RICE UNIVERSITY

Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems

by

Alan E. Mislove

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Approved, Thesis Committee:

Peter Druschel, Chair
Professor of Computer Science

T. S. Eugene Ng
Assistant Professor of Computer Science

Krishna P. Gummadi
Assistant Professor of Computer Science

Houston, Texas

April, 2009

Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems

Alan E. Mislove

Abstract

Recently, online social networking sites have exploded in popularity. Numerous sites are dedicated to finding and maintaining contacts and to locating and sharing different types of content. Online social networks represent a new kind of information network that differs significantly from existing networks like the Web. For example, in the Web, hyperlinks between content form a graph that is used to organize, navigate, and rank information. The properties of the Web graph have been studied extensively, and have lead to useful algorithms such as PageRank. In contrast, few links exist between content in online social networks and instead, the links exist between content and users, and between users themselves. However, little is known in the research community about the properties of online social network graphs at scale, the factors that shape their structure, or the ways they can be leveraged in information systems.

In this thesis, we use novel measurement techniques to study online social networks at scale, and use the resulting insights to design innovative new information systems. First, we examine the structure and growth patterns of online social net-

works, focusing on how users are connecting to one another. We conduct the first large-scale measurement study of multiple online social networks at scale, capturing information about over 50 million users and 400 million links. Our analysis identifies a common structure across multiple networks, characterizes the underlying processes that are shaping the network structure, and exposes the rich community structure.

Second, we leverage our understanding of the properties of online social networks to design new information systems. Specifically, we build two distinct applications that leverage different properties of online social networks. We present and evaluate Ostra, a novel system for preventing unwanted communication that leverages the difficulty in establishing and maintaining relationships in social networks. We also present, deploy, and evaluate PeerSpective, a system for enhancing Web search using the natural community structure in social networks. Each of these systems has been evaluated on data from real online social networks or in a deployment with real users.

Acknowledgments

First and foremost, I would like to thank my advisors, Peter Druschel and Krishna P. Gummadi, for their help, advice, and mentoring during my graduate career. Without their support and guidance, none of the work presented in this thesis would have been possible. Moreover, I am deeply indebted to them both for showing me how to do successful research, how to mentor students, and how to communicate research results effectively. I suspect that this debt will only grow over time, as I use these skills in my own research career.

I would also like to thank Eugene Ng for his service on my thesis committee. His insight and advice proved very useful during the preparation of this thesis, and in my search for a tenure-track job. I am also grateful to have worked with Bobby Bhattacharjee – his advice and enthusiasm played no small part in my decision to continue a career in academia.

I am extremely grateful to have worked with and mentored numerous talented students during my research career. Working with Bimal, Malveeka, and Hema was a pleasure, and the excitement and energy they each brought to their research was both refreshing and invigorating. I hope that I am lucky enough to work with students of a similar caliber in the future.

I am deeply indebted to Brigitta Hansen, Claudia Richter, and Belia Martinez, whose assistance with many administrative matters proved invaluable. They all made living in Germany while finishing a Ph.D. at Rice a much easier experience.

I would also like to thank my colleagues and friends in Saarbrücken: Ansley, Animesh, Atul, Andreas, Jeff, Jim, Rodrigo, Andrey, Derek, Rose, Marcel, Max, Nuno, Pedro, Mia, and Ashu. They all made MPI-SWS a wonderful place to be, and being in Germany is an experience that I will always treasure.

I am also grateful for my close friendship with Rebecca. Her contagious excitement and enthusiasm was always refreshing, and I benefited greatly from her insight and advice. Additionally, I am grateful for my friendship with Stephanie – our travels and adventures often provided a needed break from research.

Finally, I would like to express my deep gratitude to my family, and especially my parents, for their love and support during the ups and downs of graduate school. I am grateful beyond words for all that they have given me.

Contents

Illustrations

Tables

Chapter 1

Introduction

Since its creation, the Internet has spawned many information sharing networks, the most well-known of which is the World Wide Web. Recently, a new class of information networks called "online social networks" have exploded in popularity and now rival the traditional Web in terms of usage [112]. Social networking sites such as MySpace (over 246 million users)[1], Facebook (over 124 million users), Orkut (over 67 million users), and LinkedIn (over 9 million "professionals") are examples of wildly popular networks used to find and organize contacts. Other social networks such as Flickr, YouTube, and Google Video, are used to share multimedia content, and others such as LiveJournal and BlogSpot are used to share blogs.

Unlike the traditional Web, which is largely organized by content, online social networks embody users as first-class entities. Users join a network, publish their own content, and create links to other users in the network called "friends". This basic user-to-user link structure facilitates online interaction by providing a mechanism for organizing both real-world and virtual contacts, for finding other users with similar interests, and for locating content and knowledge that has been contributed or

[1]The number of users refers to the number of identities as of November 2008 as published by each social networking site.

endorsed by "friends".

The extreme popularity and rapid growth of these online social networks represents a unique opportunity to study, understand, and leverage their properties. Not only can an in-depth understanding of online social network structure and growth aid in designing and evaluating current systems, it can lead to better designs of future online social network based systems and to a deeper understanding of the impact of online social networks on the Internet. Online social networks also offer many useful properties that can be leveraged to enhance information systems, such as enhancements to controlling information propagation, new directions for information search and retrieval, and new ways of reasoning about trust.

Thus, the goals of this thesis are two-fold, aiming to both understand the properties of online social networks and leverage those properties in information systems. We describe each of these in more detail below.

Our first goal is to understand the structure of online social networks, focusing on the social network graph that connects users. To this end, we conduct the first large-scale measurement study of online social networks, capturing information about users in multiple networks at scale. By examining more than one network, we can determine which structural features are unique to one network, and which are common across all networks. While the operators themselves obviously have complete data about their social networks, this data is not generally available to researchers due to competitive and privacy concerns. Thus, we chose to collect the data ourselves by querying the

public interface provided. Our collected data (which we have made available to the research community in anonymized form) represents the first large-scale data set available to researchers for many of these systems.

Our second goal is to apply our understanding of the properties of online social networks to build applications that leverage the information contained in social networks in innovative ways. In this thesis, we present two applications that address important challenges for information systems. The first challenge we address is the problem of unwanted communication, such as unsolicited marketing, propaganda, or spam. We demonstrate how to leverage the effort required to create and maintain social relationships in a social network to effectively block users from sending such communication without impeding legitimate communication. The second challenge we address is the problem of finding Web pages that are either new, not publicly visible, or of interest to only a small set of users. Due to the rapid growth of the Web, such pages are often not included in traditional Web search engines; we demonstrate how to leverage the shared interest between users in a social network to find such pages.

At a high level, this thesis is divided into four distinct parts: (a) discussions of background, related work, and data collection methodology; (b) detailed studies on the structure and growth of online social networks; (c) an examination of the communities in online social networks; and (d) the presentation of two new systems that leverage properties of online social networks. In the next few subsections, we

describe each part in detail.

1.1 Background, related work, and methodology

This thesis begins with a brief history of online social networks, a discussion of research related to this thesis, and a description of our data collection methodology. In Chapter 2, we provide background on online social networks and motivate this thesis. We describe the rapid rise in popularity of these networks and catalog the various mechanisms that today's online social networks provide for information sharing. In Chapter 3 we detail related work from computer science, sociology, graph theory, and theoretical physics. We describe previous approaches to studying the structure of social networks, the various data sets that have been collected so far, and applications that been built on top of social networks.

In Chapter 4 we describe the data sets we have collected for study. We obtained data from six different online social networks covering over 50 million users connected together by over 400 million links. We describe the procedures used to collect these data sets and discuss ways in which our collection methodology limits our analysis. We also describe which of the data sets we have made publicly available and provide instructions for accessing them.

1.2 Network structure and growth

To begin our analysis, we first focus on the graph formed by users in online social networks. In Chapter 5, we present an analysis of the structure of four popular online social networks: Flickr, YouTube, LiveJournal, and Orkut. Ours is the first study to examine multiple online social networks at scale; in contrast, previous studies have generally relied on proprietary data obtained from the operators of a single large network. Data gathered from multiple sites enables our analysis to identify common structural properties of online social networks.

This analysis allows us to validate the power-law, small-world and scale-free properties previously observed in offline social networks, as well as to provide new insights into the properties of the social network graphs. We observe a high degree of reciprocity in directed user links, leading to a strong correlation between user indegree and outdegree. This differs from content graphs such as the graph formed by Web hyperlinks, where the popular pages (*authorities*) and the pages with many references (*hubs*) are distinct [74]. Our analysis also shows that online social networks contain a large, strongly connected core of high-degree nodes, surrounded by many small clusters of low-degree nodes. This suggests that high-degree nodes in the core are critical for connectivity and information flow in these networks.

We observe an intriguing similarity between the structure of multiple networks, despite different mechanisms, policies, and conventions for creating links. This suggests that links are created in a similar manner across all of the networks. Thus, in

Chapter 6, we use empirical data to understand the growth processes that lead to the observed network structure. Our analysis of large-scale growth data shows that new links are created and received by users in direct proportion to their current number of links, and that users tend to quickly respond to incoming links by creating a link in the reverse direction. Additionally, our analysis reveals a strong proximity bias when users select other users to link to: users tend to connect to nearby users in the network much more often than would be expected from previously proposed growth models.

Our work on studying network growth is an important first step towards understanding the processes that shape the structure of online social networks. Our work enables the creation of synthetic networks that reflect both global and local characteristics of online social networks. Moreover, our collected data may lead to more accurate structural and growth models, which are useful for network analysis and planning. Such models can be used in the design of search algorithms (e.g., by pre-identifying users that are likely to be hubs), in data mining (e.g., by identifying candidate users to monitor), and in system evaluation (e.g., by allowing networks to be simulated over a wide range of sizes).

1.3 Communities in online social networks

Next, in Chapter 7, we focus on how users form communities in online social networks.[2]. Communities are interesting for a variety of reasons. For example, users in a community tend to interact frequently, often share interests, and trust each other to some extent. Therefore, communities are useful, for instance, for guiding information dissemination and acquisition, in recommending or introducing people who would likely benefit from direct interaction, and in expressing access control policies. Many algorithms for automatically detecting communities in social networks have been proposed [14, 31, 58, 99, 118, 131, 153, 160]. However, these algorithms have never been tested over real online social networks at scale.

We use detailed data from an online social network to study the effectiveness of existing approaches for detecting communities. We collect detailed data about the members of a university in the Facebook social network [49] and analyze the structure of communities in our data. We find that users are often members of multiple overlapping communities. We then examine whether these multiple communities can be automatically detected. Most existing algorithms have only been evaluated on non-social networks, and we find that they do not perform well in detecting the multiple overlapping communities that exist in current social networks.

We propose and evaluate a new algorithm that can infer memberships of multiple,

[2]A community is a subset of the users in a social network that is more tightly interconnected than the overall network [119].

potentially overlapping communities, when given information about a small subset of the community members. The algorithm uses the ratio between the number of links within a community and number of links between the community and the rest of the network. We demonstrate that the algorithm works well in practice: even if community membership information is only known for as few as 20% of the users, the remaining members of the community can be determined with high accuracy.

1.4 Ostra: Leveraging relationships

Finally, we present systems that leverage social networks to solve open systems challenges. In Chapter 8, we present a system that exploits the the difficulty in establishing and maintaining relationships in social networks to address the problem of unwanted communication. Internet-based communication systems such as email, instant messaging (IM), voice-over-IP (VoIP), online social networks, and content-sharing sites allow communication at near zero marginal cost to users. Any user with an inexpensive Internet connection has the potential to reach millions of users. This property has democratized content publication: anyone can publish content, and anyone interested in the content can obtain it. Unfortunately, the same property can be abused for the purpose of unsolicited marketing, propaganda, or disruption of legitimate communication.

We describe a method that exploits existing relationships among users in an online social network to impose a cost on the senders of unwanted communication. Our

system, Ostra, (i) relies on existing social networks to connect senders and receivers via chains of pairwise relationships; (ii) uses a pairwise, link-based credit scheme that imposes a cost on senders of unwanted communication without requiring sender authentication or global identities; and (iii) relies on feedback from receivers to classify unwanted communication. Ostra ensures that unwanted communication strains the sender's relationships, even if the sender has no direct relationship with the ultimate recipient of the communication. A user who continues to send unwanted communication will eventually lose the ability to communicate. An evaluation of Ostra on traces from an online social network demonstrate that it can effectively block unwanted communication.

1.5 Peerspective: Leveraging shared interest

In Chapter 9, we demonstrate how to leverage communities in online social networks to help users find interesting content. Users increasingly share content, recommendations, opinions, and ratings using online social networks. However, the growing number of users and the increasing variety and volume of shared information on these sites aggravates two fundamental problems in information sharing: *privacy* and *relevance*. Since users are often sharing personal information, privacy and access control is critical. Additionally, since the volume of shared content is growing at an enormous rate, finding relevant information is becoming increasingly difficult. We argue that communities are an important concept that can offer a solution to this

growing dilemma.

