populate the histogram in Fig. 21 were chosen through random sampling from a set of training instances.

The second metric, the Restricted Clustering Coefficient, abbreviated as CC, is based off the idea of the general equation for the clustering coefficient of a vertex in a graph. The clustering coefficient for a single vertex v measures the ratio of links between vertices in the local neighborhood of v to the total number of possible links that could exist. The clustering coefficient concept is applied to a restricted local neighborhood of a potential new friend. Precisely, let x be the user in the study group for which future friendships are being predicted. Let y be the potential new friend of x that is currently a member of x's $k2$ neighborhood $(k2_x)$. Let N_y be a sub set of the local neighborhood of y where $N_y = \{v_z | e_{yz} \in E \wedge v_z \notin k2_x\}$. Note that the local neighborhood of y is restricted so that it does not contain members of $k2_{ix}$. The CC metric is shown as $Eq2$ below and is simply the clustering coefficient calculated on the restricted local neighborhood N_y which is simply the ratio of edges existing between members of N_y over the total number of edges that could possibly exist in N_y.

$$(Eq2)\quad m2_{xy} = \frac{|\{e_{jk}\}|}{|N_y| \cdot (|N_y| - 1)} : \; v_j, v_k \in N_y, e_{jk} \in E$$

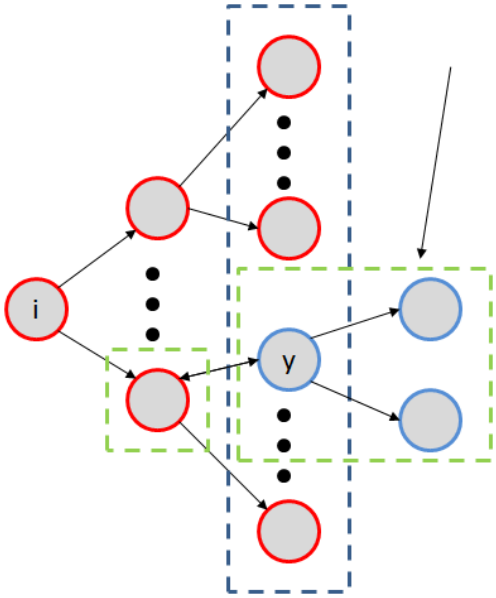

Fig. 22. Graphical View of CC Metric

Fig. 22 provides a graphical view of the calculation of the CC Metric to help clarify the calculation. The figure shows a graph configuration that would exist in a test set where y is a potential new friend of x residing in the $k2_x$ space represented by the blue dashed line. The CC metric consists of a clustering coefficient calculated on the N_y space outlined by the light green dashed line. Fig. 23 provides a transparent histogram chart of values for the CC metric calculated from training data from 04-15-07 to 04-31-07 in the same way as Fig. 21 did for the AA metric.

In Subsection IV.4.4 link prediction results from the application of two different classifiers to the network model are presented. Two different classifiers are examined.

The first classifier uses only the AA metric. The second uses both the AA metric and the CC metric.

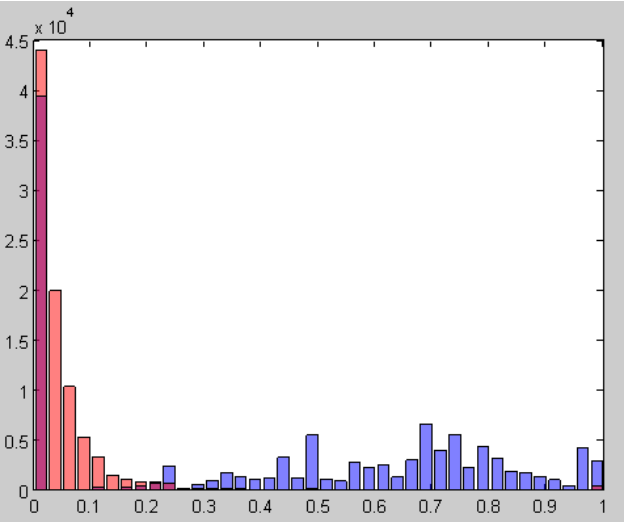

Fig. 23. Example Values for CC Metric

IV.4.2 Link Prediction Classifier

A supervised machine learning approach to link prediction was used that requires the training of a Naïve Bayes classifier applied to identification of future friendships between users in the study group and members of these users' $k2$ neighborhood. Other learning models such as decision trees and neural networks were explored, but the best results were obtained using a simple Naïve Bayes classifier with the two metrics from Subsection IV.4.1. The Weka open-source data mining and machine learning API was used for the Naïve Bayes classifier implementation. In the case of classifier 2 which uses both AA and CC metrics the input to train the classifier is simply a vector containing three values. The vector contains a true/false value indicating if the vector is an example of positive or negative example of friendship and it contains values for the two metrics.

If the training vector is a positive example then the values of the metrics in the vector were computed between a member of the study group and one of the member's current friends. If the training vector is a negative example then the values of the metrics in the vector were computed between a user of the study group and a user from its $k2$ neighborhoods. Negative examples determined by randomly sampling the $k2$ neighborhood of users in the study group in proportion to the number of positive examples in the training data.

IV.4.3 Training and Testing Data

Conceptually our training data is divided into vectors of positive examples of friendships and vectors of negative examples of friendship. For each daily graph snapshot over the 10 month period the values of the two metrics between the study group and their current friends are computed to form the positive examples of friendship. The same calculations are performed between the study group and their $k2$ neighborhood to generate our negative examples. From these calculations the training set is constructed based on the width of time the training data spans. If for instance the desire is to predict future friendships at day 04-01-07 and it is desired to use the previous two weeks of link topology to build our classifier, then all of the calculate metrics for positive examples of friendship for the prior two weeks from 03-15-07 to 03-31-07 from each graph snapshot within the range would be selected. The negative training examples are sampled from $k2$ neighborhood during the period from 03-15-07 to 03-31-07 from each graph snapshot within the range in proportion to the size of the positive examples. Together the positive and negative examples make up a single training set that is used to train the Naïve Bayes

classifier. The vectors making up the test set simply contain the metrics calculated between all members of the $k2$ neighborhood (the potential new friends) for each user in the study group for the day the prediction is being performed. The vectors in the test set are applied to the trained Naïve Bayes classifier which computes a probability for each vector either being positive, yes a future friendship will occur between the two users, or negative, a future friendship will not occur between the two users. Maximum likelihood determines the class label.

IV.4.4 Experimentation

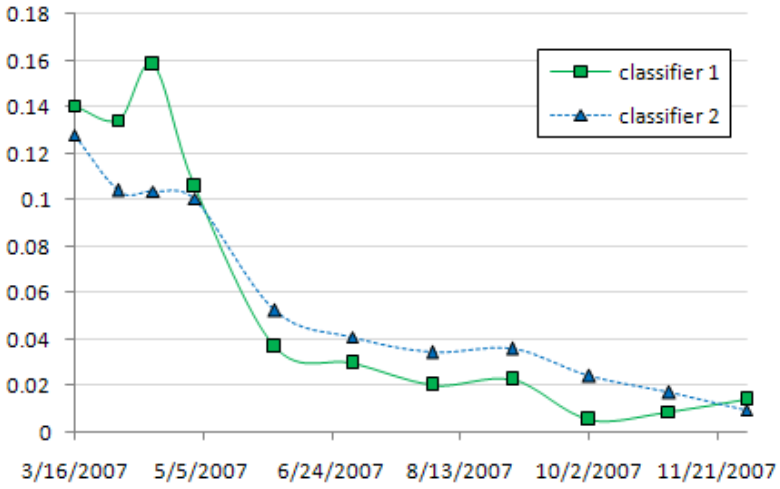

Fig. 24. Precision

Precision in the context of link prediction is much lower than it is in search activities where users are expressing a need explicitly rather than it being inferred based on their prior actions. Fig. 24 shows the precision performance using the two separate classifiers and a training window of 2 weeks. Classifier 1 uses only the AA metric.

Classifier 2 uses both the AA metric in combination with the CC metric. Precision in this context is defined to be the ratio of correctly predicted friendship links that are instantiated by the end of the 10 month period (TP-true positives) to all new friendships predicted by the classifiers. Precisely, precision is calculated as $TP/(TP + FP)$ where FP are false positives, predicted new friendships that did not materialize. This is a very conservative definition of precision in the context of predicting new friends for users. Precision scores based on having users respond to suggestions or assess the quality of link predictions would likely be higher but such evaluation is not possible in an after-the-fact analysis.

Two trends are immediately obvious. The first trend is that the precision value drops over time showing that the longer a user has been in the network the more difficult it is to predict future friendships. This may be due to users filling in the more obvious friendship links relatively early, making predictions more difficult over time. Additionally, as users add new friends the size of their $k2$ neighborhood grows (potentially exponentially) creating a larger space of potential new friends requiring the classifier to reject a larger space of users who will not eventually become friends. Classifier 1 performs best during the first month and half of predictions with the maximum precision score of 0.16. After the first month and a half the best precision results are obtained with Classifier 2, with the exception of the last data point. Each classifier provides the best precision performance at different periods of time as the study group evolves through the network. Both classifiers show their best precision performance soon after the study group first enters the network with results for both

diminishing over time. The choice of a 2 week window of training data for results in Fig. 24 was made by examining how Classifier 2 performed with different amounts of training data. Fig. 25 shows precision results for four different windows of training data.

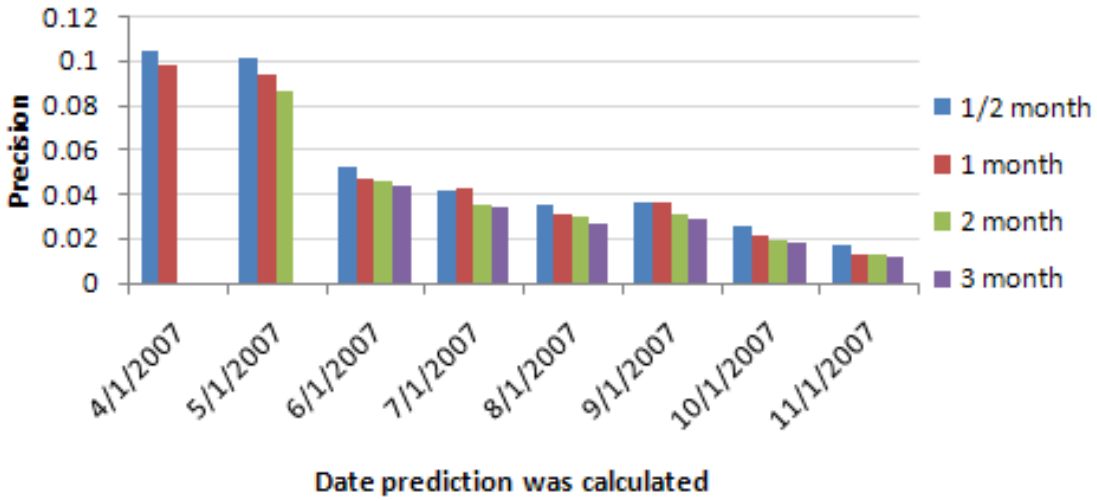

Fig. 25. Precision vs. Size of Training Data

For most dates where predictions were calculated, a window size of 2 weeks (1/2 month) of training data used to train the Naïve Bayes classifier performed best. An apparent trend is that as the size of training data increases (i.e. the training data ranges over a longer period of time) the precision value drops slightly. One interpretation is that the types of links being created vary over time so more recent activity is more predictive than older activity.

Recall is viewed as the ratio of eventually instantiated friendship links to those predicted by the classifier. Precisely, recall is calculated as $TP/(TP + FN)$ where FN are false negatives, new friendships that were not predicted by the classifier and TP are

true positives, correctly predicted friendships. To calculate a recall score there is a requirement to choose some cutoff point into the future by when one would expect the predicted friendship links to have formed thus determining the values of both TP and FN. For example, if the cutoff point is set at 1 week then only the TP and FN values from the day the predictions were calculated out to 1 week are considered. Recall in this context is a measure of the how well the classifier identifies future friends for a specific time span into the future. The examination of recall here seeks to determine how far out into the future link prediction classifiers are capable of identifying future friendships. To establish an optimal time span for the model different cut off points were examined for bounding the recall calculation where only the new links actually made by the users in the study group within the cut off period are considered. For this analysis, link predictions were calculated from 03-16-07 to 12-01-07 with multiple recall scores calculated for each day's predictions using different future cutoff values. For this analysis a training window of 2 weeks was used.

Fig. 26 was constructed by averaging the individual recall scores for each different cutoff value used for the predictions made from 03-16-07 to 12-01-7. The figure shows fairly stable recall scores for cut off values of 1 to 10 days with recall beginning to drop after 10 days. The figure shows that a classifier is best at predicting new links, from the perspective of recall, approximately 8 to 10 days into the future. Since 8 was the largest day value before the drop begins this value was chosen as the recall cutoff value used to characterize recall performance over the 10 month period of analysis for the results presented here.

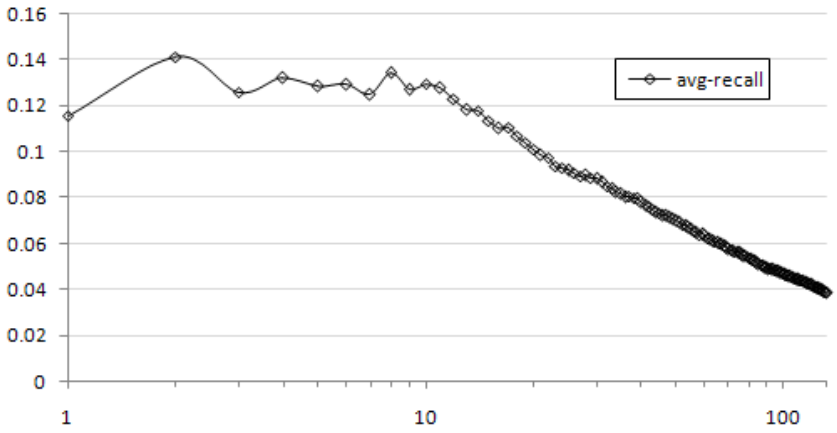

Fig. 26. Average Recall Using Different Cutoff Values

Fig. 27 provides the recall scores at each prediction point when the recall threshold is set at 8 days. Fig. 24 shows Metric 2 providing a significant increase in recall capability over the classifier that uses only the AA metric, Classifier 1.

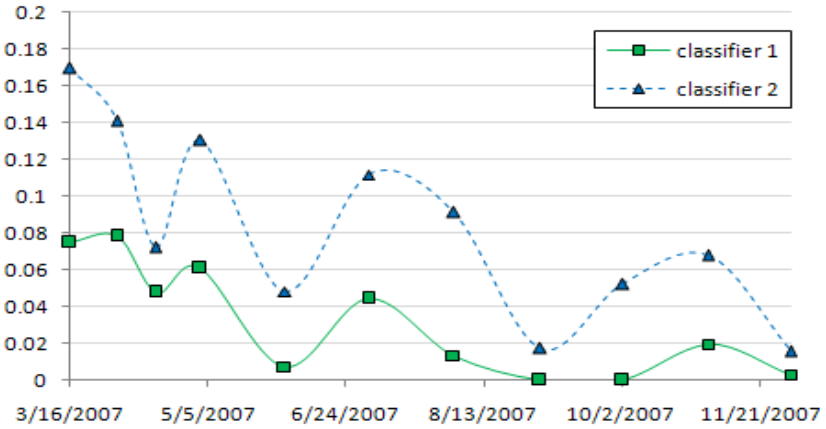

Fig. 27. Recall

Fig. 28 provides the effect of using different window sizes of training data on recall with a recall window of one week. The trend shows that more training data provides an increase in recall.