Most online social networks today allow only very coarse-grained content sharing policies: users typically have the options of sharing content with (subsets of) their direct friends or with everyone. Communities can provide a natural middle ground, allowing convenient sharing among groups of users who do not necessarily know each other but who are close together in the social network. Also, communities often represent sets of users with common interests, a fact that can be naturally leveraged by systems to provide information that is relevant at a local, rather than global, scope.

Using empirical data from an online social network and from a system deployment, we demonstrate the potential for using communities in online social networks. We have built and deployed PeerSpective, a system that leverages communities in a social network in order to aid Web search. PeerSpective automatically indexes browsed pages and transparently inserts relevant pages viewed by friends into Web search results. The results are aggregated over the community and presented alongside the normal search results. Using data from a PeerSpective deployment, we demonstrate that communities represent groups of users with shared interests, and that PeerSpective provides a measurable improvement to Web search.

Finally, Chapter 10 presents concluding remarks, discusses the implications of our work, and describes future research directions.

Chapter 2

Background

In this chapter, we first give an overview of online social networks, describing their characteristics, the reasons behind the growth in their popularity, and the range of user interactions they allow. Then, we describe applications of online social networks, motivating why understanding their structure and properties is a necessary step to building future applications. Finally, we provide background on metrics for the analysis for complex graphs.

2.1 What are online social networks?

We begin by defining online social networks, providing a brief history of their growth in popularity, and detail the mechanisms that today's online social networks provide for users to connect and share content.

2.1.1 Definition and purpose

For the purposes of this thesis, we define an online social network to be a system where (a) users are first class entities with a semi-public profile, (b) users can create explicit links to other users or content items, and (c) users can navigate the social network by browsing the links and profiles of other users. This definition is consistent

with that used in previous studies [38].

Online social networks serve a number of purposes, but three primary roles stand out as common across all sites. First, online social networks are used to maintain and strengthen existing social ties, or make new social connections. The sites allow users to "articulate and make visible their social networks", thereby "communicating with people who are already a part of their extended social network" [38]. Second, online social networks are used by each member to upload her own content. Note that the content shared often varies from site to site, and sometimes is only the user's profile itself. Third, online social networks are used to find new, interesting content by filtering, recommending, and organizing the content uploaded by users.

2.1.2 A brief history

We now give a brief history of online social networks. The site Classmates.com [30] is regarded as the first web site that allowed users to connect to other users. It began in 1995 as a site for users to reconnect with previous classmates and currently it has over 40 million registered users. However, Classmates.com did not allow users to create links to other users; rather, it allowed users to link to each other only via schools they had attended. In 1997, the site SixDegrees.com [145] was created, which was the first social networking site that allowed users to create links directly to other users. As such, SixDegrees.com is the first site that meets the definition of an online social network from above.

Online social networks began to grow in popularity as more users became connected to the Internet. In the early 2000s, a number of general-purpose sites for finding friends were established, the most notable of which is Friendster. Friendster was focused on allowing friends-of-friends to meet, beginning as a rival to the online dating site Match.com. Other, similar sites created in the same timeframe include CyWorld [34], Ryze [140], and LinkedIn [95].

In 2003, MySpace [111] was created as an alternative to Friendster and the others. MySpace allowed users to heavily customize the appearance of their profile, which proved very popular with users, causing MySpace to quickly become the largest online social network. As of this writing, MySpace has 247 million user accounts, over twice as many as the second most popular network, Facebook. For a more complete history and analysis of the evolution of online social networks, we refer the reader to the numerous papers by boyd [35, 36, 38, 39].

With the rise in popularity of online social networks, many other types of sites began to include social networking features. Examples include multimedia content sharing sites (Flickr [52], YouTube [167], and Zoomr [174]), blogging sites (LiveJournal [97] and BlogSpot [20]), professional networking sites (LinkedIn [95] and Ryze [140]), and news aggregation sites (Digg [41], Reddit [132], and del.icio.us [40]). All of these sites have different goals but employ the common strategy of exploiting the social network to improve their sites. The list above is not meant to be exhaustive, as new sites are being created regularly. For a more complete and up-to-date list

of the notable online social networking sites, we refer the reader to Wikipedia [96].

The sociological aspects behind the rapid growth and adoption of social networking sites are also the subject of much scholarship. One of the primary reasons that has been noted for popularity of social networking sites is their user-centric nature. The content that is shared on social networking sites is often information about the users themselves, such as their status, photos, and so forth. For more details, we refer the reader to the work by boyd [37].

2.1.3 Mechanisms and policies

We now give a brief overview of the mechanisms and policies that most online social networks provide.

Users

Full participation in online social networks requires users to register a (pseudo) identity[1] with the network, though some sites do allow browsing public data without explicit sign-on. Users may volunteer information about themselves (e.g., their birthday, place of residence, interests, etc.), all of which constitutes the user's *profile*.

The social network itself is composed of links between users. Some sites allow users to link to any other user (without consent from the link recipient), while other

[1]In the rest of this thesis, we use the term "user" to denote a single unique identity in a social network. Clearly, a single human may create multiple identities, and may even create links between their own identities. We consider each of these identities as separate users.

sites follow a two-phase procedure that only allows a link to be established when both parties agree. Certain sites, such as Flickr, have social networks with directed links (meaning a link from A to B does not imply the presence of a reverse link), whereas others, such as Orkut, have social networks with undirected links.

Users link to other users for numerous reasons. The target of a link may be a real-world acquaintance, a business contact, a virtual acquaintance, someone who shares the same interests, someone who uploads interesting content, and so on. In fact, some users even consider the acquisition of many links to be a goal in itself [36]. When compared to links in the Web, links in online social networks combine the functionality of both hyperlinks and bookmarks.

A user's links, along with her profile, are usually visible to those who visit the user's account. Thus, users are able to navigate the social network by following user-to-user links, browsing the profile information and any contributed content of visited users as they go. Certain sites, including LinkedIn, only allow browsing of profiles within the user's own neighborhood (i.e., a user can only view other users that are within two hops), while other sites, such as Flickr, allow users to view any other user in the system.

Groups

Most sites also enable users to create special interest *groups*, which are akin to Usenet [127] newsgroups. Users can post messages to groups (visible to all group

members) and even upload shared content to the group. Certain groups are *moderated*, and admission to the group is controlled by a single group maintainer, while other groups are open for any member to join. All sites today require explicit group declaration by users; users must manually create groups, appoint administrators (if necessary), and declare which groups they are a member of. Certain sites (such as Facebook) create a few pre-populated groups based on the domain of users' email addresses, but the majority of groups do not fall into this category.

The primary use of groups in today's networks is to either express access control policies or to provide a forum for shared content. Examples of the former include sites like Facebook, which, by default, allows only users located in the same geographic location or organization to view each other's profiles. Examples of the latter are more common, including Flickr's shared photo groups and Orkut's communities feature.

Content

Once an identity is created, users of content-sharing sites can upload content onto their account. Many such sites enable users to mark content as public (visible to anyone) or private (visible only to their immediate "friends"), and to tag content with labels. Many sites, such as YouTube, allow users to upload an unlimited amount of content, while other sites, such as Flickr, require that users either pay a subscription fee or be subject to an upload limit. All of the content uploaded by a given user is listed in the user's profile, allowing other users to browse through the social network

to discover new content. Typically, the content is automatically indexed, and, if publicly available, made accessible via a textual search. An example is Flickr's photo search, which allows users to locate photos by searching based on tags and comments.

Once on the site, users can submit their uploaded content into groups that they are a member of. The privacy settings often allow for the content to be accessible only by group members. Moreover, the sites generally allow users to browse the content uploaded to groups they are members of.

Users are also often allowed to create *favorite* lists, which link to a user's favorite content uploaded by other users. These favorite lists are also generally publicly accessible from the user's profile page. Similarly, most sites allow users to *comment* on pieces of content, much like a Usenet posting, and the comments appear alongside the piece of content itself.

Finally, many sites contain *most popular* content lists, which contain the most popular content items (in terms of the number of views, comments, or ratings) that have been recently uploaded. Users can browse these lists to find new content to view. A notable example is YouTube's top-100 lists, where popularity is based on the number of views, comments, or favorite-markings a video has recently received.

2.1.4 A new form of information exchange

To underscore how online social networks represent different information distribution systems relative to systems like the Web, we focus briefly in this section on how con-

tent is spread in today's networks. Most of the sites we study are designed for sharing content: Flickr, YouTube, and LiveJournal are used for publishing, organizing, locating, and distributing photos, videos, and blogs, respectively.

To investigate the role played by the underlying user network in organizing and locating content, we conducted a simple measurement of how users browse the Flickr system. We analyzed the HTTP requests going to the `flickr.com` domain from an HTTP trace taken at the border routers of Technical University of Munich between August 17th, 2006 and October 11th, 2006. We found 22,215 photo views from at least 1,056 distinct users. For each of these views, we examined the browser's clickstream to determine what action led the user to a given photo.

We found that 17,897 of these views (80.6%) resulted from following links in the Flickr user graph or from following links between photos within a user's collection. In other words, 80.6% of the time, the social network of Flickr users was used in browsing content. We count these views as being influenced by the social network. Of the remaining, 1,418 (6.3%) views involved Flickr search facilities. Finally, only 2,900 (13.1%) views followed a link from an external source, such as links from an external Web site or links received via email. Neither of the latter sets of views (19.4%) involved the social network.

Thus, our experiment demonstrates that the social network in Flickr plays an important role in locating content: four out of five photos were located by traversing the social network links.

2.2 Why study online social networks?

Online social networking is still very much in its infancy, yet it already forms the basis for some enormously popular applications. As this paradigm matures, we expect more sophisticated applications to naturally emerge. It is not inconceivable that social networking systems will eventually become de-facto portals for both personal and commercial online interactions. Below, we outline a few of the many potential applications that could benefit from understanding the structure of and information flow in these networks. Additionally, we speculate on how the data collected in this thesis could be relevant to researchers in other disciplines.

2.2.1 Trust

Adjacent users in a social network tend to trust each other more than random pairs of users in the network. A number of research systems have already been proposed to exploit this trust. Trust relationships are being used in the PGP web of trust [172] to eliminate the need for a trusted certificate authority. SybilGuard [169] and SybilLimit [168] uses the social network to mitigate Sybil [44] attacks in distributed systems, exploiting the fact that real people tend to have a diverse set of social relations. Re [57] determines the social network distance between the sender and the receiver of an email to aid SPAM detection. We believe that a deeper understanding of the underlying topology is an essential first step in the design and analysis of robust trust and reputation metrics for these systems.

2.2.2 Shared interest

Adjacent users in a social network also tend to share common interests. Users browse neighboring regions of their social network because they are likely to find content that is of interest to them. Systems such as Yahoo My Web [165], Google Co-op [60], and PeerSpective [104] use social networks to rank Internet search results relative to the interests of a user's social network. Using the content viewed and search results clicked on by members of a social network, these systems to rank the results of the members' future searches more accurately.

Clearly, understanding the structure of online social networks, as well as the processes that shape them, is important for these applications. For example, efficient algorithms are needed for inferring the actual degree of shared interest between two users, or the reliability of a user (as perceived by other users). It is also important to understand the robustness of such networks to deliberate attempts of manipulation. These topics are beyond the scope of this thesis; however, a fundamental understanding of online social network structure is likely to be a necessary first step.

2.2.3 Content exchange

The phenomenal popularity of social networking sites like YouTube, Flickr, and MySpace represents a shift in how content is published, located, and distributed on the Internet. Understanding how content diffuses through these networks and becomes popular over time is not only of academic interest, but is increasingly important in

commercial advertising, in political campaigning, and ultimately to society. In fact, a number of research efforts [42, 43, 66, 72, 137, 158] have proposed viral marketing campaigns to leverage the word-of-mouth effect. In 2007 alone, $1.2 billion was spent on advertisement in online social networks worldwide, and this is expected to triple by 2011 [147]. Understanding how information flows among users of online communities is an important step toward the design and analysis of future information dissemination systems.

Understanding how information flows in online social networks can also aid designers of current social networking systems. If, for example, one can predict the relative popularity of newly introduced objects, caching and pre-fetching schemes can be created to reduce the latency and bandwidth required by the site. Since many of the currently popular sites rely primarily on advertising for revenue, reducing distribution costs for multimedia content is clearly a pressing issue.

Understanding how content flows through social networks also has the potential to improve search algorithms. By examining the content that users view or mark as a favorite, sites may be able to suggest other content that may be of interest to the user. Many have noted [10] that the age of the Internet has enabled much greater diversity in preferences and tastes; using online social networks appears to be a natural approach to further discover and refine tastes.