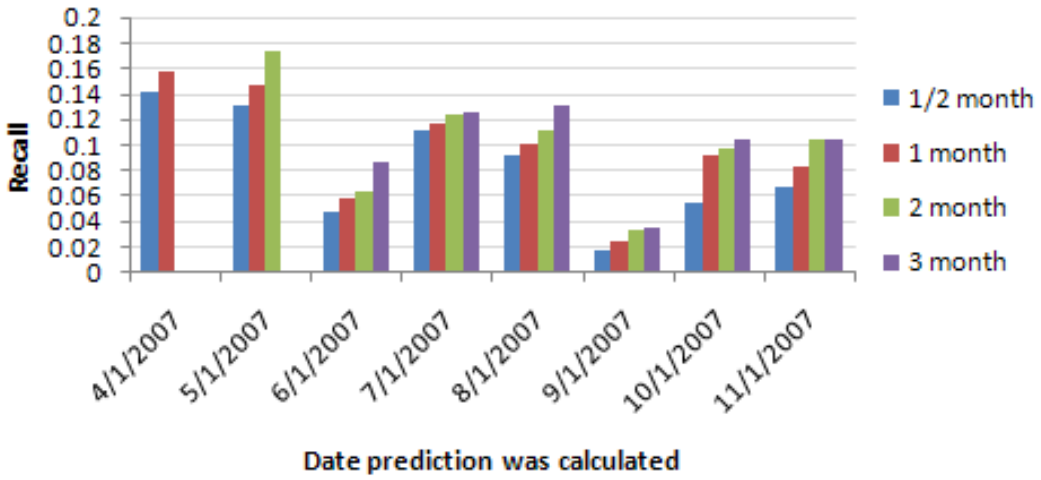

Fig. 28. Recall vs. Size of Training Data

Table 6
Counts for the First Five Test Points for Classifier 1

test date	TP	FP	TP2	FN2	precision	recall	f-measure
3/16/2007	175	1122	101	1463	0.135	0.065	0.087
4/2/2007	225	1518	58	734	0.129	0.073	0.093
4/16/2007	192	1047	36	702	0.155	0.049	0.074
5/2/2007	110	960	75	1238	0.103	0.057	0.073
6/2/2007	36	1302	7	419	0.027	0.016	0.020

Table 6 provides the true positive and false positive counts for Classifier 1 for the first five test dates for Fig. 24. Table 7 provides the true positive and false positive counts for Classifier 2 for the first five test dates for Fig. 24. True positives are the correctly predicted links that will materialize by 12-31-07, the last data of analysis. False positives are predicted links that will not materialize before 12-31-07. The column

labeled TP2 corresponds to the true positives predicted that materialized 1 week (8days) after the point of prediction. TP2 represents the predicted true positives that occurred during the limited recall window. The column FN2 corresponds to the false negatives for the same recall window of 1 week after the point of prediction. These FN2 value represents the new friendships that materialize during the 1 week recall window that the link prediction model missed. The F-measure value is computed using precision P and recall R as $(2 \cdot P \cdot R)/(P + R)$.

Table 7
Counts for the First Five Test Points for Classifier 2

test date	TP	FP	TP2	FN2	precision	recall	f-measure
3/16/2007	580	3787	268	1296	0.133	0.171	0.150
4/2/2007	558	4926	109	683	0.102	0.138	0.117
4/16/2007	473	4008	58	680	0.106	0.079	0.090
5/2/2007	393	3495	176	1137	0.101	0.134	0.115
6/2/2007	263	4768	21	405	0.052	0.049	0.051

IV.5 Conclusions

In this work the application of friendship link prediction applied to the open large-scale online social network LiveJournal was examined. The analysis and approach takes into account the open nature of the network where users are entering and exiting the network. A study group of approximately 10000 users joining over a period of 3 days was followed from the time they first entered the network out to approximately 10 months. An analysis of the link dynamics of new users entering the network for the first time revealed a chaotic period during the group's first 10 days where almost 90 percent of the new links were made to other users who were disconnected or 6 or more hops

away. Close to one third of all of the new links established over the 10 month period were made during this initial 10 day period. Shortly after the initial 10 day period the primary source of new friends quickly shifted to users that were 2 hops away eventually becoming the source of approximately 50% of all new friendships. The existence of this period implies that approaches to link prediction that strictly use topology would be greatly challenged during the first 10 days. It was found easier to predict future friendships for users the less time they had been members of the network. The ability to predict peaked at about a month or so after users in the study group joined the network suggesting the potential existence of an optimal time to predict future friendships. The notion of recall was addressed showing that the link prediction model's maximum recall scores were achieved when scoring the model with a recall window of approximately 8 days, approximately 1 week. It was shown that more training data lead to better recall in Fig. 28 but to less precision for both of classifier 1 and 2 in Figure 25. One new metric, the CC metric, showed an ability to slightly increase precision after the first month and after link prediction was applied to the study group and it showed a significant ability to increase recall over the entire period of analysis. Although link prediction is a hard problem and is likely to remain a hard problem, the study of the link prediction problem is presented as a way to gain more insight into the dynamics exhibited by open large-scale online social networks.

CHAPTER V

EXPERIMENTS WITH CONTENT-BASED LINK PREDICTION

V.1 Introduction

The link prediction metrics explored in Chapter IV were limited to topology. They were based on the abstract view of the network as a graph structure with nodes representing users and edges representing friendships between users. The AA metric was chosen because it was found to be the strongest metric for link prediction in social network like data in prior work that examined the major known topological metrics (Liben-Nowell and Kleinberg, 2003). The CC metric from Chapter IV was discovered by examining the clustering coefficient of the potential new friends for the user in which link prediction was being performed. It was discovered that the clustering coefficient calculated on a restricted sub network of the potential new friend provided a significant increase in link prediction recall without sacrificing very much in terms of precision. In this chapter the focus shifts to examining the content shared between users in an attempt to derive link prediction metrics to augment the topological metrics from Chapter IV. Additionally, there is interest in better understanding the influence of content on the link dynamics of the network and its relationship to the topological metrics identified in Chapter IV. The main content shared between users of the LiveJournal network is the textual blogs each user maintains. Each user receives an aggregated feed of blog posts from those on their friend lists in near real-time. This feature is very similar to the “Top News” and “Most Recent” features in Facebook.

A number of different content-based metrics to support link prediction were examined during the research for this dissertation. All of the metrics investigated are based on the standard $tf - idf$ weighting calculation from the information extraction field of study that views documents as a vector of terms with each term in the vector weighted by both its frequency within a document and inversely by its frequency throughout the document collection. The main difference that separates each of the eight content-based metrics in this chapter is the choice of what constitutes a term in the document vector representation. All content-based metrics are referenced as CBM# where CBM stands for content-based metric and the # character is a number.

The CBM1 metric uses single tokens separated by a single space as a term in the document vector representation for blogs. Most likely a single token is a word but no restriction is placed that requires a token to be a word. CBM2 uses noun phrases of length two or more as a term in the $tf - idf$ document vector representation. If the noun phrase "Robert Gates" is recognized in the blog text it will be transformed into the term robert_gates and will occupy one dimension of the document vector for the blog. Both CBM3 and CBM4 map extracted noun phrases into Freebase "domains" and "types" which can be viewed loosely as ontological concepts. In the case of "Robert Gates" being extracted as a noun phrase from a blog, the noun phrase would receive a mapping into all available types in the Freebase ontology associated with "Robert Gates". One such type for "Robert Gates" is the type military_person. The term military_person would become a dimension of the document vector representation for that blog. CBM5 through CBM8 examine attaching sentiment to the extracted noun phrases. Four

approaches to recognizing sentiment are explored. Two of the approaches seek to only recognize a noun phrase as having been referred to in a subjective way without regard to the polarity of sentiment. Under these first two approaches if the noun phrase "Robert Gates" was referred to with either positive or negative sentiment it would be included in the document vector representation for the blog in which it was extracted. If no sentiment was detected it would not be included. Two further approaches were examined that sought to determine the polarity of sentiment associated with a noun phrase referred to in a subjective way. If "Robert Gates" was referred to with positive sentiment the term robert_gates_pos would be generated and occupy a dimension in the document vector. All metrics seek to map extractions into a single term that will occupy a dimension in a document vector representation. Each document vector for a user is constructed using two weeks of prior blog posts by that user from the point at which prediction is being calculated. Once document vectors have been computed for users, their current friends and their potential new friends, cosine similarity is computed between these groups to construct both training and test metric data. The training and test metric data is applied to the same Naïve Bayes classifier discussed in Chapter IV to train and compute predictions.

Equation 3 (Eq3) below provides the standard $tf - idf$ weighting function where idf_t is the term frequency of term t in document d. The idf_t term in the equation below is the inverse document frequency for term t and is defined in Equation 4 (Eq4). The value N is the number of documents in the collection and df_t is the number of documents where the term t is seen.

$$(Eq3) \quad tf - idf_{t,d} = tf_{t,d} \cdot idf_t$$

$$(Eq4) \quad idf_t = log \frac{N}{df_t}$$

One of the goals of the $tf - idf$ calculation is to filter away terms that are common across the document collection and less likely to provide discrimination between different documents. Another goal of the $tf - idf$ calculation is to amplify terms based on their appearance within a document. For further discussion on the $tf - idf$ calculation the reader is recommended to consult (Manning et al., 2008). This source is also good for a background on representing documents as vectors. Once documents have been converted into vector representations using the $tf - idf$ weighting they can then be compared using the standard cosine similarity function. If we allow $\overrightarrow{V}(d_1)$ to represent the vector derived from the document using the $tf - idf$ calculation in Equation 3, then cosine similarity is defined in Equation 5 (Eq5). The term $\left|\vec{V}(d_1)\right| =$ is the Euclidean length of d which is defined as $\sqrt{\sum_{i=1}^{M} \vec{V}_i^2(d)}$.

$$(Eq5) \quad sim(d_1, d_2) = \frac{\vec{V}(d_1) \cdot \vec{V}(d_2)}{\left|\vec{V}(d_1)\right|\left|\vec{V}(d_2)\right|}$$

The cosine similarity function defined in Equation 5 is the primary function used for all of the metrics in this chapter. In this sense, the content-based metrics are based in the notion of document similarity.

All content-based metrics are calculated between two users of the LiveJournal network where the document for each user consists of all previous blog posts for that user made over a two week period prior to the point at which link prediction is being calculated. Longer ranges of time were explored and it was found that longer ranges of time did provide a very slight improvement in performance for the metrics discussed in this Chapter. However, given the computational resources available for this dissertation, the small improvement achieved did not out weight the increase in the computation cost of processing an extra two weeks worth of blogs data for each of the five analysis points.

In summary, the primary difference between each of the content-based metrics explored is the choice for what constitutes a term in the vector representation of documents that are ultimately used as part of the cosine similarity function calculation. In CBM1 (content-based metric 1) a term is simply a space delimited token taken from blog text. It is most likely that each token in the CBM1 metric is a word but there is nothing preventing it from being just a collection of non-space characters separated by a space. CBM2 makes use of the Natural Language Processing (NLP) library, OpenNLP, to identify noun phrases of length 2 or more. The noun phrases extracted from the blog text become the terms for the vector representation of the document used in the cosine similarity calculation between users. The CBM2 metric is the first metric where restricting the text that is processed to English text becomes important as the OpenNLP libraries are not configured for multi-lingual text. The next section will examine a fuzzy English classifier that was used to reduce the original study group from Chapter IV to a set of users who posted primarily in English.

The remaining metrics, CBM3 through CBM8, involve the use of a ontology and what is commonly referred to as sentiment analysis to attach semantics to the noun phrases recognized during the calculation of the CBM2 noun phrases. Metrics CBM3 and CBM4 involve the use of the Freebase ontology. Presently the Freebase ontology only supports English entity types. The Freebase ontology was used to map extracted noun phrases into Freebase types which can be loosely viewed as ontological categories. This mapping was only possible when Freebase contained an ontological entry for a recognized noun phrase. Freebase is considered a "folksonomy" in the sense that ontological categories for entities (i.e. noun phrases) are suggested by registered users of Freebase. Freebase is essentially a wiki with the current goal of constructing a common sense ontology of entities, most of which have an article in Wikipedia. It is a system that is and will remain under constant evolution as the ontological categories of entities are refined by the users of Freebase.

The remaining content-based metrics, CBM5 through CBM8, are influenced by what is known as sentiment analysis. Sentiment analysis attempts to identify sentiment directed towards entities or topics and to determine the polarity (positive or negative) directed at the recognized entities or topics. For these metrics sentiment analysis is applied to the noun phrases identified as part of CBM2. A model of subjectivity clues is used as part of the algorithms used to identify sentiment which also requires the restriction that the processed blogs are primarily in English. The next section will examine in detail the fuzzy English classifier used to select a subset of users from the original study group of Chapter IV for use in evaluating CBM2 through CBM8.

V.2 English Language Classifier

LiveJournal is used by people all over the world and therefore the content posted by users is both multilingual and composed of multiple character sets. This adds complexity to the analysis of textual blog content to support content-based link prediction metrics. An additional complexity is that some blog posts can contain a mixture of both different languages and different character sets. For example, a blogger in Russia may post English text composed of Basic Latin characters and in the same post make use of the Cyrillic character set. Handling all languages and all character sets was not a practical option for this dissertation. Only Basic Latin characters are considered in the study of content-based metrics derived from the blog posts of public users. It was necessary to make an effort to ensure that the users that were studied posted primarily in English. To achieve this goal a fuzzy classifier was required that was capable of scoring the degree to which English was used by each user. Users who did not post mostly in English relative to the score produced by the classifier were not considered in most of the analysis that follows. The fuzzy English classifier was only applied to 9600 users from the original study group of Chapter IV for which link prediction was being directly applied. The classifier was not used to filter the study groups' current friends or the new friend candidates.

While user provided location information can provide a probability of the language used by a blogger it is not an exact indication of the language chosen by the blogger. Additionally, not all users report location information. This dissertation takes

the view that location alone is not sufficient to determine the language used by the blogger and therefore the language type needs to be derived from the blog content.

The approach used to construct a classifier to detect English usage in the LiveJournal blogs has its basis in the idea of "the principle of least effort" and how it manifests itself in human communication. This principle argues that "people will act to minimize their probable average rate of work (i.e., not only to minimize the work that they would have to do immediately, buy taking due consideration of future work that might result from doing work poorly in the short term)" (Manning and Schutze, 1999). The evolution of language is not something that this dissertation addresses. However, this "principle of least effort" seems to have had an influence on written and spoken language by there being few words that are used with great frequency and many words that are used sparingly. A relationship exists between a word's frequency of use and its rank or position in a sorted list of words from an analyzed text. (Manning and Schutze, 1999). This relationship says that there is a constant k such that $f \cdot r = k$, where f is the frequency of the word and r is the word's rank. This relationship between word frequency and rank is often referred to as Zipf's law (Manning and Schutze, 1999). What Zipf's law essentially says is that there are a few high frequency words that dominate in usage within a language, at least the Indo-European languages. An additional empirical observation that can be made from looking at ranked lists of English word frequency taken from texts is that the most frequent words are smaller in length than the less frequent words. The classifier used in this study makes use of a very small set of words

containing the highest frequency words in the English language of lengths 2, 3, and 4 that appeared in the blogs of users.

The output of the fuzzy English classifier used to determine the subset of users from the study group who post "mostly" in English simply provides the percentage of all of the words of length 2, 3, and 4 seen in the blogs of users that are in the set of the most frequent English words of length 2, 3, and 4. Only text composed entirely of Basic Latin characters was considered. Once this percentage is calculated a threshold can be chosen that determines if a user is labeled as "English" or "non English". There are four main steps in the algorithm used to construct and apply the fuzzy English classifier. The four steps used are listed below.

Table 8
Ranked Lists of Words of Length 2, 3, and 4

length 2	frequency	length 3	frequency	length 4	frequency
to	1440984	the	2305725	that	611903
of	1103426	and	1355785	have	355774
in	787721	for	477211	with	336828
is	753429	not	456098	this	294035
it	528403	you	412286	will	213136
on	384149	was	351393	they	185594
my	364394	are	284237	from	183570
be	284610	but	274892	what	153527
de	265105	all	167670	just	153496
as	260063	can	157238	like	146611

Step 1 involves developing a ranked list of word frequencies taken from the blog collection. All blog posts that were gathered through the monitoring of the "always-on"

Atom feed for the period of 2007-03-15 to 2007-05-02 were analyzed by counting the frequency of all words of length 2, 3 and 4 and creating a ranked list of these words. Table 8 on the previous page provides the top ranked words for each length.

In Step 2, the top ranked English words are selected from the ranked lists constructed in Step 1 to represent the most frequent English words of length 2, 3 and 4. Table 9 below shows the number of unique words of each length chosen. Of the 2835 unique words of length 2 identified in Step1, 20 words were selected representing 0.70% of the 2835 unique words identified. These 20 words represented 77% of all of the words of length 2 seen in the blog posts for the period of 2007-03-15 to 2007-05-02.

Table 9
Chosen Ranked Words

length	num of unique words	word counts
2	2835	10399224
3	25339	10995214
4	77289	8220448

length	num of top words used	word counts
2	20	8055914
3	37	8319477
4	58	5392014

length	% of unique words used	% of word counts
2	0.70%	77.00%
3	0.15%	76.00%
4	0.08%	66.00%

In Step3, using the reduced list of words identified in Step2, all blogs for all users over the period of 2007-03-15 to 2007-05-02 are analyzed to determine the percentage of

words of length 2, 3, and 4 that are on the list of the most frequent English words of length 2, 3, 4 identified in step 2. Fig. 29 below shows the percentage of 2, 3 and 4 letter words in each user's blogs that are included in the reduced list of English words identified in Step2.

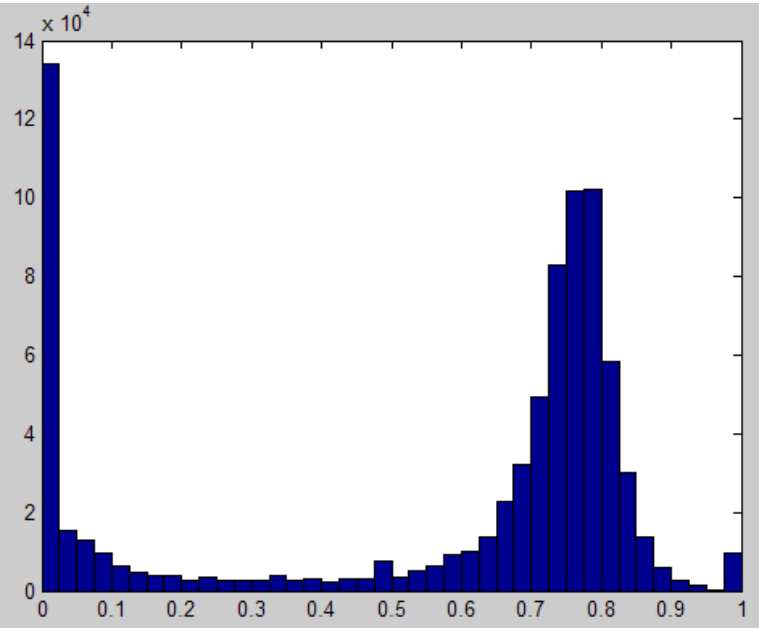

Fig. 29. Distribution of Word Usage

The use of the English language clearly dominates in the LiveJournal blogs that use a Basic Latin character set during the period of analysis. After calculating the percentages shown in the above figure subsets of users can now be chosen based on how much the 2, 3, and 4 letter words used in their blogs are the most common 2, 3, and 4 English words identified in Step2. This is an "after the fact" analysis in the sense that at the time of prediction future information is being used to select a subset of the users in the study group when scoring link prediction approaches. However, many other

approaches could provide the language identification feature that is being mimicked through the use of the fuzzy English classifier that would be equal to or more precise. The use of the fuzzy English classifier is to ease the burden of analysis and does not introduce computation that could not be reproduced at the time of prediction. The goal is to assess the performance of the content-based metrics. Making sure that the set of users analyzed are blogging in English allows for a more true characterization of the content-based metrics.

In Step 4, based on the percentage of appearance of the most common 2, 3, and 4 letter English words, subsets of users can be created from the original study group used in Chapter IV by applying a threshold to the percentage values calculated in Step3 representing the amount of "English" used in their blogs. Different values for the threshold value are explored in Experiment 2 in the next section.

V.3 Experiments

All of the link prediction experiments using content-based metrics were performed using two weeks of training data and a recall window of 1 week. Setting the recall window to 1 week means that recall is scored based on the number of correct future friendships that were predicted for the week after prediction was performed. This experimental setup is the same as the setup used in Chapter IV which focused strictly on topological metrics. The original study group outlined in Chapter IV is the same one used in this chapter. For many of the content-based metric experiments the original study group is filtered based on the use of English in the blogs of users. A fuzzy English

language classifier was used to perform filtering of the original study group. In all of the experiments that follow a user document consists of all the blog posts generated by a user over a two week period prior to the point of link prediction. For example, if link prediction is being performed on 2007-03-16 then a document for each user would consist of all of a user's blog posts over the period from 2007-03-02 to 2007-03-16. Documents representing longer periods of posting (i.e. 1 month, 1 & ½ month) were examined but only marginal increases in performance were gained from a longer window at the cost of significant computation. Therefore, to make the computation reasonable given the available resources all user documents consist of two weeks of prior blog posts.

V.3.1 Experiment1, AA + CC + CBM1

In the first experiment the terms that make up the document vector representing a two week collection of user's blog posts are simply space delimited tokens extracted from the blog posts. Most likely each term is a single word. The creation of training and test sets is exactly the same as the experiments shown in Chapter IV. Additionally, the same Naïve Bayes classifier from the Weka took-kit which uses kernel density estimation for bin estimation was used in all of the link prediction experiments that follow. Use of the same Naïve Bayes classifier was chosen because the primary focus is on comparing the performance of different metric sets and not the performance of different classifiers. Early on during research different classifiers were examined such as Decision Trees, Neural Networks and Bayes Networks; however the best performance was achieved using the simple Naïve Bayes classifier. It is possible that other classifier

approaches may provide better link prediction. The exploration of this is reserved for future work.

The scoring of prediction results is performed using the standard definitions of precision and recall provided in Chapter IV. Precision is calculated as $TP/(TP + FP)$. The variable TP corresponds to true positive and reflects correctly predicted future friendships that will eventually materialize by 2007-12-31. The variable FP corresponds to false positives referring to friendships that were predicted and never materialize. Recall is calculated as $TP/(TP + FN)$. The variable FN corresponds to false negatives representing friendships that will materialize during the 1 week recall window but were not predicted by the classifier.

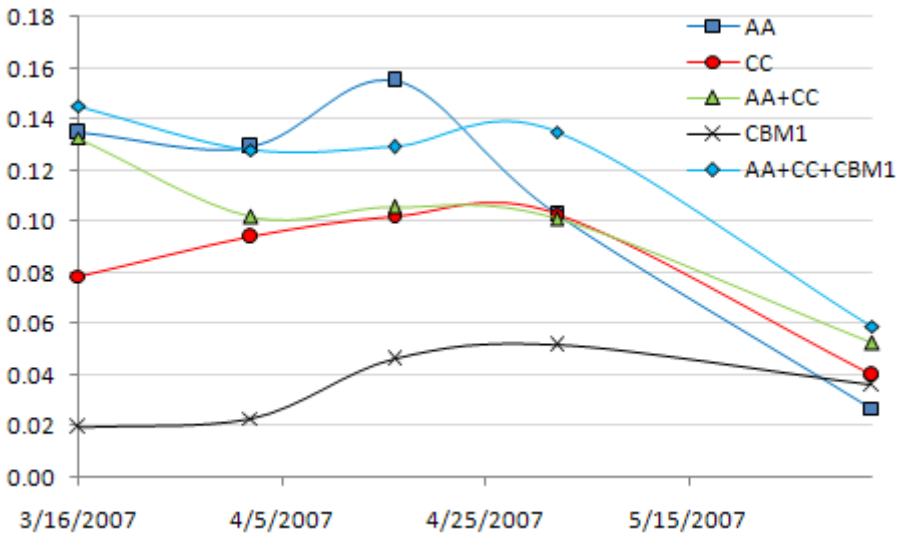

Fig. 30. CBM1 Precision Comparison 1

Fig. 30 shows the precision scores for five separate sets of metrics. The metric sets in this experiment were chosen to contrast the results achieved in Chapter IV with the addition of the first content-based metric, CBM1. The first metric set labeled AA uses only the Adamic Adar metric defined in Chapter IV and is equivalent to Classifier 1 from Chapter IV. The second set labeled CC uses only the CC metric originally defined in Chapter IV. The third set uses both AA and CC together and is equivalent to Classifier 2 from Chapter IV. The fourth metric set uses only the CBM1 metric without the aid of any of the topological metrics. The fifth metric set is composed of AA, CC, and CBM1. The fifth metric set containing CBM1 and the two topological metrics AA and CC provides better precision over the range of analysis than the other metrics sets. However, metric set 1 beats the fifth metric on one of the five test points at 2007-04-16.

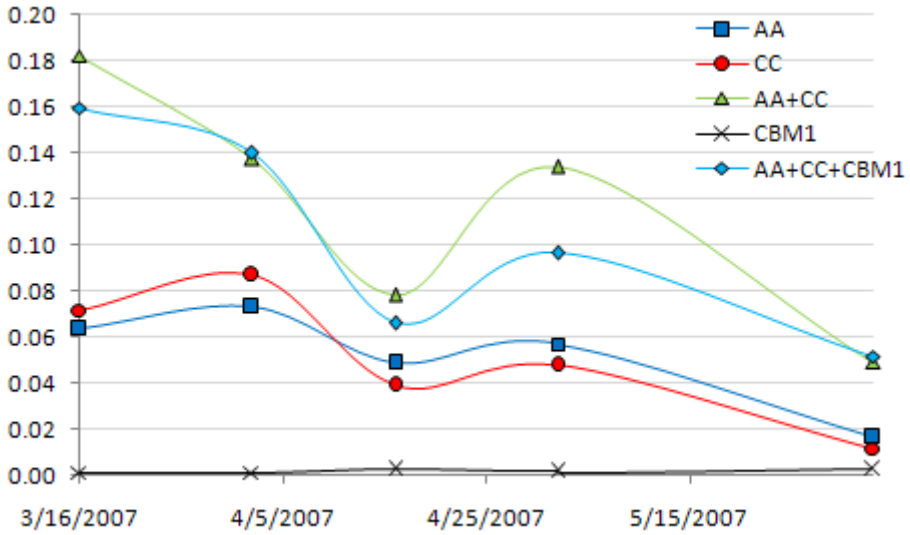

Fig. 31. CBM1 Recall Comparison 1

To assess the true impact of the CBM1 metric, recall needs to be examined. Fig. 31 provides the recall for the same five sets of metrics. The second set of metrics AA+CC provides the best recall. This is the same set of metrics, Classifier 2, identified in Chapter IV as providing the best performance. The fifth metric set provides an increase in recall over the first metric set composed only of the AA metric. Clearly the fifth metric set (AA+CC+CBM1) provides a tradeoff between precision and recall with precision being higher with metric set five and with recall being higher with metric set three. F-Measure is defined as $F = (2 \cdot P \cdot R)/(P + R)$ were P is precision and R is recall. F-Measure provides a combined score where precision and recall are equally weighted. Fig. 32 provides the F-Measure for the five metric sets.

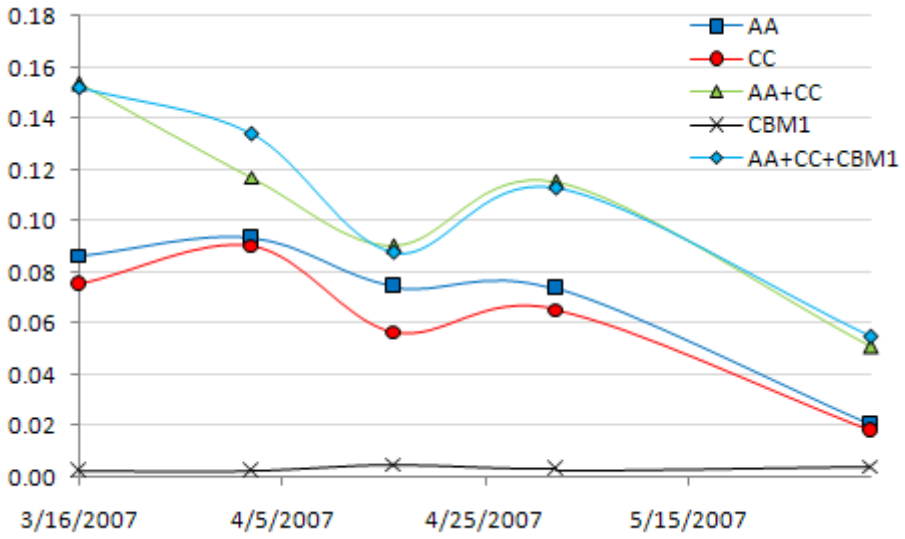

Fig. 32. CBM1 F-Measure Comparison 1

The F-Measure score of metric set three (AA+CC) and five (AA+CC+CBM1) are essentially equal with a slight advantage to metric set five at the second point of analysis. The introduction of the content-based metric CBM1 provides a choice between choosing more precision at the expense of recall. The addition of the CBM1 metric provides an average of 20% increase in precision for the range of analysis over the best performing metric set of Chapter IV, metric set three. One observation is that metric set one (AA), in terms of precision, peaks earlier than metric set four (CBM1). This could be reflecting the notion that structure is a more dominate force earlier in the network dynamics than the content in user blogs, at least for the $k2$ neighborhood. Additionally the content-based metrics when used alone (i.e. metric set four using only CBM1) perform very poorly without the addition of the topological metrics. This provides evidence that topological structure might be a more dominant force in the network dynamics than the content communicated through the blogs by users.

V.3.2 Experiment 2, Determination of a Fuzzy English Classifier Threshold

As mentioned, the original study group contains a broad spectrum of languages used by its members. Metrics which make use of Natural Language Processing techniques would be best served by reducing the original study group down to a set of users were English is the primary language used in blog postings. While the first content-based metric CBM1 used only tokens separated by a single space, CBM2 uses noun phrases of length two or more extracted from blog text using the OpenNLP library. CBM2 is more likely to be sensitive to the language used in the blog posts then CBM1.

The CBM2 metric is calculated in the same way as CBM1 with the exception that a term t is now a noun phrase of length 2 or more instead of a token of length 1. CBM2 is computed using cosine similarity calculated between two document vectors d_1 and d_2 where each document vector is computed from the collection of blog posts for each user over the prior two weeks before the point of prediction. A term in the document vector for a user is the frequency count of a noun phrase of length 2 or more located in the previous 2 weeks of a user's blog posts. CBM2 forms the basis for all of the remaining content-based metrics. Metrics CBM3 and CBM4 involve the mapping of individual noun phrases into ontological concepts using in the Freebase ontology. Metrics CBM5 through CBM8 involve the application of sentiment analysis applied to the extracted noun phrases.

The goal of Experiment 2 is to better understand the effect of varying the threshold of the fuzzy English classifier used to detect English usage in user blogs. Once a proper threshold value is established it can be used uniformly across all of the remaining content-based metric experiments presented. To gain an understanding of the effect of different threshold values, a number of experiments were carried out where the threshold value was slowly incremented and the prediction results provided by CBM2 for the reduced set of study group members were examined. Another area of concern is the size of the reduced study group. While it is desirable to achieve high precision when identifying English users it would be desirable to not reduce the size of the original study group more than necessary.

Figure 33 shows the percentage of the original user group remaining after different threshold values are applied. The y-axis corresponds to the percent of the study group remaining when a threshold value from the x-axis is applied. The x-axis corresponds to the percentage value used for the fuzzy English classifier threshold. The biggest drop occurs from the 70% to 80% for the threshold value. From 70% to 80% the size of the original study group that remains moves from 23% down to only 7%.

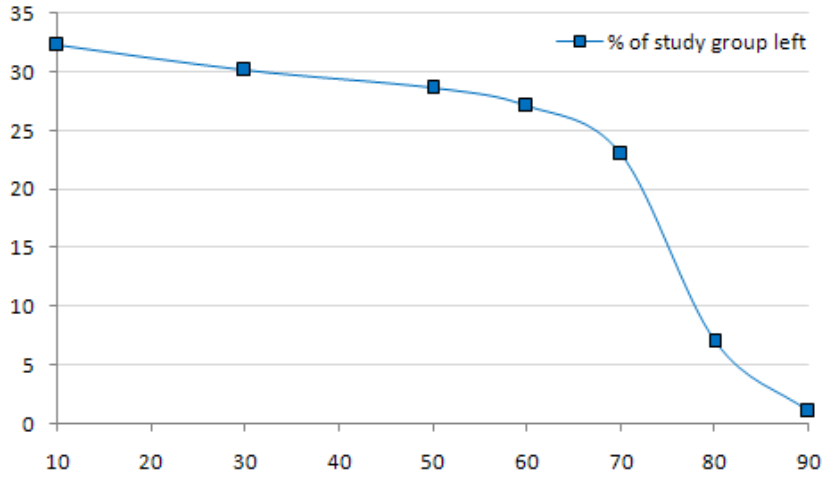

Fig. 33. Effect of Threshold on Study Group Size

To further understand what an appropriate threshold value should be, the CBM2 metric is examined using different reduced study groups based on applying different threshold values. Figs. 34, 35 and 36 provide the precision, recall and f-measure scores using threshold values ranging from 0% to 70%. The set of metrics used for the results below include the use of the two topological metrics of both the AA and CC metrics

used in Chapter IV and in Experiment 1. A reference to CBM2 in Figs. 34, 35, and 36 implies the inclusion of the AA and CC metrics.