Finally, understanding how content is exchanged in online social networks can help guide the designers of future systems. Social networks have already proven to be

useful in a number of different contexts, and we are seeing new sites popping using social networks to predict music preferences, find potential job applications, and share content. By understanding the user structure and the properties of information flow, designers of future systems have a empirical basis for designing and provisioning their systems.

2.2.4 Other disciplines

As mentioned before, our work has relevance beyond computer systems. To sociologists, online social networks offer an unprecedented amount of data. These systems represent the complete evolution of a large, contained online social network, with the accompanying timeline of every event that occurred within them. Sociologists can examine this data to validate existing theories of communication, as well as to look for new forms of communication.

To political scientists and marketing specialists, studying how information flows through social networks may help improve techniques such as targeted advertising and viral marketing. Political candidates have already realized the importance of blogs in recent elections [133]. Similarly, marketing specialists are already experimenting with paid viral marketing [125] to better promote products and companies. Clearly, a better understanding of how content is currently being exchanged in these systems holds the potential to improve these approaches.

2.3 How do we analyze complex networks?

We now discuss the various ways of analyzing and characterizing the shape of large networks, and conclude with a discussion of the various classes of graphs that have been observed in the real world.

2.3.1 Preliminaries

We assume that we have a network which can be viewed as a graph $G = (V, E)$. In the context of an online social network, for example, the vertices represent users and the edges represent relationships among users. The links in the graph can either be *directed*, meaning each link is sourced at one node and terminated at another node, or *undirected*, meaning each link is between two nodes without a source and destination.

Consistent with previous work, we define a node i's *degree*, denoted by d_i, to be the number of links the node has to other nodes. For directed networks, we distinguish between *indegree* (the number of incoming links) and *outdegree* (the number of outgoing links). Also for directed networks, we consider the level of *symmetry* in the network to be the fraction of links that have a corresponding reverse link.

2.3.2 Radius and diameter

We now discuss the radius and diameter of a graph, which represents how far away nodes are from each other in the network. First, the *eccentricity* of a node v is the maximal shortest path distance between v and any other node. The *radius* of a graph

is then the minimum eccentricity across all vertices, and the *diameter* is the maximum eccentricity across all vertices. Thus, the radius represents the maximal distance from the most "central" node in the graph to all other nodes, and the diameter represents the maximal distance from the least "central" node in the graph to all other nodes.

Due to the computational complexity associated with determining the actual radius and diameter, the radius and diameter of a graph is often estimated by calculating the eccentricity of a large random sample of nodes in the network. In such cases, the diameter should be viewed as a lower bound of the true diameter, and the radius as an upper bound of the true radius.

2.3.3 Degree distribution

The degree distribution of a graph is a function $P(k)$ which describes the fraction of the network's nodes which have degree k. The degree distribution describes how the links in the graph are distributed among the nodes. For example, the degree distribution of a graph with randomly placed edges among n nodes follows a binomial distribution of

$$P(k) = \binom{n-1}{k} p^k (1-p)^{n-1-k} \tag{2.1}$$

where p represents the probability that any two nodes are connected. Most real-world networks have been shown to deviate from random graphs, and instead, have a bias whereby a few high-degree nodes hold a large fraction of the links.

2.3.4 Joint degree distribution

In addition to the degree distribution $P(k)$, we also focus on the *joint degree distribution* (JDD) represented as $J(k, m)$. The function $J(k, m)$ represents what fraction of the links in the graph are between nodes of degree k and degree m. In the case of a directed network, $J(k, m)$ represents the fraction of links that are from a node with outdegree k and to a node with indegree m. Thus, the JDD represents how often nodes of different degrees connect to each other. This property is also referred to as the 2K-distribution [101] or the mixing patterns [116].

The JDD provides many insights into the structural properties of networks. For example, networks where high-degree nodes tend to connect to other high-degree nodes are more likely to be subject to epidemics, as a single infected high-degree node will quickly infect other high-degree nodes. On the other hand, networks where high-degree nodes tend to connect to low-degree nodes show the opposite behavior; a single infected high-degree node will not spread an epidemic very fast.

The JDD can be approximated by the degree correlation function k_{nn}, which is a mapping between outdegree and the average indegree of all nodes connected to nodes of that outdegree. Clearly, an increasing k_{nn} indicates a tendency of higher-degree nodes to connect to other high-degree nodes; a decreasing k_{nn} represents the opposite trend.

2.3.5 Scale-free behavior

The scale-free metric $s(G)$ [91] of a graph is a value calculated directly from the joint degree distribution of a graph. The scale-free metric ranges between 0 and 1, and measures the extent to which the graph has a hub-like core. To define the $s(G)$, we first define $s'(G)$ as

$$s'(G) = \sum_{(i,j)\in E} d_i d_j \tag{2.2}$$

Then, we define the scale-free metric $s(G)$ as

$$s(G) = \frac{s'(G)}{s'_{max}} \tag{2.3}$$

where s'_{max} represents the maximum value of s' over all graphs with the same degree distribution of G. A high scale-free metric means that high-degree nodes tend to connect to other high-degree nodes, while a low scale-free metric means that high-degree nodes tend to connect to low-degree nodes.

2.3.6 Assortativity

The scale-free metric is related to the assortativity coefficient r, which is a measure of the likelihood for nodes to connect to other nodes with similar degrees. The assortativity is defined as the Pearson correlation coefficient between the degrees of all pairs of nodes connected by an edge. Thus, the assortativity coefficient ranges between -1 and 1; a high assortativity coefficient means that nodes tend to connect to nodes of similar degree, while a negative coefficient means that nodes likely connect to nodes with very different degree from their own.

2.3.7 Clustering coefficient

The *clustering coefficient* of a node i, denoted by $c(i)$, is defined as the number of directed links that exist between the node's neighbors, divided by the number of possible directed links that could exist between the node's neighbors. Thus, if a node i's neighbors have n directed links between them, then the clustering coefficient of i is defined as

$$c(i) = \frac{n}{d_i(d_i - 1)} \tag{2.4}$$

The clustering coefficient of a graph is the average clustering coefficient of all its nodes, and we denote it as $C(G)$, or

$$C(G) = \frac{\sum_{v \in V} c(v)}{|V|} \tag{2.5}$$

Thus, the clustering coefficient of a graph ranges between 0 and 1, with higher values representing a higher degree of "cliquishness" between the nodes. In particular, a graph with clustering coefficient of 0 contains no "triangles" of connected nodes, whereas a graph with clustering coefficient of 1 is a perfect clique.

2.3.8 Betweenness centrality

The *betweenness centrality* B of an edge, originally proposed by Girvan and Newman [119], is defined as the number of shortest paths between all pairs of vertices in the graph that cross the edge. If a pair of vertices have multiple shortest paths between them, then each path is assigned a weight such that the sum over all paths

is one. Thus, betweenness centrality for an edge e can be expressed as

$$B(e) = \sum_{u \in V, v \in V} \frac{\sigma_e(u,v)}{\sigma(u,v)} \tag{2.6}$$

where $\sigma(u,v)$ represents the number of shortest paths between u and v, and $\sigma_e(u,v)$ represents the number of shortest paths between u and v that include e. The betweenness centrality of an edge can be viewed as a metric for the importance of an edge in a graph, as edges with a higher betweenness centrality fall on more shortest paths, and are therefore more important for the structure of the graph.

2.3.9 Modularity

When examining communities in networks, one often requires an objective metric to evaluate how "good" a particular division of the network into communities is. One such metric is the the *modularity* measure proposed by Newman [118]. Consider a community structure of k communities. Let $\mathbf{e}$ be a symmetric $k \times k$ matrix, whose element e_{ij} is the fraction of edges in the network that connect vertices in community i to community j by considering all the edges in the original network. Also, we define $a_i = \sum_j e_{ij}$ be the fraction of edges that touch vertices in community i. Then, the trace of the matrix $\text{Tr } \mathbf{e} = \sum_i e_{ii}$ gives the fraction of edges in the network within the same community. Hence, modularity is defined as

$$Q = \sum_i (e_{ii} - a_i^2) = \text{Tr } \mathbf{e} - ||\mathbf{e}^2|| \tag{2.7}$$

where $||\mathbf{y}||$ indicates the sum of elements of matrix $\mathbf{y}$. Modularity is then a measure of the fraction of intra-community edges minus the expected value of the same quan-

tity in a network with the same community divisions, but with edges placed without regard for communities. Modularity therefore ranges from -1 to 1, with 0 representing no more community structure than would be expected in a random graph, and significantly positive values representing the presence of community structure. In practice, a modularity over 0.3 or higher is observed in real-world networks with significant community structure [118].

2.3.10 Connected components

Finally, we discuss the notion of connected components in graphs. For an undirected graph, a *connected component* as a subset of the nodes such that there is a path in the network between all pairs of nodes in the set. For a directed graph, we distinguish between a *strongly connected component* and a *weakly connected component*. A strongly connected component (SCC) is defined as a set of nodes such that there is a path in the network between all pairs of nodes in the set. In contrast, a weakly connected component (WCC) is defined as a set of nodes such that there is a path in the network between all pairs of nodes in set if the all links in the network were viewed as undirected.

Studies of real-world networks, such as the Web and the Internet topology, has shown that there often exists a single, dominating SCC which is orders of magnitude larger than all other SCCs [23]. In this thesis, we refer to this dominating component as the dominant SCC .

2.3.11 Classes of studied networks

Now, we detail classes of complex networks that have been observed in the real world. For more detail on these networks, we refer the reader to the survey by Newman [117].

Random networks have been heavily studied, starting with the seminal paper by Erdös and Réyni [48]. These graphs are usually constructed by randomly adding links to a static set of nodes. Researchers have shown that random graphs tend to exhibit very short average path lengths between any two nodes [78]. More recent work on random graphs has provided mechanisms to construct graphs with specified degree distributions [108] and characterized the size of the large strongly connected component [109].

Power-law networks are networks where the probability that a node has degree k is proportional to $k^{-\alpha}$, for large k and $\alpha > 1$. Thus, the degree distribution of a power-law network follows an exponential decay. The parameter α is called the *power-law coefficient*. Researchers have shown many real-world networks are power-law networks, including Internet topologies [51], the Web [15, 83], social networks [4], neural networks [22], and power grids [128].

Scale-free networks are a class of power-law networks where the high-degree nodes tend to be connected to other high-degree nodes. Scale-free graphs are discussed in detail by Li et al. [91], and are defined as networks with a significant scale-free metric $s(G)$.

Small-world networks have a small diameter and exhibit a high clustering coefficient. Studies have shown that the Web [7, 23], scientific collaboration [115], film actors [9], and general social networks [4] have small-world properties. Kleinberg [76] proposed a model to explain the small-world phenomenon in social networks, and also examined navigability in these networks [75]. The online social networks examined in this thesis demonstrate small-world properties much like their real-world counterparts.

2.3.12 Preferential attachment

Preferential attachment [15], also known as *cumulative advantage* [170] or the *rich get richer phenomenon*, is a property of link formation in a graph. In short, preferential attachment says that the likelihood of a node being attached to a new link is in proportion to the node's degree. Preferential attachment in a given network can be characterized as linear, if the probability of a node receiving a link is in linear proportion to the node's degree, or sub-linear, if the probability of a node receiving a link is, for example, in proportion to the log of the node's degree. Under certain circumstances, preferential attachment has been shown to result in power-law networks [15].

Chapter 3

Related Work

In this chapter, we describe prior work related to the topics presented in this thesis. As this thesis covers a number of different topics, the related work has been grouped into sections detailing (a) work that examines the static structural properties of complex networks, (b) work that examines how complex networks evolve, (c) work that identifies and uses communities in online social networks, (d) work that tries to solve the problem of unwanted communication, and (e) work that tries to personalize web search.

3.1 Complex network structure

We begin by examining work that characterizes the structure of static snapshots of large scale networks. In following chapters, we examine the static snapshots of multiple online social networks. In order to ground our analysis, we compare our results to those from other large-scale complex networks such as the Web and the Internet. Thus, we describe related work that studies these networks after describing work that studies social networks.

3.1.1 Social networks

Sociologists have studied many of the properties of offline social networks, and we only briefly describe a few of the relevant findings. For a more complete overview of offline social networks and associated analysis techniques, we refer the reader to the book by Wasserman [157]. Milgram [102] showed that the average path length between two Americans was six hops, demonstrating that social networks can be classified as small-world. Pool and Kochen [129] provided an analysis of how the small-world property of social networks affects contacts and influence. The influential paper by Granovetter [63] argued that a social network can be partitioned into 'strong' and 'weak' ties, and that the strong ties are tightly clustered, while the weak ties represent longer-distance relationships. We were able to verify that online social networks have similar properties, with short path lengths and strong clusters connected by long-distance links.