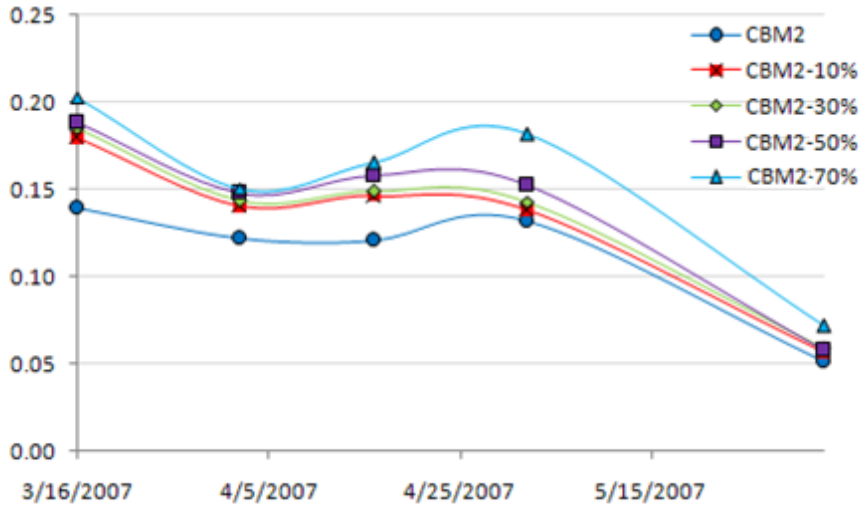

Fig. 34. Precision vs. Threshold Value

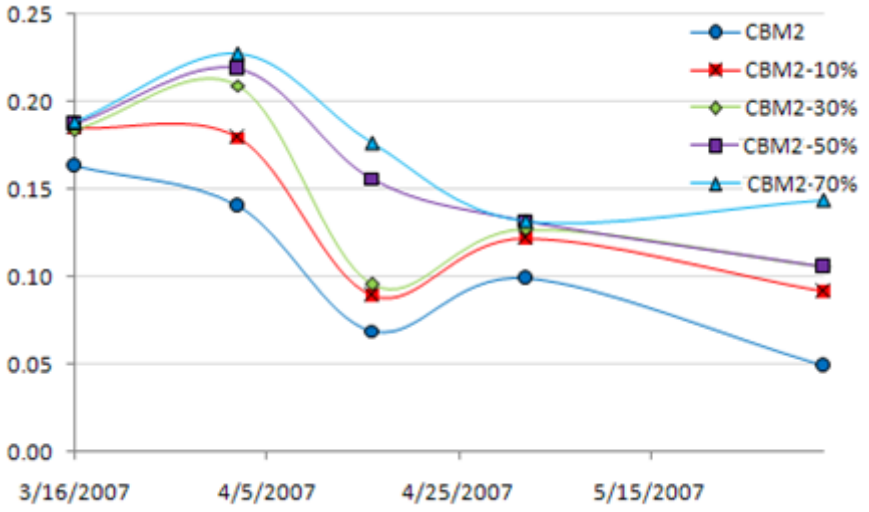

Fig. 35. Recall vs. Threshold Value

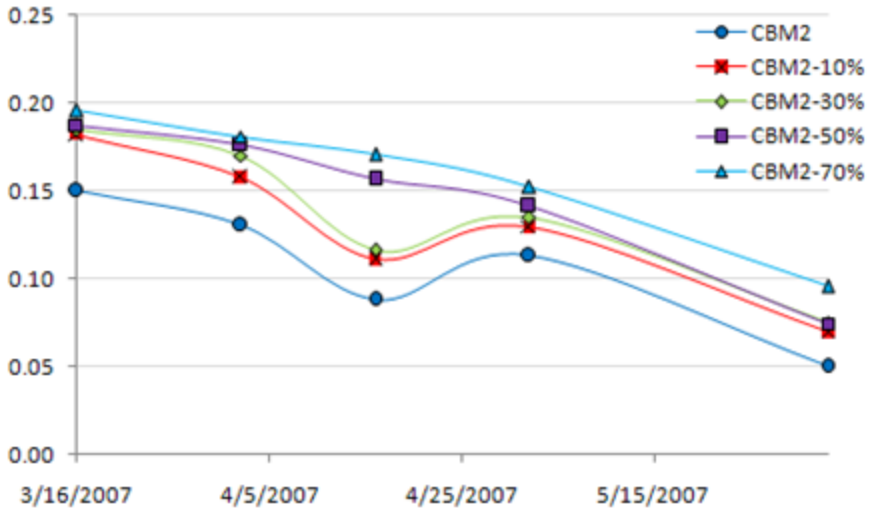

Fig. 36. F-Measure vs. Threshold Value

With the goal of maximizing the performance of the CBM2 metric without reducing the size of the study group beyond what is necessary the threshold value of 0.70 or 70% was chosen. A threshold value of 0.70 reduces the size of the study group to approximately 23% of its original size.

Table 10
Application of English Classifier to English Texts

Books	% of ELM
Moby Dick (Melville)	77
Dracula (Stoker)	78
Sherlock Holmes (Doyle)	79
Tom Sawyer (Twain)	75
Avg	77.25

To further test the choice of 0.70 as a threshold capable of recognizing text containing English text, the fuzzy English classifier was applied to some common

English texts to see how these texts would be classified. Table 10 provides output values of the classifier applied to the common English texts. The figure shows that on average 77% of the words of length 2, 3 and 4 in the common English texts were contained in the language model represented by the fuzzy English classifier. The value of 0.70 as a threshold appears as a reasonable value to ensure that a majority of the blog text of users in the final study group is in English.

V.3.3 Experiment 3, CMB2 using a 0.70 Threshold Value

The experiments in this section used the threshold value of 0.70 to determine a subset of the original study group that will be used for the remaining experiments involving content-based metrics. The original study group is the study group presented in Chapter IV which consisted of all of the users who joined the LiveJournal network on 2007-03-01, 2007-03-02, and 2007-03-03. The numbers for the original study group are repeated in Table 11.

Table 11
Size of Network Sample at 12-31-07

group	# of users	# of links at 12-31-07
3/1/2007	3222	25011
3/2/2007	3251	23490
3/3/2007	3127	21851

In Chapter III it was shown that approximately 15% of users were active but private and in Chapter IV it was shown that approximately 30% of the users who joined

during 03-01-07 to 03-03-07 and were part of the study group do not return after their first day and thus can be considered inactive. The total number of users who joined over the three days that were used for the original study group equals 9600 users. Of the 9600 users seen in the profile data collected, 5287 users were seen in the active user data collected from the live Atom feed over the 10 month period. With 15% of the 9600 users equal to 1440 and with 30% of the 9600 users equal to 2880, approximately 5280 users would be expected to be seen in the LiveJournal Atom feed. This number is very close to the 5287 users actually seen in the LiveJournal Atom feed. This number reflects the actual size of the active public portion of the original study group used for analysis in Chapter IV. A threshold of 0.70 for the fuzzy English classifier applied to the original study groups' blogs for the period 03-15-07 to 05-02-07 reducing the 5287 users down to 1216 users. These 1216 users comprise the new study group used to examine the application of content-based link prediction metrics. The fuzzy English classifier was applied to users over the period of 03-15-07 to 05-02-07; users who did not post during this period of time were filtered out. Some users who scored less than 0.70 when the fuzzy English classifier was applied to their blogs may have been English posters. Additionally, some posters post in multiple languages, including English, and may have been filtered out.

Figs. 37, 38 and 39 compare the precision, recall and f-measure results from Chapter IV, Experiment 1 in this chapter and the application of the CBM2 metric. Comparisons are made using both the original and reduced study group. The first four metric sets referenced in the legends of Figs. 37, 38 and 39 involved link predictions

applied to the original study group from Chapter IV. The AA+CC+CBM2-70% reference corresponds to the application of the AA+CC+CBM2 model applied to the reduced study group created with the application of the fuzzy English classifier.

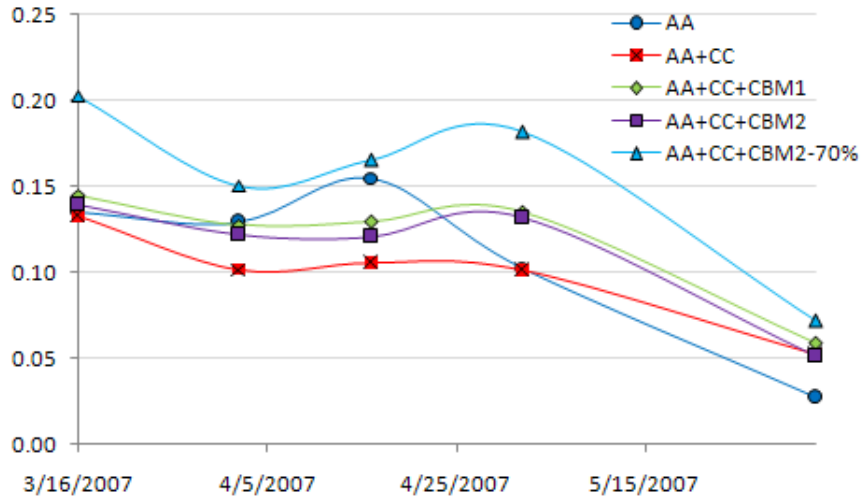

Fig. 37. Comparison of CBM2 Precision for Reduced Study Group

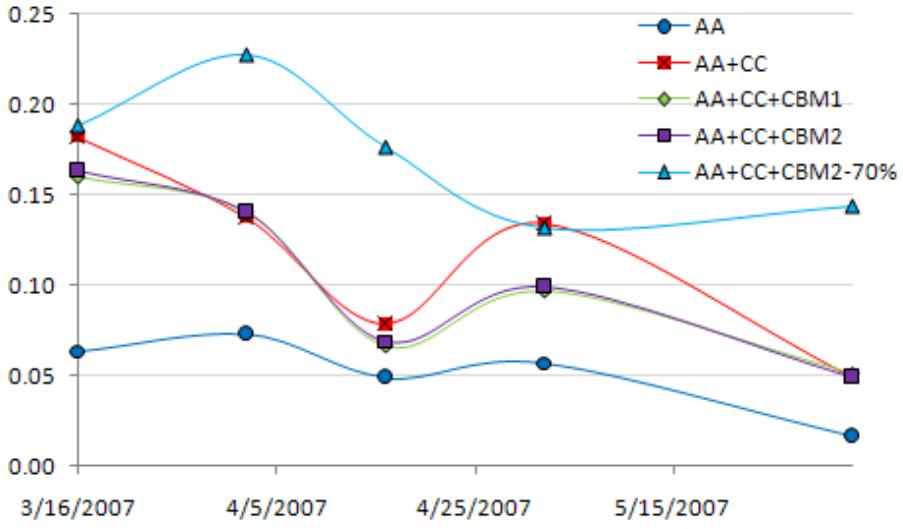

Fig. 38. Comparison of CBM2 Recall for Reduced Study Group

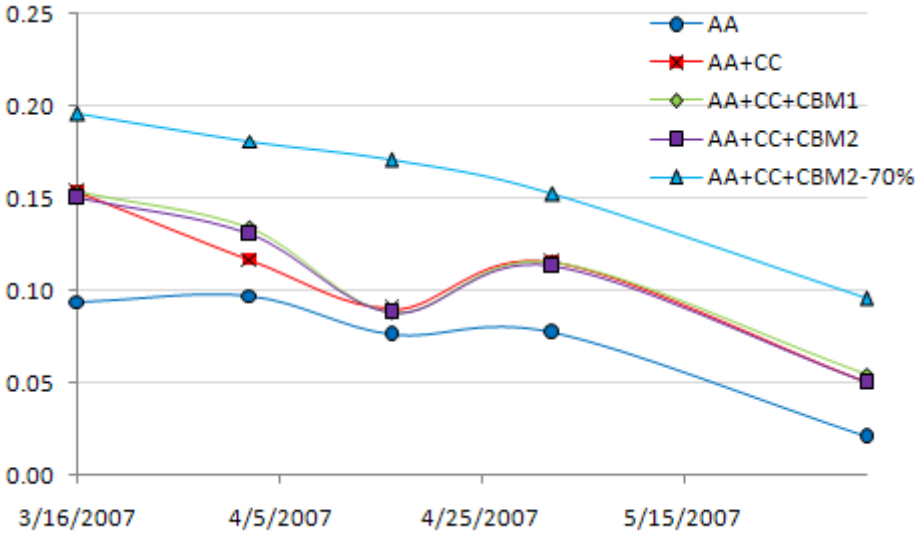

Fig. 39. Comparison of CBM2 F-Measure for Reduced Study Group

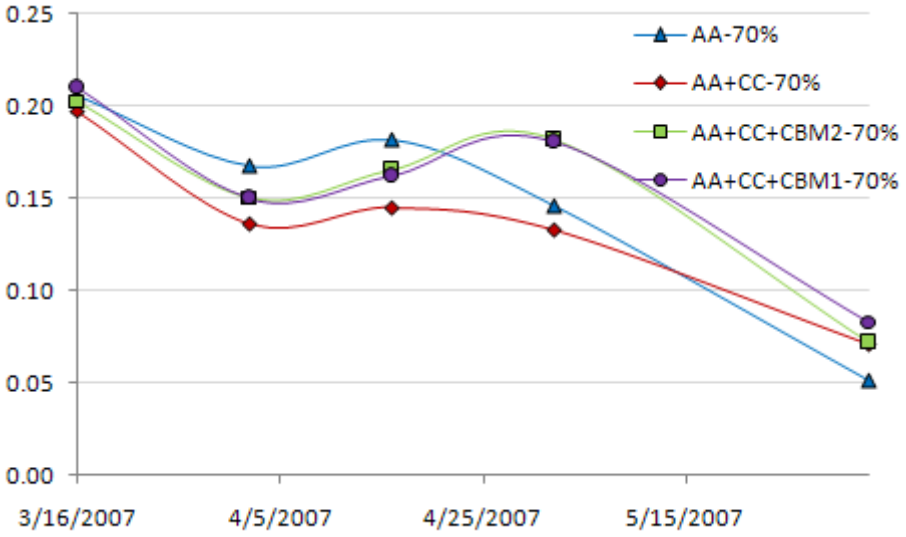

Fig. 40. Precision for All Metrics with Reduced Study Group

The AA+CC+CBM2 metric set performs better when applied to the reduced study group than the original study group both for precision and recall. However, to better understand the impact of the AA+CC+CBM2 metric set it needs to be compared

against the AA and AA+CC metric sets when these two metric sets are applied to the reduced study group. Figs. 40, 41 and 42 provide this comparison. To signify that these metrics sets are being applied to the reduced study group created with the use of the fuzzy English classifier the metric sets are referenced with a "-70%".

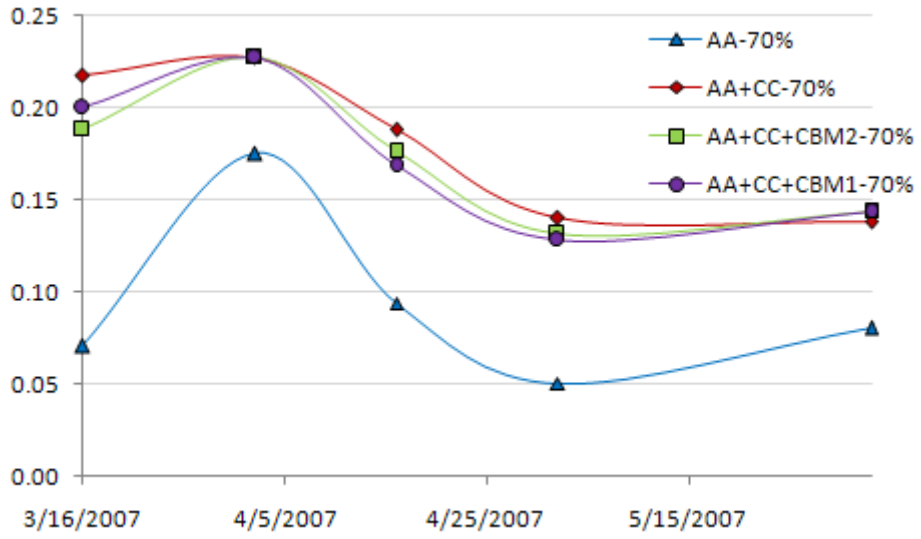

Fig. 41. Recall for All Metrics with Reduced Study Group

The metric set AA+CC no longer has the large advantage in recall over AA+CC+CBM1 and AA+CC+CBM2 shown in Fig. 38 making both of these metrics sets containing the content-based metrics dominant due to the increase in precision obtained. AA+CC+CBM1 provides results very close to equal to that of AA+CC+CBM2 showing that single words or tokens are as strong as noun phrases of length 2 or more in providing a predictive capability. This was a somewhat disappointing result as it was hoped that the increased specificity of a noun phrase of length 2 or more over a single word would have provided a stronger link prediction metric. However, a positive

observation is that the CBM2 performed as well as the CBM1 and since the CBM2 metric is the foundation for the remaining content-based metrics investigated it is good to know that the CBM2 metric itself is capable of providing equal performance to that of CBM1, both of which provided the best link prediction performance.

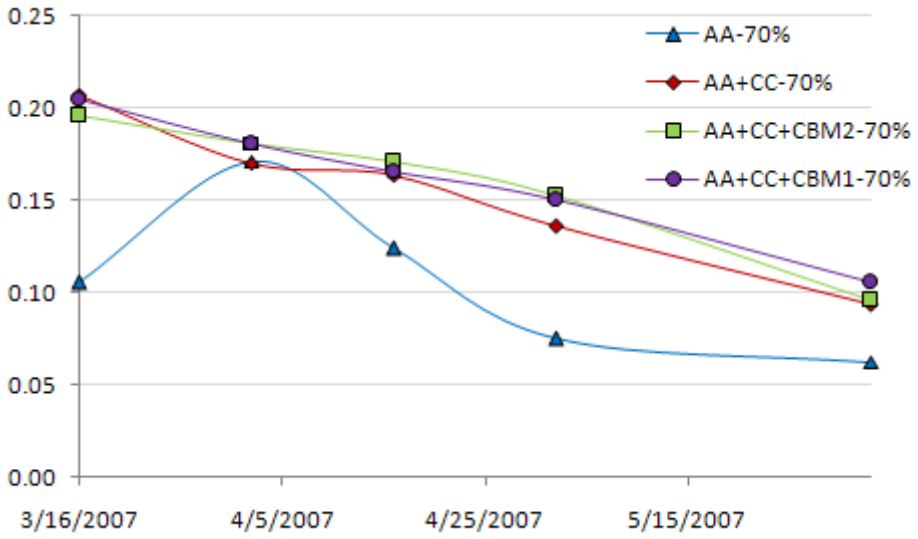

Fig. 42. F-Measure for All Metrics with Reduced Study Group

The remaining experiments will examine what boost is obtained from attempting to add different semantic dimensions to the CBM2 metric. Evident in the three previous figures is that the (AA) and (AA+CC) metric sets also experience a boost in recall performance when applied to the reduced study group generated by applying the fuzzy English classifier to the original study group. This is most likely due to the fuzzy English classifier also having the effect of selecting more active users. A correlation coefficient of 0.67 was calculated between the number of new links added during the recall analysis

window and the recall score obtained. The classifier was applied to users over the period of 03-15 to 05-02. Users from the original study group who did not post during this period of time or were not classified as English users based on their posts during this period were removed from the analysis. This is not seen as an issue since CBM2 through CBM8 will be compared using this new reduced study group. The goal of the fuzzy English classifier was only to identify users that posted primarily in English to limit the impact of non-English posters on the analysis of content based metrics. The increase in recall performance speaks to the existence of subsets of users who are easier to predict future links. Identification of these subsets of users is reserved for future work.