As online social networks gained popularity, researchers have begun to investigate their properties. Adamic et al. [4] studied an early online social network at Stanford University, and found that the network has small-world characteristics as well as a significant clustering coefficient. Liben-Nowell et al. [94] found a strong correlation between friendship and the geographic location of users by using data from LiveJournal. Kumar et al. [82] examined two online social networks from Yahoo! and found that both possessed a dominant SCC. Girvan and Newman observed that users in online social networks tend to form tightly knit groups [58], evidenced by a high

clustering coefficient. We were able to verify all of these properties on multiple sites and on a much larger scale in our study.

In more recent work, Ahn et al. [6] analyzed complete data from the large South Korean social networking site Cyworld [34], along with data from small sample crawls of MySpace and Orkut. The authors obtained data directly from CyWorld operators, and the volume of available data allows the authors to conduct an in-depth study of that site using some of the same metrics that we use in this thesis. The comparison with different networks, on the other hand, is limited by the small crawled data samples of MySpace and Orkut. Our study is largely complementary: the data available to us for any one site is less detailed, but we are able to compare large crawled data sets from multiple sites.

Finally, researchers have also examined how the activity network, or the pattern of interactions between users, compares with the social network. In particular, Wilson et al. [163] studied the activity network of samples of the Facebook network and found that, in contrast to the social network, the activity network is much more sparse and has a significantly lower maximal degree. Chun et al. [29] found similar properties for the interaction network in CyWorld. In our work, we focus only on the social network, but our approach and methods could be naturally applied to the activity network as well.

3.1.2 Other information networks

A long thread of research examines the structure of information networks like the graph of Web pages and the Internet's routing topology. A prominent study of Web structure [23] showed that the Web has a "bow-tie" shape, consisting of a dominant SCC, and groups of nodes that can either reach the SCC or can be reached from the SCC. We show that the social networks share a similar dominant SCC, but that this component is relatively much larger than that of the Web. Faloutsos et al. [51] found that the degree distribution of the Internet's routing topology follows a power-law, and Siganos et al. [144] demonstrated that the high-level structure of the Internet resembles a "jellyfish".

Kleinberg [77] showed that high-degree nodes can be observed in the Web that function either as hubs (pages containing useful references on a subject) or authorities (pages containing relevant information on a subject). Kleinberg also presented an algorithm [74], which, when given a graph of Web pages, can infer pages function as hubs and as authorities. The well-known PageRank algorithm [122] uses the Web structure to determine pages that are considered reputable. Our results indicate that in online social networks, the high degree of link symmetry may prevent such algorithms from working, since the hubs are automatically also the authorities.

3.2 Complex network growth

In addition to the study of the static structure of various information networks, researchers have also examined the evolution of networks, looking at the processes by which links are formed and removed. Consistent with previous work, we refer to these processes as *growth models*. In our work, we collect detailed data on the growth of online social networks. Thus, in this section, we describe related work on various growth models, and detail the extent to which they have been validated on real data.

3.2.1 Growth models

Growth models for complex networks can be partitioned into structural models (i.e., models that only take into account the structure of the network to predict link formation or removal) and explanatory models (i.e., models that consider external factors, such as human factors in online social networks, to predict links). We describe each of these types of models below.

Structural growth models

Researchers sought to explain the intriguing similarity in the high-level structural properties across networks of very different scales and types by hypothesizing that the networks are the result of a few common structural growth processes at work. Many models of these processes have been proposed and analyzed to explain the structure of complex networks.

The well-known Barabási-Albert (BA) model [15], based on preferential attachment, has been shown to result in networks with power-law degree distributions. In the BA model, new links are attached to nodes using a probability distribution weighted by node degree, resulting in linear preferential attachment. Many extensions to the BA model have been proposed (e.g., to add a tunable level of clustering [67]). We are able to verify that the growth of online social networks follows linear preferential attachment, but not in the way that the BA model proposes.

Another class of models that produce power-law networks are based on local rules, such as the random walk model [142,154], where nodes select new neighbors by taking random walks; the common neighbors model [114], where nodes select new neighbors by picking nodes with whom they share many friends in common; and the finite memory model [79], where nodes eventually become inactive and stop receiving any new links. All of these models exhibit preferential attachment (since high-degree nodes end up being selected more often), but with higher levels of local clustering than the BA model [154]. We demonstrate that, while these models are more accurate at predicting the destination of new links in our data than the BA model, the overall accuracy of these models remains very low. For a more detailed treatment all of these models and others, we refer the reader to a paper by Mitzenmacher [106].

Explanatory growth models

Some recent studies, particularly on online social networks, have proposed explanatory models of the network growth. Unlike structural growth models, which try to model growth solely as a function of the network structure, explanatory models seek to account for the underlying sociological factors that cause the links to be established. For example, an explanatory growth model for Flickr, a photo-sharing social network, could be based on an understanding of how users behave when sharing pictures.

Examples of work on explanatory growth models include Kumar [82], who divided users into ones who are active and passive, and presented a model describing their behavior in an online social network. Jin et al. [70] presented a model of social networks based on known human interactions. Backstrom et al. [12] looked at snapshots of group membership in LiveJournal, and presented a model for the growth of user groups over time based on understandings of peer pressure. Finally, Chang et al. [27] proposed a model for the growth in connectivity of the Internet topology, modeling the decision processes of the administrators of autonomous systems.

Compared to structural growth models, explanatory models are more detailed, but they also tend to be specific to the network being investigated. For example, the reasons why autonomous systems connect to each other in the Internet topology are very different from the reasons why users in Flickr connect to each other. By being agnostic to these factors, structural growth models are inherently less accurate. But, they are far more general, and can be compared across different types of networks.

In this thesis, we focus only on structural growth models.

Validation of growth models

It is important to note here that both structural and explanatory growth models are, by and large, intuitive models that can explain the observed structural properties of the networks. But, they have not been significantly validated using empirical data. Mitzenmacher [107] poses this as one of the biggest challenges facing the future of power-law research. One of the contributions of this thesis lies in collecting data that can be used to determine how well these processes predict what actually occurs in different real-world networks at scale.

3.2.2 Observations of network growth

With the growth in popularity of online social networks, a few studies have examined the properties of the networks over time. We briefly describe these studies below.

A few studies have looked at how links are formed in social networks. Kossinets and Watts [80] used an inferred social network from an email trace to show that new links in the network are more likely to be established between nodes close to each other. Nowell et al. [93] investigated co-authorship networks in physics to test how well different graph proximity metrics can predict future collaborations. Newman [114] and Jeong et al. [69] examined the properties of scientific collaboration networks and found evidence of preferential attachment. Peltomäki and Alava [126] examined a scientific collaboration network and a movie-actor network and found

evidence of sub-linear preferential attachment.

Our work shares similar goals and methodology as the above studies. However, the data sets we use are orders of magnitude larger than the ones used before. Moreover, our data allows us to analyze network growth over a large number of samples. We analyze daily snapshots of Flickr and YouTube networks, and weekly snapshots of the Internet topology. For Wikipedia, we have sufficient data to create a snapshot of the network at the precise second a new link is established. Since the growth models rely solely on the current network structure to predict new link formation, having frequent snapshots of the network is crucial to validating the models with high accuracy.

Researchers have also studied the high-level properties of graph evolution, looking for evolution trends at the global level. For instance, Leskovec et al. [87] examined the evolution of a number of real-world graphs, including collaboration networks and recommendation networks. They found that the graphs tend to densify, and that the average path length tends to shrink (instead of growing in proportion to the number of nodes). Additionally, Kumar et al. [81] observed the early evolution of the blogosphere, and found that it is rapidly increasing in both scale and connectedness. This line of work is largely complementary to our work, as we focus on the local link formation phenomena which might lead to these global observations.

3.3 Detecting communities

We now turn our attention to the detection of communities in online social networks. A community is a subset of the users in a social network that is more tightly interconnected than the overall network [119]. Thus, all of the work described in this section tries to detect densely connected components of graphs. At a high level, the approaches can be divided into *global* approaches, which assume knowledge of the entire graph, or *local* approaches, which only assume detailed knowledge of a region of the network. After briefly describing how communities were detected classically in sociology, we describe the global and local approaches. Then, we describe empirical studies of social networks that have looked for the presence of communities.

3.3.1 Classical community detection

Classical community detection in sociology took the approach of partitioning the vertices in a social network into different communities while minimizing the number of edges between communities. Within this approach, there are two main algorithms: spectral bisection [130] and the Kernighan-Lin algorithm [73]. Both algorithms partition the graph into the best two communities possible, and then further subdivide those two until reaching the user-specified number of communities. However, both algorithms require the user to specify the sizes of the two communities initially, as well as the final number of communities desired.

3.3.2 Global community detection

One of the first community detection algorithms that did not assume pre-existing knowledge of the community structure was proposed by Girvan and Newman [119]. In short, their algorithm works by calculating the "most important" link in the network, and then removing it. The algorithm then repeats this step until the social network graph becomes partitioned, at which point the various partitions are considered as communities. Continuing to run the algorithm over the various partitions will produce even finer communities, until all of the links are removed from the network.

From the above description, it is clear that the selection of the most important link is integral to the functioning of the algorithm. A good metric of importance can quickly partition the graph into its various communities, while a bad metric can simply disconnect nodes one-by-one and produce degenerate partitions. Girvan and Newman suggested using the metric of betweenness centrality. The intuition behind Girvan and Newman's algorithm is simple: if we assume that the social network is divided into densely connected communities, the betweenness centrality metric looks for links that bridge communities. Since communities are, by definition, more dense than the graph as a whole, these bridging links will naturally have a higher betweenness centrality. Once they are removed from the graph, the underlying community structure emerges.

Newman [118] later proposed a faster, alternate approach, based on the greedy optimization of modularity. The algorithm starts with each vertex in a separate community, and merges pairs of communities, choosing at each stage the pair that

would yield the highest increase or smallest decrease in modularity. Clauset et al. [32] proposed a faster variant of this algorithm by further optimizing the operations with the use of more efficient data structures. These improvements in speed are important, as the running time of the original algorithm prohibited it from being used on graphs with more than a few thousand links.

Tyler et al. [153] presented a variant of the algorithm of Girvan and Newman, which improved the speed of the algorithm at the cost of accuracy. Instead of calculating the total betweenness centrality score by considering all paths starting at every vertex in the graph, Tyler et al. suggest that the betweenness centrality be calculated by summing over only a subset of the vertices, thereby obtaining a partial betweenness centrality score for all edges. The algorithm is run multiple times, yielding multiple community partitionings and are then aggregated into a single community partitioning using the technique proposed by Wilkinson et al. [161].

Radicchi et al. [131] proposed another algorithm based the approach of Girvan and Newman. It uses a local approximation to select the edges to be removed, which can be calculated quickly and, hence, runs faster. For each edge, it approximates the betweenness centrality by the number of loops of length three (i.e., triangles) that include the edge. Inter-community edges are unlikely to belong to many triangles, because they require another edge between the communities to complete the loop.

Other approaches have looked at finding multiple, overlapping community structures from a global perspective. This is in contrast to the previously discussed ap-

proaches, which were only concerned with finding the best way to partition the nodes into single, non-overlapping set of communities. The overlapping approaches include work by Palla et al. [123], which used k-cliques to find overlapping communities at different scales. Baumes et al. [16, 17] proposed a similar approach for finding overlapping communities by first looking for dense collections of nodes in the graph. Du et al. [46] presented an algorithm to detect communities in large-scale social networks by considering the overlapping nature of communities. Finally, Li et al. [92] proposed a separate approach for overlapping community detection based on triangle formation and clustering based on text similarity.

3.3.3 Local community detection

One potential downside of the global approaches to community detection is that the structure of the entire graph must be known; this is often prohibitively expensive (as many real-world graphs are extremely large) or hard to obtain (for example, the graph of Web pages). As an alternative, a number of researchers have looked at local approaches to detecting communities, which use only local knowledge to build a community around a set of source nodes. In contrast with the global approaches, local approaches have the potential to be significantly more scalable and applicable to much larger graphs, as well as graphs which are not completely visible due to privacy restrictions. Moreover, local approaches to community detection also hold the potential to detect multiple community structures – global approaches assign each

node to exactly one community, even if multiple such structures exist. Finally, local approaches allow for natural decentralization, as the computation can be trivially divided up and distributed.

Clauset [31] proposed one of the first local approaches to community detection, which was based on the greedy construction of a community around a source node. The algorithm creates a community by adding vertices one-by-one, choosing the vertex at each step that maximizes the ratio of intra-community edges to inter-community edges for the nodes on the "fringe" of the community. Thus, this algorithm tries to create a strong community by greedily picking nodes that have many links inside the community. Bagrow et al. [14] proposed an alternative algorithm, which adds all of the k-hop vertices at each step, until the ratio of inter-community to intra-community links falls below a threshold. Both of these were shown to detect communities in synthetic graphs, as well as a real-world product recommendation network. Recently, Wakita et al. [155] proposed a modification to the Clauset algorithm, which is capable of identifying communities in social networks with up to 5 million users. However, their work does not provide any validation of the community structure inferred from the network.

Additionally, two new local community detection algorithms have been proposed to improve the speed and performance of community detection. Luo et al. [99] proposed an algorithm similar to Clauset's, with the exception that it iteratively adds and removes nodes, continuing until adding or removing any single vertex would not

result in a better community. Bagrow [13] evaluated the performance of the various algorithms (and one additional newly proposed one), and found that the algorithm of Leo et al. performed the best on synthetically generated graphs.

It is important to note that none of these algorithms has been validated on a large-scale social network, primarily due to the lack of data availability. Typically, the algorithms are evaluated on synthetically generated graphs, product recommendation networks, or very small social networks such as Zachary's karate club [171], consisting of 34 members. Thus, it is not known whether they can detect communities in online social networks. In this thesis, we demonstrate the limitations of these approaches on data taken from an online social network, and present a new algorithm that addresses these limitations.

3.3.4 Observations of communities

A few studies have examined the community structures that exist in online social networks. The most notable of these is the work by Nazir et al. [113], which presented a large-scale measurement study of usage characteristics of applications in Facebook. They launched a few Facebook applications and, using the data collected from these applications, they characterized the workload, the structure of user interactions, and the existence of communities in the network. Their results, however, do not paint a complete picture of the Facebook network as they are only able to collect data on users who installed one of their applications. In contrast, we are able to collect data

on the majority of members of multiple Facebook networks, and do not require users to install any applications.

3.4 Preventing unwanted communication

Later in this thesis, we present Ostra, a system that leverages the difficulty in establishing and maintaining relationships in social networks to prevent malicious users from sending unwanted communication. Thus, we now describe related approaches, encompassing work that aims to prevent unwanted communication and work that leverages relationships in online social networks for other purposes.

Unwanted communication has long been a problem in the form of email spam, and many strategies have been proposed to deal with it. However, the problem increasingly afflicts other communication media such as IM, VoIP, and social networking and content-sharing sites. At a high level, there are three approaches for preventing unwanted communication in unicast systems. First, one can examine the content of the communication itself, looking for messages that are likely to be unwanted. Second, one can look at the reputation of the sender, focusing on users who send lots of unwanted communication. Third, one can impose a cost on the sender, with the hope that this cost will discourage the sending of unwanted messages. Additionally, in systems that allow multiple recipients, one can also look at allowing recipients to vote on whether content is wanted or not. These four approaches are discussed in detail below.

3.4.1 Content-based filtering

The most widespread approach to fighting unwanted communication is content-based filtering, where recipients have software that uses heuristics to automatically classify communication based on its contents. Popular examples are email filtering systems like SpamAssassin [148] and dSPAM [45]. Since content-based filters are installed at the email receiver, they lend themselves to customization and incremental deployment. Today's state-of-the-art filters are effective at blocking most spam, with some reporting correct classification of over 99% of messages [45]. Content-based filters are also used for other types of unwanted communication, such as blog spam [103] and network-based security attacks [84].

Content-based filtering, however, is subject to both false positives and false negatives. False negatives — that is, when unwanted communication is classified as wanted — are a mere inconvenience. False positives [5] are a much more serious concern, because relevant messages are marked as unwanted and thus may not be received [65]. Moreover, there is a continual arms race [64] between spammers and filter developers, because the cognitive and visual capabilities of humans allow spammers to encode their message in a way that users can recognize but filtering programs have difficulty detecting. When early spam filters looked for certain keywords and text patterns in messages, spammers started to misspell telltale words and include random text to escape detection. Nowadays, spammers embed text in images, requiring spam filters to use optical character recognition (OCR) components.

3.4.2 Originator-based filtering

Another approach for eliminating unwanted communication is to classify content based on the originator's history and reputation. One technique is whitelisting, where each user specifies a list of users who they are willing to receive content from.

Whitelisting is commonly deployed in IM applications such as iChat and AIM, in VoIP systems such as Skype, and in social networking sites such as LinkedIn. In these cases, users have to be on each other's whitelists (i.e., their lists of contacts) to be able to exchange messages. To get on each other's whitelists, two users must exchange a special invitation. If the invitation is accepted, the two parties are added to each other's whitelists. If the invitation is rejected, then the inviter is added to the invitee's blacklist, which prevents the inviter from contacting the invitee again. RE [57] extends whitelists to automatically and securely include friends of friends.

To be effective, whitelisting requires that users have unique identifiers and that content can be authenticated; otherwise, it is easy for malicious users to make their communication seem to come from a whitelisted source. In most deployed email systems, messages cannot be reliably authenticated. However, secure email services, IM, VoIP services, and social networking sites can authenticate content. Whitelisting, however, cannot eliminate unwanted invitations, which represents another form of unwanted communication.

Another, similar approach is blacklisting, where each user or site specifies a list of users who they are unwilling to receive content from. To be effective, though,

blacklisting requires strong user identities, as malicious users could otherwise trivially change identities. However, to date, no such large-scale strong identity system has been deployed, and the political, legal, and social issues associated with such a system make deployment in the near future unlikely.

3.4.3 Imposing a cost on the sender

A third approach taken to discourage unwanted communication is imposing a cost on the originators of either all communication or unwanted communication. The cost can be monetary or in terms of another limited resource. Optionally, systems can impose a cost only on the senders of unwanted communications, instead of imposing a cost on senders of both wanted and unwanted communication. We discuss some specific proposals below.

Quotas and micropayments

Quota- and payment-based approaches attempt to change the economics of unwanted communication by imposing a marginal cost on the transmission of an (unwanted) message.

Systems have been proposed where senders must commit to paying a per-message fee prior to sending digital communication [50, 138], and, occasionally, one-time fees imposed by a trusted organization [59]. These solutions attempt to model the offline postal service; they are based on the assumption that the cost will discourage mass distribution of unwanted messages. There are a few examples of deployed systems

that charge a per-message fee, such as LinkedIn [95] and Goodmail [59].

In some of the proposed systems, the per-message fee is charged only if the receiver classifies the messages as unwanted. This feature is desirable because it preserves the ability of any legitimate content originator to reach a large audience at low cost. Otherwise, users who send a large amount of wanted communication are subject to prohibitively high fees, thereby reducing the usefulness of the communication medium.

In general, one significant disadvantage of these proposals is that they typically require an extensive micropayment infrastructure, which some have claimed is impractical [2, 88, 120]. Additionally, the cost could be in terms of a resource other than money [139]. However, it may be difficult to ensure that the resource is readily available to anyone in quantities sufficient for legitimate communication, but hard to acquire in quantities needed for unwanted communication at a large scale.

Alternative systems impose a per-user quota on sending messages [156], limiting each user to sending only a certain number of messages per day. Systems based on quotas do not need micropayments but still require a market for the distribution of the user quotas. This market must ensure that legitimate senders can obtain a sufficient quota at reasonable cost, while the cost for spammers must be high enough to discourage large-scale unwanted communication.

Challenge-response systems

In so-called challenge-response systems, the sender of a message must prove she is a human (as opposed to a computer program) before the message is delivered. Like micropayment systems, challenge-response systems impose a cost on senders, but unlike micropayment systems, the cost is the human attention necessary to complete the challenge, rather than money.

Although challenge-response systems can limit the amount of unwanted communication, these systems have several disadvantages. One disadvantage is that they eliminate all automatically generated email messages, even when such messages are wanted. The typical way that this issue is handled is by having users specify a special email address to avoid false positives. Moreover, the need to complete a challenge may annoy and discourage some legitimate senders, as has been observed with captchas today [166]. Finally, if challenge-response systems were widely deployed, their designers could face an arms race to develop challenges that can be easily answered by most human users, yet cannot be answered by a program.

Legislation

Some countries have passed legislation that mandates a penalty for those sending unwanted communication. There is little evidence that it has significantly reduced the level of email spam received by users. There are at least two reasons why such laws have had limited impact. First, spammers are using technical means to obscure

the origin of unwanted communication. Second, the global nature of the Internet makes it easy for spammers to operate from a location where they face little or no risk of prosecution.

3.4.4 Content rating

Finally, a few proposals have looked at detecting unwanted communication in one-to-many communication systems (unlike one-to-one communication systems that we have discussed before). Many content-sharing sites (e.g., YouTube [167]) use content rating. Users can indicate the level of interest, relevance, and appropriateness of a content item they have viewed. The content is then tagged with the aggregated user ratings. Content ratings can help users to identify relevant content and avoid unwanted content. These ratings can also help system administrators to identify potentially inappropriate content, which they can then inspect and possibly remove. Moreover, content-rating systems can be manipulated, particularly in a system where new user identities can be created without significant cost. A recent proposal has looked at limiting the impact of multiple identities in the context of content rating [151], but has not been evaluating in a real-world system as of this writing.

3.4.5 Leveraging relationships

As Ostra uses the difficulty of establishing and maintaining links between users in a social network, we now briefly describe other work which leverages social network relationships. Trust between participating users has been used to replace certain

centralized functions. For example, the use of transitive trust has been leveraged in the PGP Web of Trust [172] to eliminate the need for a central certificate authority. SybilGuard [169] and SybilLimit [168] use social network links to identify users with many fake identities (Sybils). In brief, these systems look for large subsets of the network that are connected to the rest of the network by just a few edges, suggesting that the subset contains a number of fake identities.

Online social relationships are also used in several web-based applications to perform other tasks, such as content sharing [167], socializing [49], and professional networking [95]. LinkedIn uses implicit whitelisting of a user's friends and offers a (manual) introduction service based on the social network. Similarly, F2F [89] uses trust between users to provide a reliable storage system for user backups. However, none of these systems leverage the social network to automatically enable legitimate communication among users who have not had prior contact, while thwarting unwanted communication.

3.5 Personalized web search

Finally, in the last chapter of this thesis, we present the PeerSpective system, which integrates Web search with online social networks and displays pages browsed by friends in the results of a Web search. In this section, we describe systems which have similar aims. Several projects have looked at replacing the functionality of the large centralized Web search engines with a decentralized system. This architecture is

similar to PeerSpective, as they are built from contributing users' desktops [90]. Both Minerva [19] and YaCy [164] implement a peer-to-peer Web search engine without any points of centralization. Additionally, other projects [124, 141] have examined replacing the centralized PageRank computation of Google with a decentralized approach. All of these projects, though, are primarily focused on replacing the functionality of existing centralized search engines with a decentralized architecture, rather than improving or personalizing Web search.

A few systems have looked at personalizing responses to queries by taking a user's preferences and interests into account when ranking pages. Most notably, A9 [1] and Google Personalized Search [61] allow users to create profiles to which search results are tailored. However, these systems require users to create detailed profiles to perform well, which represents a significant burden on users. There has also been much research into methods for automatically personalizing search queries by inferring user interests [68, 150]. While these projects are concerned with personalization, the approach taken PeerSpective is complementary in that we obtain personalization by using social links to improve search results.

The MAAY [100] project has examined combining both of the above approaches, providing a decentralized and personalized search engine. In ISpy [146], organizations deploy a single web proxy which records the results of past queries and uses these to influence future ones. Their approach is limited to the single social group consisting of an organization, in contrast, PeerSpective is able to use an arbitrary social network

graph to influence search results.

Chapter 4

Measurement Methodology

We now describe the data used in this thesis, and the methodology used to collect it. We were not able to obtain data directly from the respective site operators, as most sites are hesitant to provide even anonymized data. Instead, we chose to crawl the user graphs by accessing the public web interface provided by the sites. This methodology gives us access to large data sets from multiple sites.

We discuss the data collected for each of the three measurement studies separately. First, we detail general challenges faced when crawling large social networks. Second, we describe the data collected for our analysis of the structure of online social networks, which includes data from multiple online social networks and the Web. We then describe the data collected to study how online social networks grow, and finally, we detail the data collected to study community structures.

4.1 Challenges in crawling large graphs

Crawling large, complex graphs presents unique challenges. In this section, we discussing the details of how we crawled each network after we describe our general approach. Most real-world graphs have been shown to have a dominant large connected component [23], which we call the large WCC. Therefore, we focus our methodology

on crawling this component of the graph.

4.1.1 Crawling the entire large WCC

The primary challenge in crawling large graphs is covering the entire giant connected component. At each step, one can generally only obtain the set of links into or out of a specified node. In the case of online social networks, crawling the graph efficiently is important since the graphs are large and highly dynamic. Common algorithms for crawling graphs include breadth-first search (BFS) and depth-first search.