V.3.4 Experiment4 CBM3 and CBM4

The purpose of metrics CBM3 and CBM4 was to explore the use of an ontology to semantically enhance the CBM2 metric by mapping extracted noun phrases to ontological concepts in the Freebase ontology. Mapping noun phrases to ontological concepts was viewed as having the potential of reducing the set of all extracted noun phrases down to a smaller set of ontological concepts and potentially increasing the cosine similarity score between users who blogged about similar topic domains. If two bloggers spent a lot of time referencing different politicians in their blog the cosine similarity between the two bloggers would be increased since all referenced politicians would be mapped to the "politician" type. Therefore the CBM3 and CBM4 metrics are concerned with capturing the domains most referenced by bloggers and to compute cosine similarity based on the domains the users spent time blogging rather than the individual words or noun phrases.

The Freebase ontology was chosen for CBM3 and CBM4 because it was at the time of research and still currently is the largest ontology of common sense knowledge freely available electronically. Many of the topics in the Freebase data which map to ontological concepts or Freebase types correspond to Wikipedia article titles. The freebase ontology is known as a "folksonomy" which can be thought of as a collaborative knowledge base consisting of topics and associated metadata composed and maintained primarily by its community of members. The version of Freebase used was downloaded on 08-06-2010 and contained 13243712 individual topics.

Each freebase topic contains any number of types. These types can be viewed somewhat loosely as ontological concepts. Types are declared with a general root type followed by a more specific type. For example the entry "John Matrix" which corresponds to a fictional character in the movie Commando played by Schwarzenegger has the following type entries: /en/john_matrix, /film/film_character, /common/topic, and /fictional_universe/fictional_character. The set of general types are en, film, common, and fictional_universe. The set of more specific types are john_matrix, film_character, topic, and fictional_character. Metric CBM3 made use of the general types while metric CBM4 made use of the more specific types. Fig. 43 provides some sample entries from the Freebase ontology. The topic is seen to the left of the arrow with the corresponding Freebase types listed to the right.

Topics can have any number of types which are determined by the Freebase community of users. One problem is illustrated by the examples above is that of ambiguity. The types listed for the each topic are potentially references to different

"White Rabbit", "Robert Kennedy", or "Robert Gates" topics. This type of problem is known as an entity resolution problem where two separate and distinct entities can carry the same name. Topics in Freebase also have properties which could be used to support entity resolution, but this is a very large problem on its own and outside the scope of this dissertation. Additionally the properties for each type within Freebase were not always complete or very well defined. Only topic types were used in the calculation of metrics CBM3 and CBM4 with the goal of mapping the individual noun phrases from CBM2 into the ontological categories of Freebase.

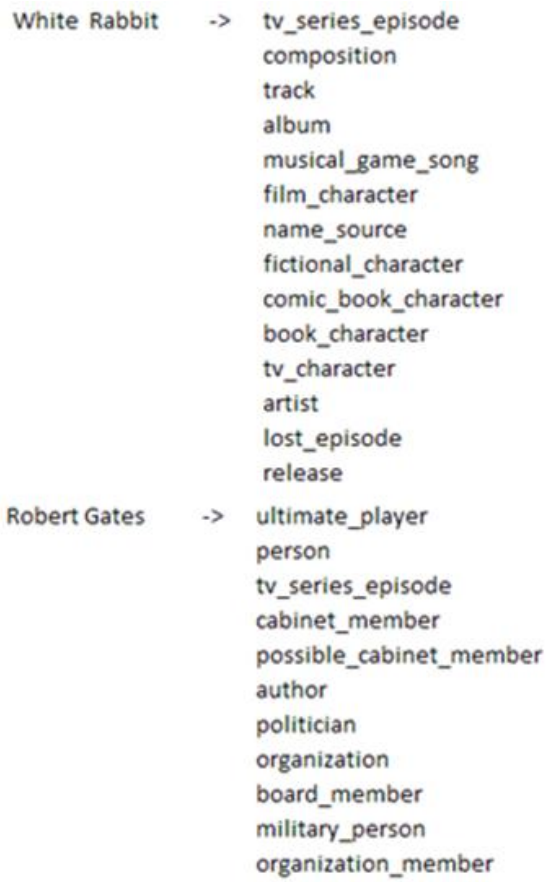

Fig. 43. Freebase Examples

The values of CBM3 and CBM4 were calculated using the same cosine similarity calculation outlined for metrics CBM1 and CBM2. The difference is that for CBM3 a term t in the document vector is a general topic value obtained by indexing the Freebase ontology with a noun phrase extracted from a user's blog. A single noun phrase can return multiple topic types and it can potentially return no type at all if the noun phrase is not contained in the Freebase ontology. CBM4 is the same as CBM3 with the exception that the term is the more specific topic value obtained by indexing the Freebase ontology.

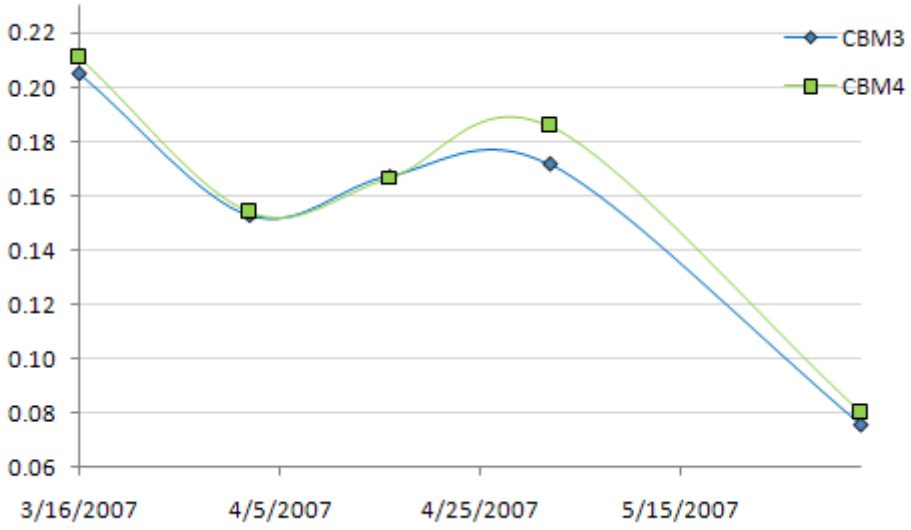

Fig. 44. Comparison of Precision for CBM3 and CBM4

Figs. 44, 45, and 46 show the precision, recall and f-Measure scores for the (AA+CC+CBM3), and (AA+CC+CBM4) metric sets. Further references to CBM3 and CBM4 in this section imply the inclusion of the AA and CC metrics. Therefore CBM3

corresponds to the metric set (AA+CC+CBM3) and CBM4 corresponds to the metrics set (AA+CC+CBM4).

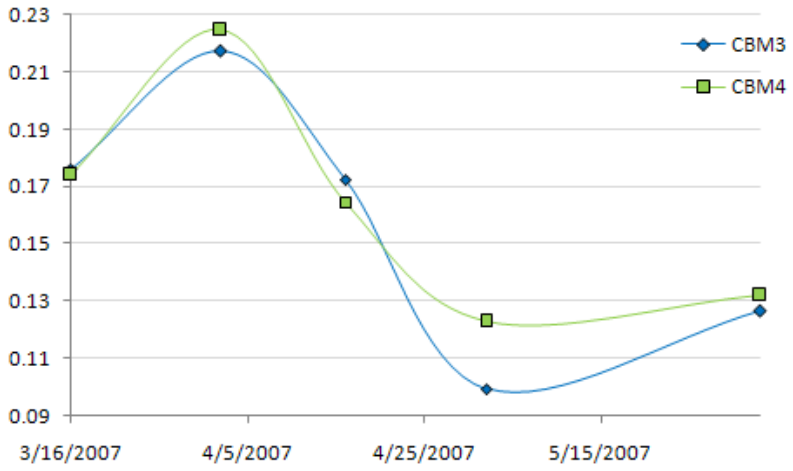

Fig. 45. Comparison of Recall for CBM3 and CBM4

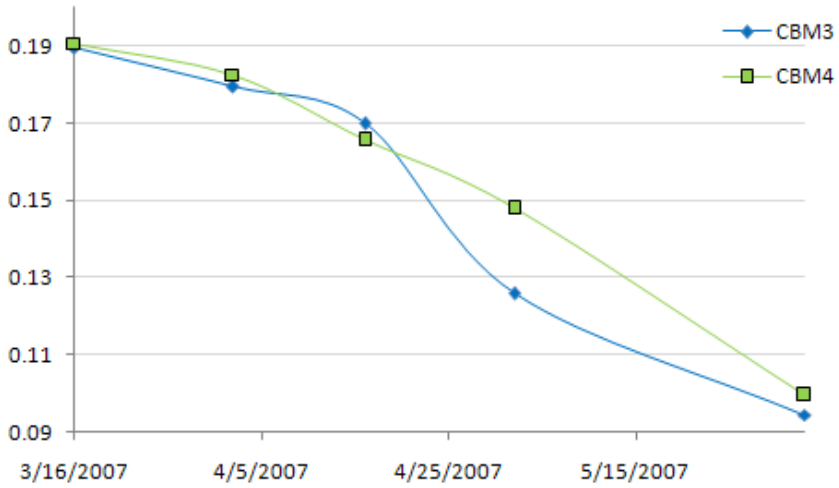

Fig. 46. Comparison of F-Measure for CBM3 and CBM4

CBM4 performs better than CBM3 on four out of the five test points showing that the more specific topic types provide a higher discriminating power. Table 12 shows

the performance of CBM4 against CBM2 which used only noun phrases and CBM1 which used only words. The comparison assumes the use of the AA and CC metrics in addition to the content-based metrics. Without these topological metrics the classifiers performance drops considerably. CBM4 provided a very slight increase in precision with a very slight decrease in recall.

Table 12
Performance Comparison for CBM2, CBM3, and CBM4

test date	CBM4	CBM2	CBM1
3/16/2007	**0.210765**	0.202978	0.210098
4/2/2007	**0.153876**	0.150482	0.149928
4/16/2007	**0.166762**	0.165099	0.162328
5/2/2007	**0.185579**	0.181726	0.180612
6/2/2007	0.079775	0.07204	**0.083093**
avg precision	**0.159352**	0.154465	0.157212
test date	CBM4	CBM2	CBM1
3/16/2007	0.17418	0.188525	0.199795
4/2/2007	0.224691	0.22716	0.22716
4/16/2007	0.164706	0.176471	0.168627
5/2/2007	0.123103	0.131535	0.128162
6/2/2007	0.132184	0.143678	0.143678
avg recall	0.163773	0.173474	0.173485

While the increase in precision achieved through the use of CBM4 is small, this increase in performance is achieved with a much smaller set of unique terms. The number of unique types for CBM3 and CBM4 respectively was 1156 and 2639. It is the unique Freebase types that make up the terms of the document vector representations which are used in the cosine similarity calculation. Reducing the space of unique terms,

and therefore the dimensions of the document vector, could potentially help other similarity calculations that could be used to support link prediction that are computationally sensitive to the dimensions of the document vector such as Latent Semantic Indexing. This line of reasoning is mentioned because it increases the relevance of achieving a document vector of reduced dimensions but is reserved for future research.

Table 13
Counts for CBM3 and CBM4

test date	CBM4-TP	CBM4-FP	CBM4-TP2	CBM4-FN2
3/16/2007	325	1217	170	806
4/2/2007	391	2150	91	314
4/16/2007	292	1459	42	213
5/2/2007	157	689	73	520
6/2/2007	142	1638	23	151
test date	CBM2-TP	CBM2-FP	CBM2-TP2	CBM2-FN2
3/16/2007	368	1445	184	792
4/2/2007	437	2467	92	313
4/16/2007	316	1598	45	210
5/2/2007	179	806	78	515
6/2/2007	160	2061	25	149
test date	CBM1-TP	CBM1-FP	CBM1-TP2	CBM1-FN2
3/16/2007	387	1455	195	781
4/2/2007	415	2353	92	313
4/16/2007	318	1641	43	212
5/2/2007	177	803	76	517
6/2/2007	173	1909	25	149

Table 13 provides the TP, FP, TP2, and FN2 counts for the CBM3 and CBM4 metrics. TP2 and FN2 correspond to the true positives and false negatives for the 1 week recall window at the point of prediction. For example for the test date of 2007-04-02 CBM4 correctly identifies 91 of the new friends made during the week after 2007-04-02 but misses 314 for the reduced study group of 1216 users.

These results were achieved with approximately 20% of all noun phrases extracted as part of the CBM2 metric achieving a mapping into the Freebase ontology. Of the unique noun phrases extracted as part of CBM2 approximately 9% achieved a mapping into Freebase. There are various factors that could affect these percentages such as erroneous noun phrases, missing topics or entities in Freebase, and misspellings. Additionally issues related to coreference could also affect whether noun phrases achieve mapping into Freebase. Often people refer to the same entity using many different lexical patterns. This can take the form of a proper noun being introduced early in the blog post and then later being referenced using a common noun phrase. One simple example would be a blog post about Barack Obama which later references this proper noun phrase with phrases such as “the president”, “president Obama”, or “Mr. Obama”. These types of references are often handled with coreference resolution algorithms.

Adding to the issue of coreference is that it is very common to use pronouns to reference a proper noun after its introduction in text. This would not address the issue of noun phrases achieving a mapping into the Freebase ontology, but it does speak to whether a document vector accurately reflects what is in the text. This issue would also

impact CBM2. In addition to addressing the issue of connecting separate references to the same entity for noun phrases of length two or more, coreference resolution could potentially provide a boost to term frequency counts of the document vector thus creating a more accurate model of what is in the blog text. Many practical approaches to this problem provide solutions with f-measures in the range of 0.6 to 0.7. The exact impact of coreference resolution on the cosine similarity calculation between users is unknown and due to the already strenuous computation required to process user blogs it was set aside for future work. Work in (Soon et al., 2001) is an excellent introduction to a practical machine learning based approach to noun phrase resolution that could both potentially improve the mapping of noun phrases into Freebase and increase the accuracy of the document vector representation.

V.3.5 Experiment5 CBM5 and CBM8

In this subsection the use of sentiment analysis was employed to indentify noun phrases that were blogged about with sentiment, first without regard to polarity and then later to identify the polarity of the sentiment associated with the noun phrase by the blogger. The same computation framework used for metrics CBM1 through CBM4, namely a document vector representation and the cosine similarity calculation, were employed for metrics CBM5 through CBM8. The difference for each of these new metrics is what constitutes a term in the document vector representation of user blogs. Two metrics, CBM5 and CBM7 seek to only identify whether an extracted noun phrase was blogged about with sentiment, they do not attempt to apply a polarity to the noun phrase. Therefore, in the construction of the document vector for these two metrics only

noun phrases identified as having been written about with sentiment are included in the final document vector. From this perspective CBM5 and CBM7 can be viewed as a filter applied to CBM2 which removes noun phrases that are considered neutral. The remaining two metrics, CBM6 and CBM8, seek to apply a positive or negative label to the extracted noun phrases. For these two metrics, sentiment analysis appends a polarity to the extracted noun phrase that makes up terms of the document vector. A brief background on sentiment analysis is now provided to better understand the approach to sentiment analysis taken in the calculation of metrics CBM5 through CBM8.

Sentiment analysis is a fairly new subfield within Natural Language Processing (NLP). NLP has its beginnings in Alan Turing's 1950 paper "Computing Machinery and Intelligence" which introduced the famous Turing test (Turing, 1950). Sentiment analysis is much younger than general NLP research with a majority of the research having been conducted with the last seven to ten years (Pang and Lee, 2008). Sentiment analysis is a very challenging field where success has often been limited to small tightly controlled document domains (Hu and Liu, 2004; Liu and Hu, 2005). When the finer details of the term "sentiment analysis" are explored it becomes clear that the term is often used to describe many different types of analysis directed primarily at unstructured natural language text. At a high-level sentiment analysis, within the context of the NLP community, focuses on identifying and summarizing subjectivity within text. Recently a large amount of research effort has been spent on locating and summarizing opinions about consumer products from customer reviews (Hu and Liu, 2004; Kushal et al., 2003; Liu and Hu, 2005). Some sentiment analysis focuses on simply classifying text as

containing subjective or objective content without regard to topic or polarity of sentiment (Finn et al., 2002; Hatzivassiloglou and Wiebe, 2000). Efforts have been placed on developing general approaches to sentiment analysis targeted at extracting both topic and associated sentiment polarity, however a majority of these approaches have been query driven (Nigam and Hurst, 2006; Kim, 2004; Kim and Hovey, 2006). A query driven approach is usually seeded in some way to focus the sentiment analysis around a predetermined topic or set of topics. A query driven approach usually means that the evaluation of the sentiment analysis is limited to a few selected domains.