Often, crawling the entire giant connected component is not feasible, and one must resort to using samples of the graph. Crawling only a subset of a graph by ending a BFS early (called the *snowball method*) is known to produce a biased sample of nodes [85]. In particular, partial BFS crawls are likely to overestimate node degree and underestimate the level of symmetry [18]. In social network graphs, collecting samples via the snowball method has been shown to underestimate the power-law coefficient, but to more closely match other metrics, including the overall clustering coefficient [85]. However, some previous studies of social networks have used small graph samples. Example studies have used samples of 0.3% of Orkut users [6], less than 1% of LiveJournal communities [12], and 0.08% of MySpace users [6]. In this thesis, we obtain and study much larger samples of the user graphs.

4.1.2 Using only forward links

Crawling directed graphs, as opposed to undirected graphs, presents additional challenges. In particular, many graphs can only be crawled by following links in the forward direction (i.e., one cannot easily determine the set of nodes which point *into* a given node). Using only forward links does not necessarily crawl an entire WCC; instead, it explores the connected component reachable from a set of seed users. This limitation is typical for studies that crawl online networks, such as the Web [23].

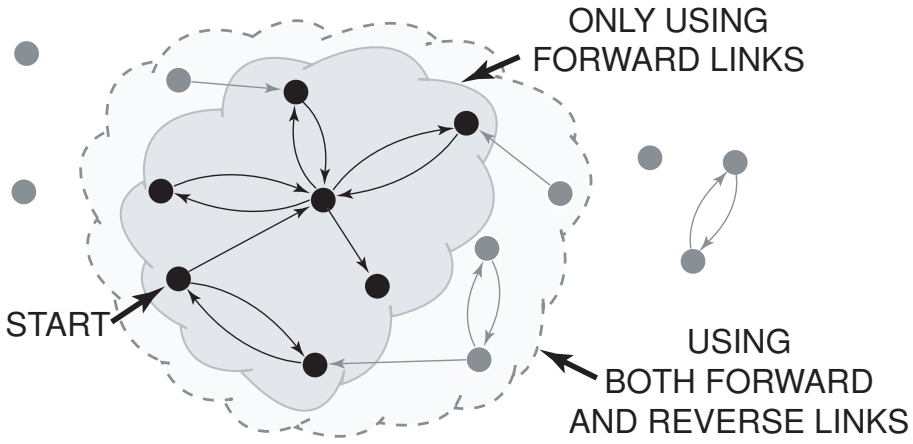


Figure 4.1 : Users reached by crawling different link types. If only forward links are used, we can reach only the inner cloud (shaded cloud); using both forward and reverse links, we can reach the entire WCC (dashed cloud).

Figure 4.1 shows an example of a directed graph crawl. The users reached by following only forward links are shown in the shaded cloud, and those reached using both forward and reverse links are shown in the dashed cloud. Using both forward and reverse links allows us to crawl the entire WCC, while using only forward links results in a subset of the WCC.

4.2 Capturing social networks' structure

We now discuss our methodology for crawling each of the networks, their limitations, and high-level statistics of the resulting data sets. Using automated scripts on a cluster of 58 machines, we crawled the social network graphs of Flickr, LiveJournal, Orkut, and YouTube. We chose these four sites because they are among the most popular social networking sites and they allow us to view the links out of any user in the network. In each step of our crawls, we retrieved the list of friends for a user we had not yet visited and added the retrieved users to the list of users to visit. We continued until we exhausted the list. This corresponds to a BFS of the social network graphs. High-level statistics of the resulting data sets are presented in Table 5.1.

Since the focus of this part is to investigate the structure of online social networks, we focus on the large WCC of the corresponding graphs in the rest of this thesis. As we show later in this section, the large WCC is structurally the most "interesting" part of the network. The nodes not included in the large WCC tend to be either part of very small, isolated clusters or are not connected to other users at all.

4.2.1 Flickr

Flickr (`www.flickr.com`) is a photo-sharing site based on a social network. The Flickr data presented in this thesis is from a crawl of the large WCC conducted on January 9th, 2007, and contains over 1.8 million users and 22 million links. Flickr exports an API for third-party developers, and we used this API to conduct the crawl. We also

obtained group membership information via Flickr's API.[1]

Flickr only allows us to query for forward links. Therefore we were unable to crawl the entire large WCC. To estimate the fraction of users who are part of the WCC but missing in our crawl, we performed the following experiment. We used the fact that the vast majority of Flickr user identifiers take the form of *[randomly selected 8 digit number]@N00*. We generated 100,000 random user identifiers of this form (from a possible pool of 90 million) and found that 6,902 (6.90%) of these were assigned usernames. These 6,902 nodes form a random sample of Flickr users.

Among these 6,902 users, 1,859 users (26.9%) had been discovered during our crawl. Focusing on the 5,043 users *not* previously discovered by our crawl, we conducted a BFS starting at each user to determine whether or not they could reach our set of previously crawled users. We found that only 250 (5.0%) of the missed users could reach our crawled set and were definitively in the WCC. While we cannot conclusively say that the remaining 4,793 (95.0%) missed users are not attached to the WCC (there could be some other user who points to them and to the WCC), the fact that 89.7% of these have no forward links suggests that many are not connected at all.

Finally, to explore how the remaining missing nodes are connected, we crawled the social network using these missing users as seeds, and compared the results with

[1]Flickr also allows users to form private groups, which do not appear in the user's profile list. We were unable to determine any information about the membership of such groups.

our initial crawl. We found only 11,468 new nodes that were not in the connected component of 1.8 million nodes discovered in the original crawl. Of these new nodes, 5,142 (44.8%) were nodes with no forward links, 3,370 (29.3%) had one link, 620 (5.4%) had two or three links, and 2,336 (20.3%) had four or more links. Thus, the nodes missing from our crawls tend to have low degree and are connected only to small clusters that are not reachable from the large connected component we crawled.

Thus, we believe that our crawl of the large WCC, although not complete, covers a large fraction of the users who are part of the WCC. Further, our experience with the randomly generated Flickr user identifiers indicates that (at least for Flickr), the nodes not in the largest WCC do not form large subgraphs.

4.2.2 LiveJournal

LiveJournal (`www.livejournal.com`) is a popular blogging site whose users form a social network. The LiveJournal data set considered in this thesis is the largest we examine: it contains over 5.2 million users and 72 million links. Due to its size, the LiveJournal crawl took several days, from December 9-11, 2006. LiveJournal offers an API that allows us to query for both forward and reverse links. We followed both link types to crawl the entire large WCC. We also obtained group membership information via LiveJournal's API.[2]

To estimate the fraction of the LiveJournal network covered by our crawl, we

[2]We inferred groups in LiveJournal by crawling the *interests* specified by users.

used a feature of LiveJournal[3] that returns random users. We selected a list of 5,000 random LiveJournal users and then checked how many of these random users our crawl had already covered. We found that we had already crawled 4,773 (95.4%) of these users, showing that our LiveJournal crawl covers the vast majority of the LiveJournal population. Finally, we started another crawl from the previously unknown 227 users to determine how many additional users could be discovered. This technique found only 73 additional users. These results suggest that our LiveJournal crawl covers almost the entire LiveJournal user population, and that the users not included in our crawl are part of small, isolated clusters.

Using the entire WCC from LiveJournal, we calculated the fraction of the WCC that is not reachable by using only forward links (as we did for the Flickr and YouTube crawls). We found that of the 5,284,457 nodes in the discovered WCC, only 404,134 (7.64%) would have been missed had we followed only forward links. Finally, we examined the 404,134 users who would have been missed to see how well these users were connected. We found that 201,694 (49.9%) of these users had a single forward link, 86,561 (21.1%) had two or three links, and 78,463 (19.4%) of the users had four or more forward links. Since, as we will show later, Flickr and YouTube share many characteristics with LiveJournal, this result suggests that the users that are missing in our Flickr and YouTube crawls tend to be small in number and have relatively small outdegree.

[3] `http://www.livejournal.com/random.bml`

4.2.3 Orkut

The next site we examined is Orkut (`www.orkut.com`), a social networking site run by Google. Orkut is a "pure" social network, as the sole purpose of the site is social networking, and no content is being shared. In Orkut, links are undirected and link creation requires consent from the target. Since, at the time of the crawl, new users had to be invited by an existing user to join the system, the Orkut graph forms a single SCC by definition.

The Orkut data considered in this thesis was collected during a crawl performed between October 3rd and November 11th, 2006. Because Orkut does not export an API, we had to resort to the bandwidth-intensive process of HTML screen-scraping to conduct our crawl. We obtained group information in a similar manner. Crawling Orkut presented other challenges, as Orkut limits the rate at which a single IP address can download information and requires a logged-in account to browse the network. As a result, it took more than a month to crawl a total of 3,072,441 users, at which point we stopped. This subset of the entire network corresponds to 11.3% of Orkut's user population of about 27 million users at the time of the crawl. The Orkut data considered in this thesis, therefore, is limited to this connected component and disregards all links from this component to other, uncrawled users.

Because our Orkut data set contains only a sample of the entire Orkut network, there are two potential concerns with the representativeness of the data. The first concern is whether the 11.3% subset of the network we gathered would be similar to

a different 11.3% subset gathered in the same way. In other words, are the properties of our sample representative of other samples of similar size? The second concern is whether the properties of our sample are representative of the properties of the network as a whole.

To explore the first of these concerns, we conducted five separate, small crawls of Orkut starting from random locations. Our random starting locations were chosen using Maximum-Degree random sampling [11] configured with a path length of 100,000 hops. Each of the five crawls was configured to cover 80,000 nodes in the same manner as our single, large crawl. We then compared the properties of the resulting samples.

We found that the properties of the five smaller crawls were similar, even though these crawls covered only 0.29% of the network each. For example, we found that the clustering coefficient of these crawls had an average of 0.284 with a standard deviation of 0.040. Similarly, we found that the scale-free metric had an average of 0.550 with a standard deviation of 0.083 (both of these metrics are discussed in more detail in the following section). Thus, we believe that the properties of our 11.3% sample of the network are likely to be similar to other crawls of similar size that are done in the same manner.

However, we caution the reader to be mindful of the second concern when extrapolating the results from our crawl to the entire Orkut network. Partial BFS crawls are known to over-sample high-degree nodes, and under-sample low-degree nodes [85].

This has been shown to overestimate the average node degree and to underestimate the level of symmetry [18]. Thus, our results may not be representative of the Orkut network as a whole.

4.2.4 YouTube

YouTube (`www.youtube.com`) is a popular video-sharing site that includes a social network. The YouTube data we present was obtained on January 15th, 2007 and consists of over 1.1 million users and 4.9 million links. Similar to Flickr, YouTube exports an API, and we used this feature to conduct our crawls.

YouTube allows links to be queried only in the forward direction, similar to Flickr. Unfortunately, YouTube's user identifiers do not follow a standard format,[4] and we were therefore unable to create a random sample of YouTube users. Also, YouTube does not export group information via the API. Instead, we obtained group membership information by screen-scraping the HTML pages attached to user profiles.

Because we were unable to crawl reverse links or estimate the size of the user population in YouTube, we advise the reader to be cautious in extrapolating the YouTube results to the entire YouTube population, as we do not know the number of users who do not participate in the social network.

[4]YouTube's user identifiers are user-specified strings.

4.2.5 Web graph

In order to compare the structure of social networks with that of the Web, we made use of data collected by the Stanford WebBase Project [149]. We employed the data from their crawl of November 2003. The crawl includes 8.6 million pages and 132 million hyperlinks collected from over 3900 crawled Web sites.

4.2.6 Summary

Our results indicate that

- The Flickr and YouTube data sets may not contain some of the nodes in the large WCC, but this fraction is likely to be very small.
- The LiveJournal data set covers almost the complete population of LiveJournal, and contains the entire large WCC.
- The Orkut data set represents a modest portion of the network, and is subject to the sampling bias resulting from a partial BFS crawl.

Moreover, the results also indicate that the vast majority of missed nodes in Flickr, LiveJournal, and YouTube have low degree and are likely to be part of small, isolated clusters.

Based on the number of user accounts each site claimed to have at the time of the crawl, we estimate the fraction of nodes our crawls cover in Table 4.1. Note that, unfortunately, we do not know the number of accounts in YouTube and were

	Users crawled	Total population	Coverage
Flickr	1,846,198	6,800,000	26.9%
LiveJournal	5,284,457	5,500,000	95.4%
Orkut	3,072,441	27,000,000	11.3%
YouTube	1,157,827	n/a	n/a

Table 4.1 : Coverage of social networking site crawls.

therefore unable to estimate the fraction of the population that our 1.1 million crawled YouTube users represent.

4.3 Capturing group membership

All the sites we considered allow users to form groups. We determined a user's group membership using corresponding APIs in Flickr and LiveJournal. On YouTube and Orkut, we determined a user's groups by screen-scraping the HTML pages that contain the user's profile. Note that since Flickr allows users to form private groups, we were unable to determine any information about the membership of such groups.

4.4 Capturing social networks' growth

We now describe the methodology for collecting information on the evolution of online social networks, and the data we collected. In order to compare our results to the evolution of other, well-understood networks, we also collected data on the evolution of the Wikipedia article network and the Internet's autonomous system network.

Descriptions of the methodology for collecting data on these are also included below.

Using automated scripts on a cluster of 58 machines, we crawled the social network graphs of Flickr and YouTube once per day. We chose these sites because they represent different types of online social networking sites and because it is possible to crawl the entire network once per day. On each day, we revisited every user we had previously discovered, in additional to all nodes that were reachable from the seed node, and recorded any newly created or removed links and nodes.

Since the sites do not provide the time of creation for any node or link, our growth data for the social networks has a granularity of one day for the links we observed being created. As a result, we cannot determine the exact time of link creation, or the order in which links were created within a single day. Moreover, new nodes cannot be observed until they become connected to one of the nodes we have already crawled. Additionally, in the rest of the thesis, we only examine links that we observed being created. In other words, we may discover a new node that has a few established links, but we do not examine these previously established links in our growth analysis, as we did not observe them being created.

4.4.1 Flickr

We crawled the Flickr network daily between November 2nd, 2006 and December 3rd, 2006, and again daily between February 3rd, 2007 and May 18th, 2007, representing a total of 104 days of growth. During that period of daily growth observations, we

observed over 10.7 million new links being formed and discovered over 680,000 new users. This represents, relative to the initial network snapshot, over 42% growth in the number of users and over 63% growth in the number of links.

4.4.2 YouTube

We crawled the YouTube network daily between December 10th, 2006 and January 15th, 2007, and again daily between February 8th, 2007 and July 23rd, 2007, representing 201 days of growth. Between the two date ranges of our crawls, YouTube changed its policy to require confirmation from the destination of a link (previously, this approval was not required). Thus, between our two observation periods, YouTube changed from a directed network to an undirected network. To properly analyze the data before and after this significant change in policy, we treat the two YouTube networks separately — we denote the first set of growth data covering the directed graph as *YouTube-D* and the second set representing the undirected graph as *YouTube-U*.

The YouTube-D data set represents the growth of a directed network over a period of 36 days. During that period of daily growth observations, we observed over 540,000 new links being formed and discovered over 130,000 new users. This represents, relative to the initial network snapshot, over 13% growth in the number of users and over 12% growth in the number of links.

The YouTube-U data set represents the growth of an undirected network over a period of 165 days. We observed the network grow by over 11.7 million links and over

1.8 million users. This represents, relative to the initial network snapshot, over 129% growth in the number of users and over 173% growth in the number of links.

4.4.3 Wikipedia

Wikipedia (`www.wikipedia.org`) is a popular online encyclopedia that allows any user to add or edit content. Wikipedia makes its entire edit history available on a monthly basis, and we downloaded the edit history of the English language Wikipedia as of April 6th, 2007.

To extract the graph of links between Wikipedia pages, we used the following method: for each link in the current snapshot, we determined the time when this link was first created. We then construct a graph using these derived links and the associated timestamps. This method allows us to remove the effects of page vandalism, where malicious users sometimes overwrite entire pages, thereby temporarily removing all of the links from vandalized pages.

Since Wikipedia allows pages to redirect to other pages, we configured our tool to follow the redirects, and treat a link to a redirect page as if it was a link to the destination page. Thus, if page A originally linked to B at time t, but later, B was set to redirect to C, we treat this like a link from A to C established at time t. This allows us to handle multiple layers of redirect pages, as well as large-scale naming convention changes.

Since the data represents the complete history of a complex network, we exclude

startup effects by limiting our analysis to the recent history. This is similar to previous studies [25, 114]. In particular, we only consider links created between January 1st, 2005 and April 6th, 2007, a period of 826 days. During this period, we observed over 1.1 million new pages and over 33 million new links, representing 169% growth in the number of pages and 500% growth in the number of links relative to the snapshot on January 1st, 2005.

4.4.4 Internet topology

The Internet can be viewed as a collection of *autonomous systems* (AS), where each AS represents a single administrative domain (typically, an ISP). The inter-domain routing protocol of the Internet, BGP, uses unique AS numbers to allow ASes to advertise their connections to their neighbors. The union of these advertisements forms an undirected graph representing the AS-level connectivity of the Internet.

We used the AS topology graphs collected by CAIDA [24] to study the evolution of the AS network. CAIDA creates weekly (monthly for the first two years) snapshots of the AS topology using a number of BGP monitoring machines. We downloaded the entire history of their measurements, which covers the period from January 5th, 2004 until July 9th, 2007. The AS topology evolution data therefore covers 1,282 days of growth. During this period, the number of ASes in the network grew from 9,978 to 25,526, a growth of 155%. Similarly, the number of AS links grew from 29,504 to 104,824, a growth of 255%.

4.5 Capturing communities

In this section, we describe the data set we collected for close analysis of community information, and we discuss its limitations.

4.5.1 Measurement methodology

Our data set was collected by crawling part of the Facebook [49] social network through the site's public web interface. We crawled the part of the Facebook social network that consists of Rice University students and alumni. We started by logging into the Facebook user account of one of the authors, who is a student at Rice University. We then conducted a breadth-first-search (BFS) of all reachable users in the Rice network, in the same manner as in previous work [105]. By default, Facebook allows all users whose email addresses have the same domain (`rice.edu` in this case) to view each others' friends, and we were thus able to crawl a large portion of the Rice Facebook network.

The data collected for this thesis is from a crawl conducted over 9 hours on May 17th, 2008. In total, we discovered 6,156 users in our crawl, who are connected together with 377,350 links. This represents a network with an average user degree of 61.29.

4.5.2 Collected data

From the Facebook crawl, we only collected the names of the users and their list of friends. We collected additional information about the users by querying the Rice University Student Directory [136] and the Rice University Alumni Directory [135]. From these two directories, we were able to determine the users' matriculation year, graduation year, residential college[5], and major(s) or department.

To correlate the Facebook user list with the directories, we first looked up each user's name in the Student Directory, and then the Alumni Directory. If a single entry was found in either directory, the information from that entry was used. If multiple entries were found that exactly matched the student's name, we disregarded the student. We used a conservative matching policy: only exact name matches were used.[6]

Overall, we found unique matches for 1,781 students in the Student Directory and 2,093 additional students in the Alumni Directory. This left us with 2,282 Facebook users who we were unable to match with a directory listing; we disregarded these users. Of the 3,874 students we were able to find records for, 1,233 (31.8%) were

[5]Rice University has nine residential colleges, to which incoming undergraduate students are randomly assigned. The colleges serve as dormitories, cafeterias, and social circles; students stay at the same college during their entire undergraduate tenure.

[6]The only exception was a list of common nicknames, such as Bob for Robert and Chris for Christopher. In these cases, a match between a name and a nickname was allowed if there was only one entry found in the Facebook crawl and in the student or alumni directory.

current undergraduate students, 548 (14.1%) were current graduate students, 1,856 (47.9%) were undergraduate alumni, and 237 (6.11%) were graduate alumni.

As a point of reference, the total number of current undergraduate and graduate students at Rice is 3,001 and 2,144, respectively [134]. Thus, we were able to locate 41.1% of the current undergraduate and 25.6% of the current graduate students in Facebook.

4.5.3 Limitations

Our Facebook crawl includes only those users who had not changed the default Facebook privacy setting, which shares the friend list with users whose email address has the same domain. During our crawl, we found that 360 of the 6,156 users (5.85%) had changed their privacy settings so that their friend list was not accessible to us.

Our crawl is also limited by our ability to match names between Facebook accounts and information in the directories. Rice students can elect to remove their information from the online directory; in this case, we would not be able to find corresponding entries in the directories. Additionally, users with all but the most common nicknames are likely to be missed by our correlation procedure. Indeed, we found that we were unable to match 37.1% of the Facebook users we discovered with entries in either online directory.

Additionally, there may be users who were not connected to the large, strongly connected component of the social network we crawled. Because Facebook does not

provide a way to select random user accounts, we are unable to estimate the fraction of Rice University Facebook accounts that we were unable to crawl.

4.6 Data availability

All of the data sets considered in this thesis, with the exception of the Facebook data from Rice University, are available to the research community. The data has been anonymized in order to ensure the privacy of the social network users. A detailed description of the data format and downloading instructions are available at

`http://socialnetworks.mpi-sws.org`

Chapter 5

Network Structure

Unlike the Web, which is largely organized around content, online social networks are organized around users. Participating users join a network, publish their profile and any content, and create links to any other users with whom they associate. The resulting social network provides a basis for maintaining social relationships, for finding users with similar interests, and for locating content and knowledge that has been contributed or endorsed by other users.

An in-depth understanding of the graph structure of online social networks is necessary to evaluate current systems, to design future online social network based systems, and to understand the impact of online social networks on the Internet. For example, understanding the structure of online social networks might lead to algorithms that can detect trusted or influential users, much like the study of the Web graph led to the discovery of algorithms for finding authoritative sources in the Web [74]. Moreover, recent work has proposed the use of social networks to mitigate email spam [57], to improve Internet search [104], and to defend against Sybil attacks [169]. However, these systems have not yet been evaluated on real social networks at scale, and little is known to date on how to synthesize realistic social network graphs.

In this chapter, we characterize the structural properties of four online social networks. We compare the networks to each other, and we compare their properties with those previously observed for the Web. The Web is one of the most well-studied online networks, and our study shares much of its methodology with previous studies of the Web. Thus, it is perhaps natural to tend to compare our results with the structure of the Web. However, we are well aware that the user graph in social networks is fundamentally different from the interconnection of web pages; our comparisons serve more to calibrate our results than to point out (expected) differences.

The focus of our work is the social network users within the sites we study. More specifically, we study the properties of the large WCC in the user graphs of four popular sites. We do not attempt to study the entire user community (which would include users who do not use the social networking features), information flow, or workload of online social networking sites. While these topics are important, they are beyond the scope of this thesis.

5.1 High-level data statistics

Table 5.1 presents the high-level statistics of the four networks we examine in this chapter. The crawled network sizes vary by almost a factor of five (1.1 million users in YouTube versus 5.2 million in LiveJournal), and the number of links varies by almost two orders of magnitude (4.9 million in YouTube versus 223 million in Orkut). Similarly, other metrics such as the average number of friend links per node and user

	Flickr	**LiveJournal**	**Orkut**	**YouTube**
Number of users	1,846,198	5,284,457	3,072,441	1,157,827
Estimated coverage	26.9%	95.4%	11.3%	unknown
Dates of crawl	Jan 9, 2007	Dec 9 - 11, 2006	Oct 3 - Nov 11, 2006	Jan 15, 2007
Number of links	22,613,981	77,402,652	223,534,301	4,945,382
Friends per user	12.24	16.97	106.1	4.29
Fraction symmetric links	62.0%	73.5%	100.0%	79.1%
Number of groups	103,648	7,489,073	8,730,859	30,087
Memberships per user	4.62	21.25	106.44	0.25

Table 5.1 : High-level statistics of social networking site crawls.

participation in shared interest groups also vary by two to three orders of magnitude. Our analysis later will show that despite these differences, these graphs share a surprisingly large number of key structural properties.

5.2 Link symmetry

The fact that links are directed can be useful for locating content in information networks. For example, in the Web graph, search algorithms such as PageRank [122] consider a directed link from a source to a destination as an endorsement of the destination by the source, but not vice-versa. For instance, numerous Web pages point to sites like `cnn.com` or `nytimes.com`, but very few pages receive pointers back from these sites. Search engines leverage this to identify reputed sources of

information, since pages with high indegree tend to be authorities [74].