There are essentially two dominant algorithmic approaches to sentiment analysis (Hu and Liu, 2004). One approach makes use of templates which are instantiated by locating instances of the slots or facets of the template within the input text. Work by (Kim and Hovey, 2006) is a good example of this approach which makes use of a set of semantic frames known as FrameNet (Filmore, 1976). Once the slots or facets of a template or frame have been satisfied the template can be instantiated. Instantiation can mean different things depending on the semantics the designer of the system attached to each template. The other dominant approach to sentiment analysis uses passage extraction (Hu and Liu, 2004). Passage extraction involves some form of NLP often seeking to identify noun phrases and sentiment clues that are pieced together in some algorithmic fashion to achieve an extraction. In this approach a final sentiment extraction is based on some general model of sentiment which is encoded into the system. One instance of this approach is seen in the system presented in (Nigam and Hurst, 2006) where sentiment is extracted based on what is termed as "the collocation assumption".

A very significant difference between the two approaches mentioned above is that the template or frame-based approach requires some form of prior background knowledge to construct templates. Template construction is expensive in terms of time and usually limiting in terms of the scope of the phenomena that can be recognized in the text. Some existing templates or frames exist that system builders can make use of such as FrameNet or PropBank (Palmer et al., 2005) but these systems do not always contain templates or frames capable of capturing the semantics of interest. Work in (Kim and Hovey, 2006) sought to expand the set of frames in FrameNet through the use of word senses in the lexical database known as WordNet (Miller et al., 1990).

Because of the nature of the blog data of the LiveJournal network and the way sentiment extractions are to be used in the support of link prediction metrics a template-based approach was not practical for many reasons. The approach to sentiment analysis must have the ability to be applied to any and all users that are being considered in the link prediction calculation. The contents of the blog data is largely an unknown and even for the relatively small number of users in the study group it would be impractical to enumerate the main topics being blogged. Even for the small reduced study group the size of the potential friend space considered in the link prediction calculation can be close to one million. Trying to assess which topics are represented or not represented and to then design templates for these topics given the time and resources available for this dissertation was not practical. Even if a few "hot topics" were located which provided an improved performance for link prediction when the sentiment analysis results for these topics were considered, these results might be misleading as the results from other topics

that were not considered may counter act or nullify the results. This statement is made not to discount investigations which seek to examine sentiment analysis applied to "hot topics" within the context of link prediction. However it was the goal of this dissertation to examine more general approaches that did not place restrictions on the topics considered where those restrictions might color the results. For these reasons the template or frame-based approach was not considered. Therefore the second approach made up of an unsupervised phrase extraction supported by a generalized model of sentiment served as the basis for the sentiment analysis used to compute metrics to support link prediction.

The definition of sentiment analysis becomes more complicated since sentiment in text can be expressed both explicitly and implicitly (Nigam and Hurst, 2006). Examples of explicit sentiment occur when the speaker or writer of the blog makes an explicit subjective reference to some entity such as "I liked that movie". Sarcasm, irony and idioms are forms of implicit sentiment and are not directly addressed in the approach to sentiment analysis used in the computation of link prediction metrics. Sentiment can also be expressed and represented at many different levels of granularity. Extracted sentiment can be assigned varying degrees of positive or negative such as strongly positive, weakly positive, neutral, weakly negative, and strongly negative. The assignment of sentiment could be numerical such as a real number from 0 to 1.

Work in (Kim and Hovey, 2006(2)) sought to first locate text from customer reviews that seemed to provide reasoning behind their sentiment and then to classify these simply as positive or negative. Sentiment can be assigned at the document,

sentence, and phrasal level. The type of sentiment assigned can be binary in the form of either positive or negative sentiment. Work in (Lin et al., 2006) sought to label text at the document level within a collection containing articles with opposing views on the Israeli and Palestinian conflict. The goal was to recognize which documents contained sentiment and then to classify those documents according to which view was supported.

Four different algorithms for sentiment extraction are explored in this subsection to support metrics for link prediction. All four approaches are based on the notion of "the collocation assumption" (Nigam and Hurst, 2006) which was found to be effective at connecting explicitly expressed sentiment to topics expressed at the phrasal level (Hu and Liu, 2004). A subjectivity lexicon was used to provide a list of words that provide an indication of sentiment and the potential polarity of the sentiment. This lexicon was part of the OpinionFinder system (Wilson et al., 2005). The subjectivity lexicon contains approximately 8000 subjectivity clues. All subjectivity clues contain a type which labels the clue as either "strong" or "weak". The polarity for each clue is either "positive", "negative", or "neutral". Only the "strong" types were used along with "positive" or "negative" polarity which provided 1718 positive clues and 3621 negative clues.

Each of the four algorithms has some common steps which include: breaking up blogs into individual sentences using the OpenNLP sentence boundary detection, extracting all noun phrases of length two or more, locating all subjectivity clues in the sentence that are contained in the subjectivity lexicon, and locating words or phrases that negate sentiment such as "not" or "never". The polarities of subjectivity clues preceded by words or phrases that negate the polarity of the sentiment were toggled to reflect the

appropriate sentiment. The remaining portions of the four different algorithms deal with the construction of the document vector representation. These portions deal with how sentiment is recognized and how the recognized sentiment impacts the construction of the document vector representation for user blogs.

In Algorithm 1 (CBM5) the goal was to identify noun phrases being referenced subjectively. For this metric no effort was made to associate the polarity of the subjective reference to the individual noun phrases. All subjective references in a sentence are associated with the closest noun phrase. This can be viewed as recognizing subjectivity at the phrasal level. This algorithm makes use of a proximity assumption that subjectivity clues are located closest to the noun phrase in which they modify or reference. This assumption was shown to provide good performance for customer reviews (Hu and Liu, 2004). The steps of Algorithm 1 are as follows: 1). Break an individual blog into sentences, 2). Extract all noun phrases of length two or more, 3). Locate all subjectivity clues, 4). Reverse the polarity of subjectivity clues preceded by words that indicate negation, 5). Associate the subjectivity clues with the closest noun phrases of length two or more, 6). Sum up the number of the associated subjectivity clues for each noun phrase where each subjectivity clue contributes 1 to the sum, 7). If the sum of the subjectivity clues for a noun phrase is greater than 1 then include the noun phrase in the document vector and set the term frequency equal to the value of the sum in Step 6.

In Algorithm 2 (CBM6) the processing was essentially the same as Algorithm 1 with the exception that it attaches polarity to the extracted noun phrases. Steps 1 through

5 are the same as Algorithm 1. In Step 6 subjectivity clues for a noun phrase are summed with positive clues contributing +1 and negative clues contributing -1 to the sum. In Step 7, if the sum of the polarity of the subjectivity clues associated with a given noun phrase is greater than one then assign a positive polarity to the noun phrase. If the sum of the polarity is less than one then assign a negative polarity to the noun phrase. Include the noun phrase as a term in the document vector with its associated sentiment polarity attached. In the final step, Step 8, when the absolute value of the sum computed in Step 7 is greater than 1 then set the term frequency for each extracted noun phrase to this absolute value.

"The second thing that bothered me was reading that **Mitt Romney**, who i consider to be a **decent** and **intelligent** man and the **best** candidate for the presidency, went on record saying that no, this **horror** does not at all **mean** that the sale of guns should be restricted and that the **second amendment** should in any way be infringed, contradicted or put on the trash-heap of history."

Fig. 47. Example Sentence

Fig. 47 provides an example sentence to help illustrate how the sentiment analysis algorithms are applied. The two noun phrases "Mitt Romney" and "second amendment" are the noun phrases recognized in the sentence. The words "decent" "intelligent" and "best" are identified as indicators of positive sentiment. The words "horror" and "mean" are identified as indicators of negative sentiment. For algorithms 1 and 2, which apply sentiment analysis at the phrasal level, the sentiment indicators are associated with the closest noun phrase. The sentiment indicator "mean" is not used as a

subjectivity clue because it is used as a verb in this example. Algorithm 1 would output "mitt_romney 3" and "second_amendment 1" for the example sentence. This output is used as part of the construction of a document vector for the user being analyzed where the "mitt_romney" is a term in the document vector with a frequency count of 3. Algorithm 2 would output "mitt_romney#pos 3" and "second_amendment#neg 1".

The first two algorithms recognized subjectivity at the phrasal level using a proximity assumption to attach subjectivity clues. The next two algorithms seek to recognize subjectivity at the sentence level. In these two algorithms the polarity value of subjectivity clues is summed across the sentence rather than within the phrase. Algorithm 3(CBM7), just like Algorithm 1 (CBM5), does not seek to attach the polarity of sentiment to the noun phrases contained in subjective sentences. Algorithm 4(CBM8) does seek to attach polarity to the noun phrases as Algorithm 2 (CBM6) did. In Algorithm 3, Steps 1 through 4 are the same as Algorithm 1. In Step 5 of Algorithm 3 subjectivity clues are associated with all of the noun phrases in the sentence. In Step 6 the subjectivity clues are summed across the entire sentence with each subjectivity clue contributing 1 to the sum. In the final step of Algorithm 3, Step 7, when the value of the sum of the subjectivity clues across the entire sentence is greater than 1 then the extracted noun phrases are included in the document vector and the term frequency for each extracted noun phrase is set to this absolute value.

For Algorithm 4, Steps 1 through 5 of Algorithm 3 are the same. In Step 6 the subjectivity clues are summed across the entire sentence with positive clues contributing a +1 to the sum and negative clues contributing a -1. If the sum is greater than or equal

to 1 then the sentence is classified as positive. If the sum is less than or equal to -1 then the sentence is classified as negative. All noun phrases in the sentence receive the same sentiment assignment. In Step 7, if the absolute value of the sum is greater than one, the polarity of the sentence is appended to each noun phrase. Each extracted noun phrase is placed into the document vector representation with the term frequency set to the absolute value of the sum calculated in Step 6. For the example sentence in Figure 47 Algorithm 3 would output "mitt_romney 4" and "second_amendment 4". Algorithm 4 would output "mitt_romney#pos 2" and "second_amendment#pos 2".

Results for metrics CBM5 through 8 were mixed. Improvements were seen with the use of some of these metrics but the improvements were very small over CBM1 and CBM2. Table 14 provides the precision, recall and f-measure for metrics CBM5 through 8 along with CBM2 for comparison. Again, as with CBM3 and CBM4, the user of metrics AA and CC were included. Therefore a reference to CBM5 through CBM8 implies the inclusion of the AA and CC metrics. The individual counts of TP, FP, and FN are shown in Table 15.

CBM5-P corresponds to precision for the CBM5 metric. CBM5-R corresponds to recall and CBM5-F corresponds to f-measure. The highest average value for precision, recall and f-measure are highlighted along with the max values for these measures on each of the test dates are in bold. CBM2 still provided the best average precision performance. CBM6 provided the best average recall performance but the increase over CBM2 was very small, on average approximately 1%.

Table 14
Performance Comparison for CBM5 through CBM8 and CBM2

test date	CBM5-P	CBM6-P	CBM7-P	CBM8-P	CBM2-P
3/16/2007	0.196	0.196	0.197	0.197	**0.203**
4/2/2007	0.149	0.142	**0.153**	0.144	0.150
4/16/2007	0.159	0.152	0.157	0.154	**0.165**
5/2/2007	0.134	0.137	0.157	0.147	**0.182**
6/2/2007	0.071	0.072	0.071	**0.073**	0.072
average P	0.142	0.140	0.147	0.143	0.154
test date	CBM5-R	CBM6-R	CBM7-R	CBM8-R	CBM2-R
3/16/2007	0.207	**0.217**	0.205	0.205	0.189
4/2/2007	0.215	0.222	0.207	0.217	**0.227**
4/16/2007	**0.188**	**0.188**	0.180	**0.188**	0.176
5/2/2007	**0.142**	**0.142**	0.135	0.135	0.132
6/2/2007	**0.138**	0.138	0.132	**0.138**	**0.144**
average R	0.178	0.181	0.172	0.177	0.173
test date	CBM5-F	CBM6-F	CBM7-F	CBM8-F	CBM2-F
3/16/2007	0.202	**0.206**	0.202	0.202	0.202
4/2/2007	**0.177**	0.173	**0.177**	0.174	0.175
4/16/2007	**0.172**	0.168	0.168	0.169	0.169
5/2/2007	0.138	0.139	**0.146**	0.141	0.143
6/2/2007	0.094	**0.095**	0.093	**0.095**	0.094
average F	0.157	0.156	0.157	0.156	0.157

Fig. 48 provides the precision scores for each metric averaged over the five test dates. CBM5 and CBM7 which only sought to identify subjectively referenced noun phrases without assigning polarity provided the best precision performance over their counterparts CBM6 and CBM8 which attempted to assign polarity to the extracted noun phrases. CBM7 which summed subjectivity across the entire sentence performed better than CBM5 which associated subjectivity clues with the closest noun phrase.

Table 15
Count for CBM5 though CBM8 and CBM2

date	CBM5-TP	CBM5-FP	CBM5-TP2	CBM5-FN2
3/16/2007	388	1595	202	774
4/2/2007	404	2307	87	318
4/16/2007	354	1873	48	207
5/2/2007	202	1308	84	509
6/2/2007	187	2433	24	150

date	CBM6-TP	CBM6-FP	CBM6-TP2	CBM6-FN2
3/16/2007	411	1682	212	764
4/2/2007	439	2654	90	315
4/16/2007	364	2029	48	207
5/2/2007	197	1241	84	509
6/2/2007	188	2425	24	150

date	CBM7-TP	CBM7-FP	CBM7-TP2	CBM7-FN2
3/16/2007	382	1554	200	776
4/2/2007	378	2092	84	321
4/16/2007	333	1791	46	209
5/2/2007	170	915	80	513
6/2/2007	177	2301	23	151

date	CBM8-TP	CBM8-FP	CBM8-TP2	CBM8-FN2
3/16/2007	382	1558	200	776
4/2/2007	425	2523	88	317
4/16/2007	352	1933	48	207
5/2/2007	181	1049	80	513
6/2/2007	188	2403	24	150

date	CBM2-TP	CBM2-FP	CBM2-TP2	CBM2-FN2
3/16/2007	368	1445	184	792
4/2/2007	437	2467	92	313
4/16/2007	316	1598	45	210
5/2/2007	179	806	78	515
6/2/2007	160	2061	25	149

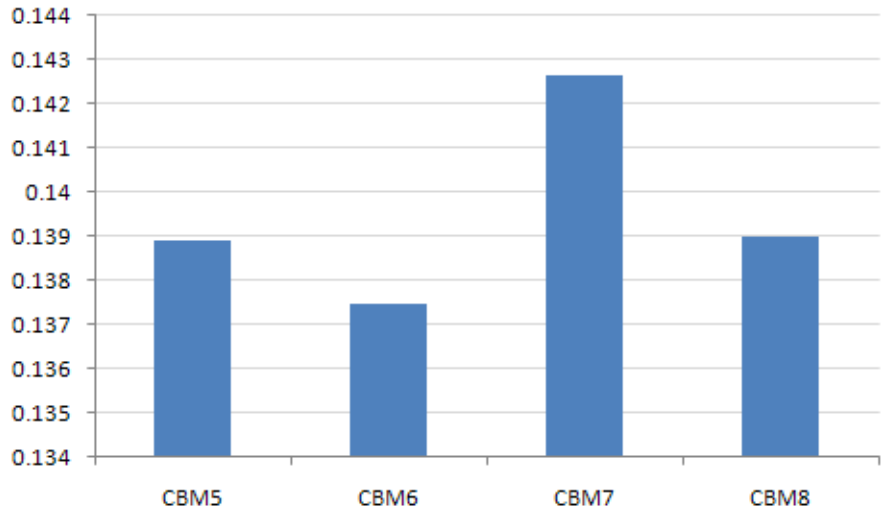

Fig. 48. Comparison of Precision for CBM5 through CBM8

Fig. 49 provides the recall scores for each metric averaged over the five test dates. CBM6 and CBM8 which attempted to label each extracted noun phrase with polarity provide better recall over CBM5 and CBM7 which did not attempt to label noun phrases with polarity. In the case of recall, the algorithms which attached subjectivity clues to the nearest noun phrase performed better. In the case of recall, CBM6 and CBM8 which attempted to determine the polarity of sentiment performed better. CBM6 which attached subjectivity clues to the closest noun phrase and labeled polarity performed the best in term of recall and as Table 14 showed provide slightly better recall than CBM2.

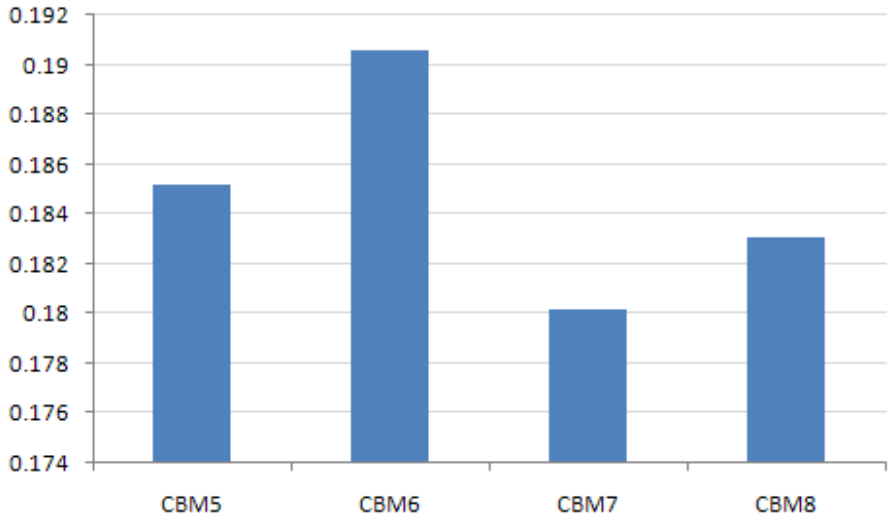

Fig. 49. Comparison of Recall for CBM5 through CBM8

While CBM6 did provide a performance improvement in terms of recall over CBM2, more work is needed to explore the space of sentiment analysis metrics if they are to be useful for enhancing link prediction.

V.4 Conclusions

This chapter explored the addition of content-based metrics to the link prediction model from Chapter IV which relied strictly on topological metrics. A definite improvement was seen with the addition of content-based metrics CBM1 and CBM2. Both of which provided a very similar improvement in precision over the best performing metric combination of AA+CC identified in Chapter IV, without a significant reduction in recall. The content-based metrics alone did not provide a strong link prediction capability and required the use of the topological metrics to achieve an increase in precision. A fuzzy English classifier was presented which identified users

who posted mostly in English. This classifier was used to reduce the original study group used in Chapter IV so that the content-based metrics which relied on Natural Language Processing techniques could be applied to users who posted primarily in English. For the reduced study group the best recall value achieved was 23% meaning that 23% of new friendships which occurred during the week following prediction were identified. The precision for this data point was 15%. An effort was made to increase these results by adding semantics to the extracted noun phrases. CBM3 and CBM4 mapped noun phrases to types within the Freebase ontology. The CBM4 mappings were to the more specific Freebase types which provided a slight increase in precision over the CBM2 metric. Mapping the extracted noun phrases from CBM2 into Freebase topics greatly reduced the number of unique terms in the resulting document vector representation for user blogs. It is possible that this reduction could aid other approaches to computing document similarity that are sensitive to the dimensions of the document vector. Research along these lines was identified as potential future work. Metrics CBM5 through CBM8 sought to apply sentiment analysis to the extracted noun phrases from CBM2. Results for CBM5 through CBM8 were mixed. CBM6 provided a slight improvement over CBM2 in terms of recall but this improvement was minor. The two algorithms (CBM7 and CBM8) which summed subjectivity across sentences provided the best precision performance while the two algorithms (CBM5 and CBM6) which summed subjectivity across phrases provided the best recall performance. The two algorithms (CBM6 and CBM8) which explicitly labeled the polarity of sentiment provide the best recall performance while the two algorithms (CBM5 and CBM7) which

did not label polarity provided the best precision performance. Topological metrics remained the dominant features to describe the link dynamics between users of the LiveJournal network. This is highlighted in Figs. 30, 31, and 32. This provides evidence that topological structure has a stronger influence on network evolution than the content exchanged between users.

CHAPTER VI

CONCLUSIONS

VI.1 Review of Dissertation

The primary goal of this dissertation was to add to the understanding of the network dynamics of open large-scale online social networks. This was accomplished by developing a methodology of acquiring dynamic social network data and the incremental development of a predictive model of the linking dynamics of the LiveJournal network.

The online social networking paradigm is arguably one of the fastest growing and most dominant trends in human computer interaction. The study of these types of networks is still in its infancy especially in the area of network dynamics. Access to complete datasets is difficult, something directly addressed in this dissertation. Additionally, there are many challenges associated with working with large dynamic graph structures. Increasing the knowledge of the dynamics of open large-scale online social networks can potentially provide many benefits. Designers of new social networking systems would be aided with more complete models of how these systems behave over time both at the macro and micro levels. Users of these social networking systems would also be aided by having a deeper understanding of structure and dynamics of these systems as they increasingly incorporate them into their lives. For social scientists, there has never been another time in history when so much human social interaction data has been available for analysis.

The remainder of this subsection highlights the main insights from each chapter. Chapter III focused the application of event-driven sampling to the LiveJournal network and provided the major characteristics of the resulting dataset. Chapter IV provided an examining of the linking dynamics seen in the network model and gave details on the development and application of a link prediction model which used only topological metrics. Chapter V focused on extending the link prediction model with the addition of content-based metrics.

Chapter III provided a detailed description of the process of sampling an open large-scale online social network using a technique termed Event Driven Sampling. The architecture of the LiveJournal site allowed for the application of this technique because of the "live Atom feed" provided by the site. The EDS technique allowed for the capture of approximately 85% of the active network link mass with approximately 98% correctness. Correctness was measured as the ratio of all additions and removals in the network model generated using EDS to all of the additions and removals seen in a verification crawl. The 15% of the active network link mass not capture belonged to private users. Along with the link dynamics all of the blog posts for public active users were collected to support analysis of content-based metrics. Chapter IV showed that the EDS technique performed well for new users entering the network with on average 86% of the links being correct over the first 30 days after new users joined with well over 90% of these links pointing to other public active uses. Analysis of the static network structure at the point of verification showed that the network model maintained the scale-free and small-world properties that would be expected of a large social network

such as LiveJournal. Results were similar to that of (Mislove et al., 2007) which provided details related to a single static snapshot of 95% of the LiveJournal network.

An examination of the linking dynamics in Chapter IV showed that for new users joining the network there is a chaotic period during the first 10 days were approximately 33% of new links that were established over the 10 month period of analysis were establish during this period. A large percentage, almost 90%, of the new links established by the study group during this chaotic period was made to users that were six of more hops away, potentially disconnected. After the first 10 days the source of new friends shifts quickly to users two hops away. For the remainder of the 10 month period 52% of new links are to users two hops away. The group of potential new friends two hops away is referred to as the k2 neighborhood. With this group being the largest source of new friends, the remaining research on link prediction focused on this group. To assess the impact of the missing 15% of the linkage mass on the calculation of the source of new links an experiment was performed which randomly removed 15% of public active users and the source of new links were recalculated. The impact of the randomly removed nodes on the source of new links calculation was negligible. Since the public users are much more strongly connected to other public users this also provided evidence that the missing 15% of link mass of active private users is also negligible.

The Adamic Adar metric was used as a baseline metric for link prediction because it was shown in past research to provide the best performance (Liben-Nowell and Kleinberg, 2003). Past efforts on link prediction made use of two static snapshots

without regard to the length of time users for which link prediction was being applied had been members of the network. The length of time before a predicted link would materialize had largely been ignored in prior research. This notion was represented as recall and was directly addressed in Chapter IV. Chapter IV showed that precision and recall values were much stronger when prediction was applied earlier after users joined the network. This analysis was possible due to the link dynamics captured for new users with the use of the EDS approach. Early experimentation, based on available classifier libraries, identified a Naïve Bayes classifier using kernel density estimation as providing the best link prediction performance. The Naïve Bayes classifier from the Weka 3.4 toolkit was used for all link prediction experiments. It is recognized that other classifiers may improve performance but the primary focus was on the analysis and comparison of different metrics and not different classifiers. Metric 2 from Chapter IV was identified providing a doubling, in some cases a tripling, of recall in most cases without sacrificing precision. It was shown that larger training windows provided better recall at the expense of precision and smaller training windows provided better precision at the expense of recall.

Chapter V presents the final area of research which focused on the addition of content-based metrics to the link prediction model presented in Chapter IV. The foundations of all of the content-based metrics were the TF-IDF document vector representation and the cosine similarity calculation to compute document similarity. Eight content-based metrics (CBM) were examined. CBM1 used single tokens delimited with a single space to populate the document vector representation. CBM2 used noun

phrases of length 2 or more to populate the document vector. Extracting noun phrases required that content of blogs to contain English text. This was accomplished with a very simple fuzzy English classifier that identified blogs containing primarily English text. The classifier was verified using commonly known English texts. It was essential that the blogs contain English with high probability since the noun phrases extracted from the text form the basis for the remaining content-based metrics CBM3-CBM8. Both CBM1 and CBM2 performed equally well. Their main contribution to the link prediction model for both metrics was an increase in precision of approximately 20% on average over the five test dates. None of the content based metrics performed well without the addition of the topological metrics identified in Chapter IV. This does not prove that topology has a larger influence on the network dynamics than the content exchanged between users but it does provide evidence. In an effort to enhance the semantics of the noun phrases communicated between users, noun phrases were mapped to the ontological types in the Freebase ontology. Freebase was chosen because it is the largest electronic ontology of common English available with approximately 13 million topics. Freebase provides both general and specific type labels for noun phrases. CBM3 explored the use of the more general types while CBM4 explored the use of the more specific types. CBM4 provide a slight improvement in precision over CBM2 with a slight reduction in recall. CBM4 performed better than CBM3. What was not seen was a drastic increase over the CBM2 metric. In a second attempt to increase the semantics of the CBM2 noun phrases sentiment analysis was applied to those noun phrases for metrics CBM5 through CBM8. Two different approaches to recognizing subjectivity were used, one approach identified

sentiment at the sentence level (CBM7 and CBM8) and the other which identified sentiment at the phrasal level (CBM5 and CBM6). Each of the two separate approaches supported the creation of two new metrics, one which simply recognized sentiment without regard to polarity (CBM6 and CBM8) and the other which attempted to determine a positive or negative polarity to the sentiment (CBM5 and CBM7). CBM6 which identified sentiment at the phrasal level and attempted to label sentiment as either positive or negative provided a slight increase in recall with a reduction in precision over the CBM2 metric.

There were three primary problems examined during the course of research for this dissertation. The first primary problem examined the challenge of gaining access to a near complete online social network dataset containing network dynamics in conjunction with the content exchanged between users. A general examination of the static structure and network dynamics of the acquired data was performed. What materialized out of the effort to acquire data was a general process of using event-based data to capture a near complete dataset of link dynamics and the associated textual blog data. This process was described in detail providing a path for other researchers to access data from the LiveJournal social networking site. Additionally, the study of the event-driven sampling approach applied to LiveJournal provides insight and a starting point for others looking at applying event-driven sampling strategies to other online social networks. Twitter, which offers a feed similar to LiveJournal, is another site where the event-driven sampling approach would be possible. Insights gained from a study of the

network dynamics of the LiveJournal data provided motivations for the investigation of a predictive model.

The second primary problem examined was the construction and application of a classifier using topological metrics to predict future links between users. A strong emphasis was placed on the development of a predictive model of network dynamics because the model can be graded on its ability to predict future changes to the network topology. The view taken is that if a component of the model provides better prediction performance, then the model is a truer description of the network dynamics. In the study of the application of topological metrics to predict future links, both the amount of time users had been members of the network and the notion of recall were examined, two aspects of the link prediction problem in general that had not received much attention.

The third primary problem examined was the addition of content-based metrics to the prediction model. The goal was to examine the relationship between the content-based metrics and the topological metrics. Topological metrics were significantly more powerful for predicting future friendships between users then content-based metrics, providing evidence that the current topology of the network is a larger factor in the topological evolution of the network than the content exchanged between users. The impact of the content-based metrics on link prediction was primarily seen as an increase in the precision of predictions.

VI.2 Importance of Research

This dissertation outlined a methodology for collecting and analyzing open large-scale online social network data. Three main contributions made by this dissertation are highlighted in this section. While the contributions of the dissertation are not limited to these three contributions, these three contributions are highlighted because they are viewed as having the greatest impact on research related to open large-scale online social networks.

The first contribution is the development of a fairly low cost path to acquiring a near complete dataset containing both link dynamics and user generated content from an open large-scale online social network. There is a struggle for researchers wishing to study these networks to acquire data while social network service providers view their data as intellectual property. The LiveJournal site is interesting because it is one of the oldest online social networking sites and many of its features are found in the two largest online social networks, Facebook and MySpace. It is a relatively smaller network which makes dealing with the entire network computationally feasible for most researchers. While being able to perform computations on the network graph and user blogs required a small five desktop PC cluster, the entire dataset was acquired with a single PC. There is little motivation for social network services providers to give their data freely to academic researchers who are interested in publishing research results openly. Work presented in Chapter III provides a path for researchers to acquire their own dataset.

The second contribution is related to the link prediction problem. Chapter IV showed that there is a period of time when users first enter the network where better

prediction results are possible. This was accomplished by following a group of users from the time they first enter the network out to approximately 10 month made possible by having a near complete dataset. The notion of recall was directly addressed showing that a majority of new predicted friendships materialize fairly early after prediction. In past studies of link prediction recall had not been addressed. The results of link prediction presented in Chapter IV were obtained from the analysis of an open large-scale online social network. Past studies on link prediction had focused primarily on research citation networks. A new metric was identified that greatly increased recall showing the existence of unidentified metrics and the possibility of that other unidentified metrics may exist.

The third contribution is related to the examination of the content-based metrics discussed in Chapter V. Simple metrics such as the use of words and noun phrases provided an improvement in link prediction precision. Some performance increase was seen with the use of the Freebase ontology and sentiment analysis; however the performance increase from these metrics was minor. By themselves the content-based metrics performed poorly when compared with the topological metrics. While this does not prove that the current topology is a stronger factor than the content exchanged between users on the future network topology, it does provide some evidence in this direction.

VI.3 Limitations of the Research

This section examines some aspects of the research that have the potential for limiting the generalizations that can be drawn from the results presented. Provided is a discussion that may be useful for other researchers interested in using results from this dissertation or extending the research. The three main areas discussed in this section relate to the architecture and interface of the LiveJournal social network in its relation to other online social networks, the choice of the study group used for examining both the topological and content-based link prediction metrics, and how the content exchanged between users of the LiveJournal network might be related to the content exchanged between users of other online social networks and in general society at large.

While the study of the LiveJournal network presented in this dissertation is large in terms of scale and timeline, care must be taken when generalizing the findings to online social networks and human social networks in general. Connecting the results in this dissertation to offline human social networks is outside of the scope of this dissertation. Although it is hoped that findings in this dissertation may help researchers that study offline human social networks and their relationship with online social networks. In terms of other online social networks, LiveJournal is one of many that currently exist. Care has been taken in the dissertation not to over state how the results obtained through the analysis of LiveJournal would generalize over all online social networks. Many of the other existing online social networks such as Facebook and Twitter contain architectural and interface choices that differ from LiveJournal and may impact dynamics between users therefore potentially limiting a direct comparison

between LiveJournal and the other online social networks. The effect of architectural and interface differences on analysis and comparisons could possibly be reduced by choosing a proper abstract representation for the network data used in analysis. This dissertation focused on representing users of the network as nodes and edges in a graph-based representation because the central area of study dealt with topology. To date this has been the primary representation chosen by other researchers examining social network like data and is most likely a good place to begin when examining a large online social network and when attempting to compare results with those provided in this dissertation.

A central goal of the research in this dissertation was to lay a foundation for future comparisons with other online social networks. This foundation seemed to be lacking in the literature for open large-scale online social networks such as LiveJournal at the time research began. The dissertation attempts to provide this foundation through the outlining of a methodology for acquiring topological dynamics in conjunction with user's content, an analysis of the topological dynamics between users, and an examination of the link prediction problem using both topological and content based metrics. As previously mentioned the Twitter website would be an excellent source for future analysis as its architecture is similar to LiveJournal's in that it would be possible to acquire both topological dynamics and user content. Because Twitter is much larger in scale than LiveJournal it would be possible to examine any dependencies of scale on the topological dynamics by comparing the analysis used in this dissertation against similar analysis applied to the Twitter social network.

In the construction of the study group used for the analysis of both topological dynamics and link prediction, care was taken to provide a group that was representative across the LiveJournal network. Approximately 10,000 users were chosen simply because they joined the network during 03-01-07 and 03-03-07. The calculations of topological dynamics and link prediction for these 10,000 users involved close to 1 million other LiveJournal users within the network model constructed. While these numbers seem to provide a fairly exhaustive examination, further analysis might examine a set of users joining over an entire week comparing the results from another study group joining over a different week. The main reason that larger study groups ranging over longer periods of time were not examined during research was largely due to limited computational resources. The periods of computation for single runs of analysis using the study group of 10,000 spanned periods from few days to few months depending on the computations involved. Computations involving the content of user blogs required much more processing time than those which examined only topology.

The present research could also benefit from a more detailed examination of both the content of user journals within the study group and the content of user journals within the entire network model to see how representative they are of society in general. To provide some insight into the relationship between the content of user journals and society in general Figure 50 is provided below. The top chart shows three topics represented simply as strings of text and their relative appearance within the journals of users in the network model for the period of 04-24-08 to 12-01-08. The three topics include the string “bailout” which was a common phrase used during the period when

the U.S. economic crisis began gaining wide public attention sometime in September of 2008. The blue line representing “bailout” in the top chart was obtained by querying for the term “bailout” across all user journals, involving millions of journals, during the period of 04-24-08 to 12-01-08. The same analysis was performed for the text strings “John McCain” and “Barack Obama” which appear as the red and orange lines respectively.

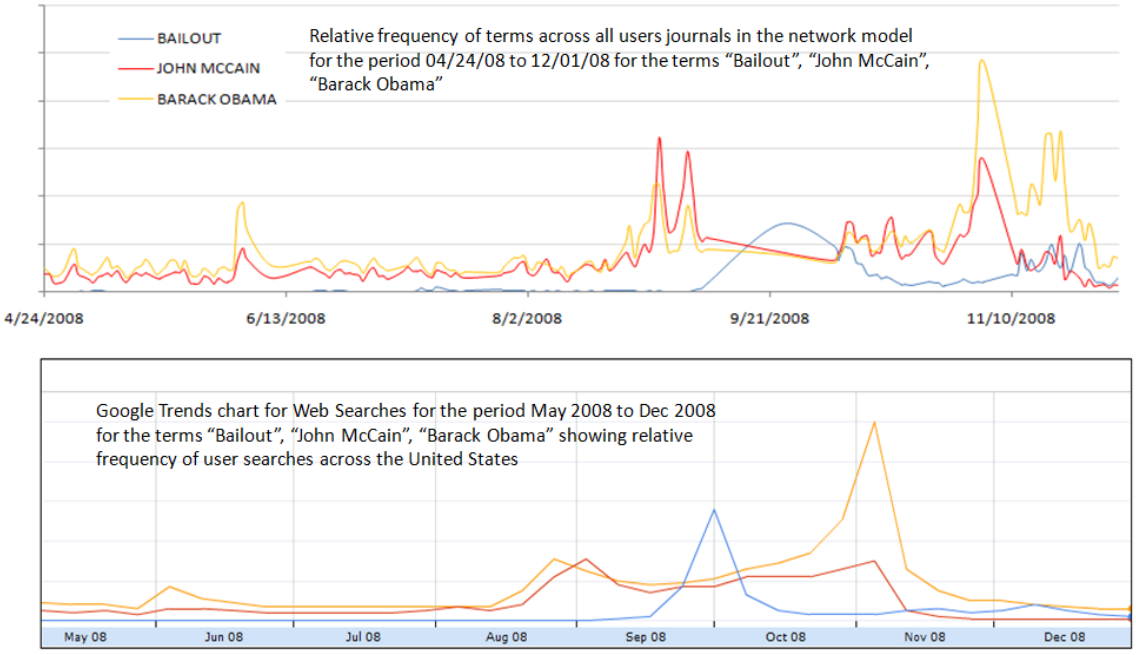

Fig. 50. Comparing LiveJournal Content with Google Searches

The bottom chart in Fig. 50 was taken from Google Trends by performing a query across all web searches within the United States using the same exact search phrases of “bailout”, “John McCain” and “Barack Obama” for the same approximate period of 04/2008 to 12/2008. The colors used to encode the search terms is the same in

both charts. The charts display a similar relative frequency for mentions of the textual search strings within journals of LiveJournal users and all web search queries entered into Google within the U.S. during the same period. While this is not proof that the distribution of topics expressed within the LiveJournal dataset matches the same distribution of topics discussed or examined within the entire U.S. population, the charts do provide evidence that the content of LiveJournal journals is representative of the interests of users of the World Wide Web within the United States. Since no restrictions were placed on the inclusion of LiveJournal users into the study group other than having joined during a certain period of time it was viewed as reasonable that the distribution of a majority of the topics within the journals of the study group would be fairly representative of a majority of the topics within the entire set of LiveJournal journals.

VI.4 Future Work

There are many areas of the research presented in this dissertation that would benefit from further research. In the area of event-driven sampling, further research could be performed to reduce the 2% of the network structure that is not captured correctly from the LiveJournal site. The event-driven sampling approach used in Chapter III was fairly simple in that sampling was driven by the appearance of a user's blog post in the Atom feed. No history of past user posting is used to inform the sampling process. Making use of user posting history may help reduce the small error experienced in the network model. Application of the event-driven sampling approach to the Twitter website would also be of value. Twitter receives significantly larger amounts of traffic

than LiveJournal and would have exhausted the resources available for this dissertation. The software that was constructed for the capturing and analysis of the LiveJournal site was optimized for use on a small cluster of personal computers.

The choice of using a Naïve Bayes classifier was made primarily because early experimentation showed that it provided the best performance. There was a stronger desire to focus research on the metrics themselves rather than on different classifier implementations that may only improve results marginally. One immediate benefit of using the Naïve Bayes approach is that it was much less sensitive to the class skew issue related to their being many more negative examples than positive examples in the training data. Additionally, the view was taken that it was easier to compare between the different link prediction metrics by consistently using the Naïve Bayes classifier throughout the analysis. Future work could explore the use of different classifier approaches to improve prediction results.

Approximately 9% of noun phrases extracted were mapped into the Freebase types. The reasons related to why more of the extracted noun phrases were not mapped are varied. There could be errors in the noun phrase extraction process. Additionally, a majority of the topics covered in the Freebase ontology are considered proper noun phrases. Many times people use multiple names to refer to the same proper noun. Therefore a proper noun such as “Barack Obama” may be included in the Freebase ontology but an extracted noun phrase such as “president Obama” is likely not included. Coreference techniques which seek to connect noun phrases referencing the same entity may help to improve the mappings between extracted noun phrases and Freebase types.

The techniques described and evaluated in this dissertation both provide insight into the collection and analysis of open large-scale on line social network data and open the door to other researchers interested in exploring social network dynamics.

REFERENCES

Adamic, L., Adar, E., 2005. Friends and neighbors on the web. Social Networks 25(3), 211-230.

Adamic, L., Glance, N., 2005. The political blogosphere and the 2004 U.S. election: divided they blog. Proceedings of the 3rd International Workshop on Link Discovery, Chicago, IL, USA, 36-43.

Alexa Top 500 Global Sites, 2010. Retrieved October 15, 2010. Available at: http://www.alexa.com/topsites

Backstrom, L., Huttenlocher, D., Kleinberg, J., Lan, X., 2006. Group formation in large scale networks: Membership, growth, and evolution. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 44-54.

Balog, K., Mishne, G., de Rijke, M., 2006. Why are they excited?: identifying and explaining spikes in blog mood levels. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations, Trento, Italy, 207-210.

Barabasi, A.L., Albert, R., Jeong, H. and Bianconi, G., 2000. Power-law distribution of the world wide web. Science 287, 2115.

Barabasi, A.L., 2002. Linked: The New Science of Networks. Perseus Publishing, Cambridge, MA.

Barabasi, A.L., 2003. Scale-free networks. Scientific American 288, 60-69.

Broder, J., Kumar, R., Maghoul, R., Raghavan, P., Rajagopalan, S., Stats, R.; Tomkins, A., Wiener, J., 2000. Graph structures in the web: experiments and models. Proceedings of the 9th International World Wide Web Conference, Amsterdam, The Netherlands, 309-320.

Caverlee, J., Webb, S., 2008. A large-scale study of myspace: observations and implications for online social networks. Proceedings of the 1st International Conference on Weblogs and Social Media, Seattle Washington, 36-44.

Clauset, A., Shalizi, C.R., Newman, M.E.J., 2009. Power-law distributions in empirical data, SIAM Review 51, 661-703.

Corlette, D., Shipman, F., 2009. Capturing on-line social network link dynamics using event-driven sampling. Proceedings of the12th IEEE International Conference. on

Computational Science and Engineering: Symposium on Social Intelligence and Networking, Vancouver, British Columbia, Canada, 284-291.

Corlette, D., Shipman, F., 2010. Link prediction applied to an open large-scale online social network. Proceedings of the 21st ACM Conference on Hypertext and Hypermedia, Toronto, Ontario, Canada, 135-140.

Ellison, N., Steinfeld, C., Lampe, C., 2007. The benefits of facebook 'friends': social capital and college students' use of online social network sites. Journal of Computer-Mediated Communication 12(4), article 1.

Fillmore, C., 1976. Frame semantics and the nature of language. Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech 280, 20-32.

Finn, A., Kushmerick, N., Smyth, B., 2002. Genre classification and domain transfer for information filtering. Proceedings of the 24th BCS-IRSG European Colloquium on IR Research: Advances in Information Retrieval, Glasgow, UK, 353-362.

Gaston, M., desJardins, M., 2005. Agent-oriented networks for dynamics team formation. Proceedings of the 9th International Conference on Autonomous Agents and Multi-agent Systems, Toronto, ON, Canada, 230-237.

Getoor, L., Diehl, C. P., 2005. Link mining: a survey. SIGKDD Explorations Newsletter 4(2), 3-12.

Google Freebase Data Dumps, 2010. Retrieved August 6, 2010 Available at: http://download.freebase.com/datadumps/

Grimes, C., Ford, D., Tassone, E., 2008. Keeping a search engine fresh: Risk and optimality in estimating refresh rates for web pages. Proceedings of the 40th Symposium on the Interface: Computing Science and Statistics, Durham, NC, USA, 1-14.

Hasan, M., A., Chaoji, V., Salem, S., Zaki, M., 2006. Link prediction using supervised learning. Proceedings of SIAM International Conference on Data Mining: Workshop on Link Analysis Counter-terrorism and Security, Bethesda, Maryland.

Hatzivassiloglou, V., Wiebe, J., 2000. Effects of adjective orientation and grade ability on sentence subjectivity. Proceedings of the 18th Conference on Computational Linguistics, Saarbrucken, Germany, 299-305.

Herring, S.C., Paolillo, J.C., Ramos-Vielba, I., Kouper, I., Wright, E., Stoerger, S., Scheidt, L.A., Clark, B., 2007. Language networks on livejournal. Proceedings of the

40th Hawai'i International Conference on System Sciences, Waikoloa, Big Island, Hawaii, USA, 79-90.

Hu, M., Liu, B., 2004. Mining opinion features in customer reviews. Association for the Advancement of Artificial Intelligence, San Jose, California, USA, 755-760.

Hu, M., Liu, B., 2004. Mining and summarizing customer reviews. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, 168-177.

Kashima, H., Abe, N., 2006. A parameterized probabilistic model of network evolution for supervised link prediction. Proceedings of the 6th International Conference on Data Mining, Hong Kong, China, 340-349.

Kim , S., 2004. Determining the sentiment of opinions, Proceedings of the 20th International Conference on Computational Linguistics, Genève, Switzerland, 1367-1373.

Kim, S., Hovy, E., 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. Proceedings of ACL the Workshop on Sentiment and Subjectivity in Text, Sydney Australia, 1-8.

Kim, S., Hovy, E., 2006. Automatic identification of pro and con reasons in online reviews. Proceedings of the COLING/ACL on Main Conference Poster Sessions, Sydney Australia, 483-490.

Kumar, R., Novak, J., Raghavan, P., Tomkins, R., 2004. Structure and evolution of blogspace. Communications of ACM 47(12), 35-39.

Kumar, R., Novak, J., Tomkins, A., 2006. Structure and evolution of online social networks. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 611-617.

Kunegis, J., Lommatzsch, A., 2009. Learning spectral graph transformation for link prediction. Proceedings of the 26th Annual International Conference on Machine Learning, Montreal, QC, Canada, 561-568.

Kushal, D., Lawrence, S., Pennock, D. M., 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. Proceedings of the 12th International World Wide Web Conference, Budapest, Hungary, 519-528.

Leskovec, L., Backstrom, L., Kumar, R., Tomkins, A., 2008. Microscopic evolution of social networks. Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV, USA, 462-470.

Li, L., Alderson, D., Doyle, C., Willinger, W., 2006. Towards a theory of scale-free graphs: definition, properties, and implications. Internet Mathematics 2(4), 431-523.

Li, X., Chen, H., 2009. Recommendation as link prediction: a graph kernel-based machine learning approach. Proceedings of the 9th Joint Conference on Digital Libraries, Austin, TX, USA, 213-216.

Liben-Nowell, D., Kleinberg, J., 2003. The link prediction problem for social networks. Proceedings of the 12th International Conference on Information and Knowledge Management, New Orleans, LA, USA, 556-559.

Lin, W., Wilson, T., Wiebe, J., Hauptmann, A., 2006. Which side are you on?: identifying perspectives at the document and sentence levels. Proceedings of 10th Conference on Natural Language Learning, New York City, USA, 109-116.

Liu, B., Hu, L., Cheng, J., 2005. Opinion observer: analyzing and comparing opinions on the Web. Proceedings of the 14th International Conference on World Wide Web, Chiba, Japan, 342-351.

Manning, C. D., Raghavan, P., Schütze, H., 2008. Introduction to Information Retrieval, Cambridge University Press, Cambridge, MA.

Manning, C., Schutze, H., 1999. Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA.

Martin, D., 2010, What Americans Do Online: Social Media and Games Dominate Activity, Retrieved: October 15, 2010, Available at: http://blog.nielsen.com/nielsenwire/online_mobile/what-americans-do-online-social-media-and-games-dominate-activity/

Miller, G.A., Beckwith, R., Fellbaum, C.D., Gross, D., Miller, K., 1990. Wordnet: an online lexical database. International Journal of Lexicograph 3(4), 235-244.

Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P. Bhattacharjee, B., 2007. Measurement and analysis of online social networks. Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, San Diego, CA, USA, 29-42.

Murata, T. and Moriyasu, S., 2007. Link prediction and social networks based on weighted proximity measures. Proceedings of the International Conference on Web Intelligence, Sydney, Australia, 85-88.

Nigam, K. Hurst, M., 2006. Towards a robust metric of polarity. In: Shanahan, J.G., Qu, Y., Wiebe, J. (Eds.), Computing Attitude and Affect in Text: Theory and Application. Springer, Dordrecht, The Netherlands, pp. 265-279.

O'Modadhain, J., Hutchins, J., Smyth, P., 2006. Prediction and ranking algorithms for event-based network data. SIGKDD Explorations 7(2), 23-30.

Palmer, M., Kingsbury, P., Gildea, D., 2005. The proposition bank: an annotated corpus of semantic roles. Computational Linguistics 31 (1), 71–106.

Pang, B., Lee, L., 2008. Opinion mining and sentiment analysis. Foundation and Trends in Information Retrieval 2(1-2), 1-135.

Rattigan, M., Jensen, D., 2005. The case for anomalous link discovery. SIGKDD Explorations 7(2), 41-47.

Scripts, J., Tan, P., Chen, F., Esfahanian, A., 2009. A matrix alignment approach for link prediction. Proceedings of The 19th International Conference on Pattern Recognition, Tampa, Florida, USA, 1-4.

Scripts, J., Tan, P., Esfahanian, A., 2009. Measuring the effects of preprocessing decision and network forces in dynamics network analysis. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Paris, France, 747-756.

Soon, W., Ng, H., Lim, D.C.Y., 2001. A machine learning approach to coreference resolution of noun phrases. Computational Linguistics 27(4), 521-544.

Stumpf, P.H., Carsten W., May, R.M., 2005. Subnets of scale-free networks are not scale-free: sampling properties of networks. Proceedings of the National Academy of Sciences 102(12), 4221-4224.

Turing, A.M., 1950. Computing machinery and intelligence. Mind 59, 433-460.

Twitter FAQ, 2009. Retrieved: January 15, 2009, Available at: http://apiwiki.twitter.com/FAQ

Tylenda, T., Angelova, R., and Bedathur, S., 2009. Towards time-aware link prediction in evolving social networks. Proceedings of the 13th International Conference on Knowledge Discovery and Data Mining , San Jose, California, USA, 1-10.

Wang, C., Satuluri, V., and Parthasarathy, S., 2007. Local probabilistic models for link prediction. Proceedings of the 7th International Conference on Data Mining, Omaha, Nebraska, USA, 322-331.

Wasserman, S., Faust, K., 1994. Social Network Analysis: Methods and Applications. Cambridge University Press, New York, NY.

Watts, D. J., Strogatz, S., 1998. Collective dynamics of small-world networks. Nature 393, 440-442.

Watts, D. J., 2003. Six Degrees. W. W. Norton & Company, New York, NY.

Weintraub, S., 2010, Schmidt: we'll pull facebook's data by hook or by crook. Retrieved: October 15, 2010, Available at: http://tech.fortune.cnn.com/2010/09/15/ schmidt-well-pull-facebooks-data-by-hook-or-by-crook/

Wilson, T., Wiebe, J., Hoffmann, P., 2005. Recognizing contextual polarity in phrase-level sentiment analysis. Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, Canada, 347-354.

Wortham, J., 2010, Facebook Tops 500 Million Users From New York Times. Retrieved: September 10, 2010, Available at: http://www.nytimes.com/2010/07/22/technology/22facebook.html

VITA

Name:	Daniel James Corlette
Address:	Department of Computer Science, Texas A&M University TAMU 3112 College Station, TX 77843-3112
Email Address:	dan@cse.tamu.edu
Education:	B.S., Computer Science, California State University Sacramento, 1999
	M.S. Computer Science, California State University Sacramento, 2003
	Ph.D., Computer Science, Texas A&M University, 2011