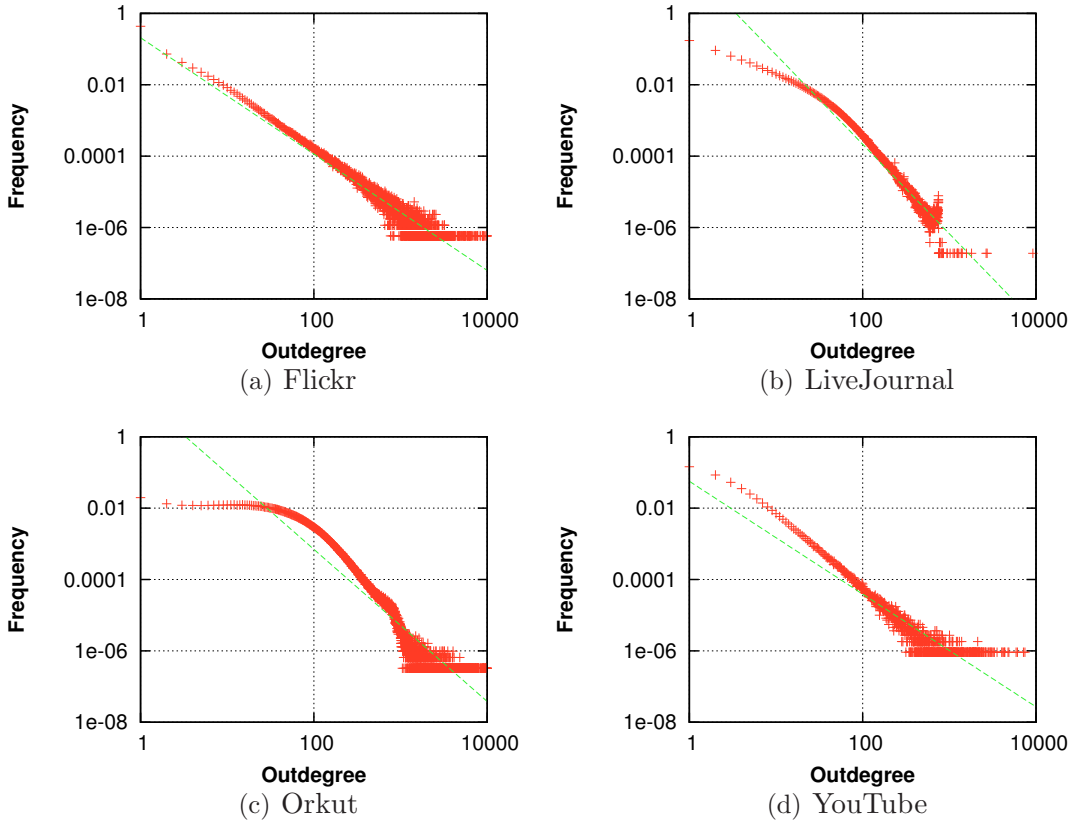
With the exception of Orkut, links in the social networks we studied are directed and users may therefore link to any other user they wish. The target of the link may reciprocate by placing a link pointing back at the source. Our analysis of the level of symmetry in social networks, shown in Table 5.1, reveals that all three social networks with directed links (Flickr, LiveJournal, and YouTube) have a significant degree of symmetry. Their high level of symmetry is consistent with that of offline social networks [63]. Furthermore, social networking sites inform users of new incoming links, which may also contribute to the high level of symmetry.

Independent of the causes, the symmetric nature of social links affects the network structure. For example, symmetry increases the overall connectivity of the network and reduces its diameter. Symmetry can also make it harder to identify reputable sources of information just by analyzing the network structure, because reputed sources tend to dilute their importance when pointing back to arbitrary users who link to them.

5.3 Power-law node degrees

We begin to examine the graph structure by considering the node degree distribution. As discussed in Chapter 3, the degree distributions of many complex networks, including offline social networks, have been shown to conform to power-laws. Thus, it may not be surprising that social networks also exhibit power-law degree distri-

butions. However, as our analysis shows, the degree distributions in social networks differ from that of other power-law networks in several ways.

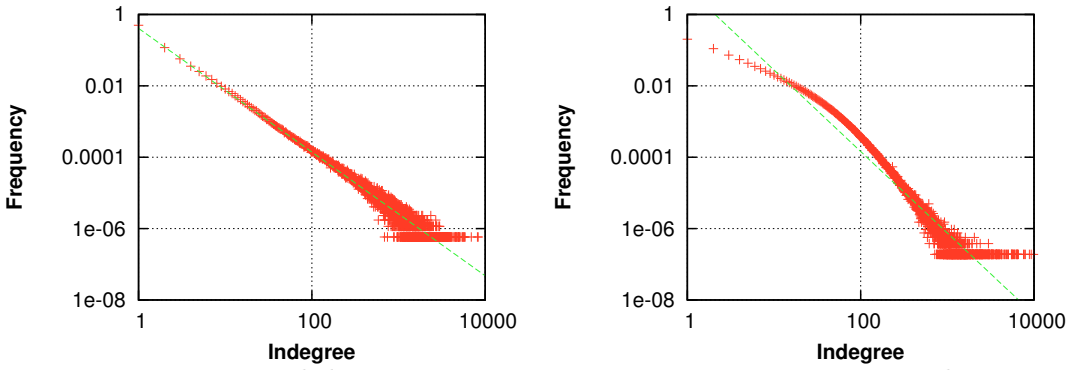


(a) Flickr (b) LiveJournal

(c) Orkut (d) YouTube

Figure 5.1 : Log-log plot of outdegree complementary cumulative distribution functions (CCDF). All social networks show properties consistent with power-law networks.

Figures 5.1 and 5.2 shows the outdegree and indegree complementary cumulative distribution functions (CCDF), respectively, for each measured social network. All of the networks show behavior consistent with a power-law network; the majority of the nodes have small degree, and a few nodes have significantly higher degree. To test how well the degree distributions are modeled by a power-law, we calculated the



(a) Flickr (b) LiveJournal

(c) Orkut (d) YouTube

Figure 5.2 : Log-log plot of indegree complementary cumulative distribution functions (CCDF). All social networks show properties consistent with power-law networks.

best power-law fit using the maximum likelihood method [33]. Table 5.2 shows the estimated power-law coefficient along with the Kolmogorov-Smirnov goodness-of-fit metric [33]. While the best power-law coefficients approximate the distributions very well for Flickr, LiveJournal, and YouTube, the Orkut data deviates significantly.

Two factors contribute to this deviation. First, our Orkut crawl reached only 11.3% of the network — partial BFS crawls tend to undersample nodes with lower degree, which can explain the flat head of the distribution [85]. Second, both LiveJournal and Orkut artificially cap a user's number of outgoing links,[1] which leads to a distortion in the distribution for high degrees.

	Outdegree		**Indegree**	
Network	α	D	α	D
Web [23]	2.67	-	2.09	-
Flickr	1.74	0.0575	1.78	0.0278
LiveJournal	1.59	0.0783	1.65	0.1037
Orkut	1.50	0.6319	1.50	0.6203
YouTube	1.63	0.1314	1.99	0.0094

Table 5.2 : Power-law coefficient estimates (α) and corresponding Kolmogorov-Smirnov goodness-of-fit metrics (D). The Flickr, LiveJournal, and YouTube networks are well approximated by a power-law.

[1]Orkut caps the outdegree at 1,000, and LiveJournal at 750. Both of these caps were instituted after the networks were established, and some users therefore exceed the caps. Also, Flickr has since instituted a cap of 3,000 *non-reciprocal* links; however, the data shown here was collected before this cap was established.

Additionally, we tested the stability of the power-law coefficient estimates by running the maximum likelihood estimator over varyingly sized subsamples of our data [162]. We found that the estimates of the power-law coefficient were remarkably stable; the estimates varied by less than 6% from those provided in Table 5.2 when we considered as few as 1,000 data points.

Table 5.2 also shows a difference between the structure of social networks and that of previously observed networks. In the Web, for example, the indegree and outdegree power-law exponents have been shown to differ significantly, while the power-law exponents for the indegree and outdegree distributions in each of our social networks are very similar. This implies that in online social networks, the distribution of outgoing links is similar to that of incoming links, while in the Web, the incoming links are significantly more concentrated on a few high-degree nodes than the outgoing links.

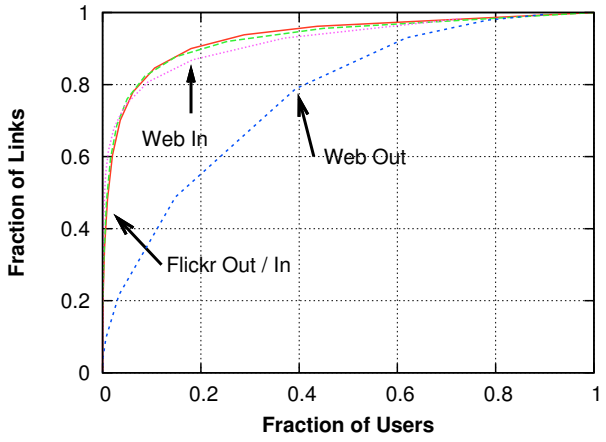


Figure 5.3 : Plot of the distribution of links across nodes. Social networks show similar distributions for outgoing and incoming links, whereas the Web links shows different distributions.

Focusing on this difference, Figure 5.3 shows the distribution of incoming and outgoing links over nodes in the Web and Flickr graphs.[2] The difference is readily apparent: 5% of the Web nodes account for 75% of all incoming links, but for only 25% of all outgoing links. In all social networks we considered, the distributions of incoming and outgoing links across the nodes are very similar. We now examine this phenomenon in more detail.

5.4 Correlation of indegree and outdegree

Studies of the indegree and outdegree distributions in the Web graph helped researchers find better ways to find relevant information in the Web. In the Web, the population of pages that are *active* (i.e., have high outdegree) is not the same as the population of pages that are *popular* (i.e., have high indegree) [74]. For example, many Web pages of individual users actively point to a few popular pages like `wikipedia.org` or `cnn.com`. Web search techniques are very effective at separating a very small set of popular pages from a much larger set of active pages.

In social networks, the nodes with very high outdegree also tend to have very high indegree. In our study, for each network, the top 1% of nodes ranked by indegree has a more than 65% overlap with the top 1% of nodes ranked by outdegree. The corresponding overlap in the Web is less than 20%. Hence, active users (i.e., those

[2]The Flickr topology is representative of all four networks; we omitted the others in the plot for readability.

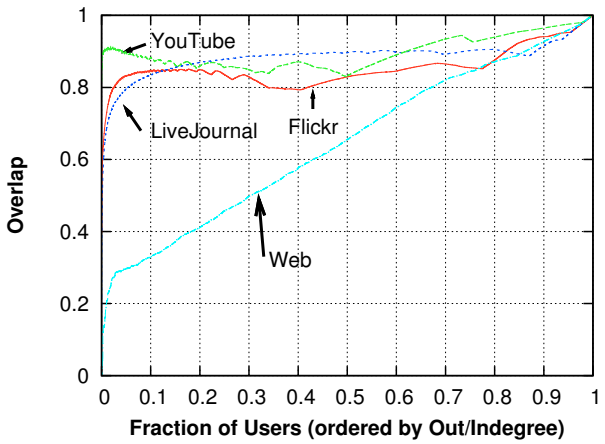


Figure 5.4 : Plot of the overlap between top $x\%$ of nodes ranked by outdegree and indegree. The high-indegree and high-outdegree nodes are often the same in social networks, but not in the Web.

who create many links) in social networks also tend to be popular (i.e., they are the target of many links). Figure 5.4 shows the extent of the overlap between the top $x\%$ of nodes ranked by indegree and outdegree.

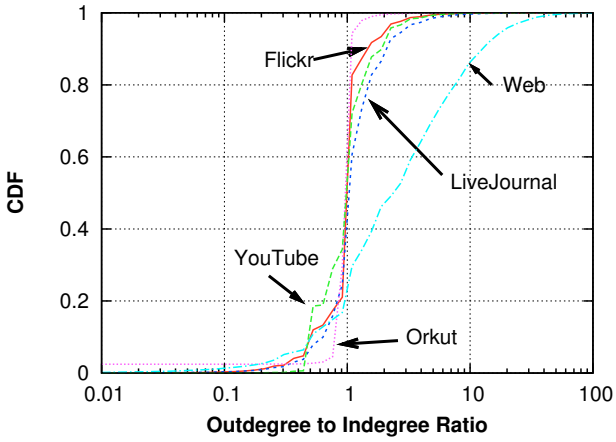


Figure 5.5 : CDF of outdegree to indegree ratio. Social networks show much stronger correlation between indegree and outdegree than the Web.

Next, we compared the indegree and outdegree of individual nodes in the social

networks. Figure 5.5 plots the cumulative distributions of the outdegree-to-indegree ratio for the four social networks and the Web. The social networks show a remarkable correspondence between indegree and outdegree; for all networks, over 50% of nodes have an indegree within 20% of their outdegree. The distribution for the Web is markedly different; most nodes have considerably higher outdegree than indegree, while a small fraction of nodes have significantly higher indegree than outdegree.

The high correlation between indegree and outdegree in social networks can be explained by the high number of symmetric links. The high symmetry may be due to the tendency of users to reciprocate links from other users who point to them. This process would result in active users (who place many outgoing links) automatically receiving many incoming links, and lead to the distributions we have observed.

5.5 Path lengths and diameter

Next, we look at the properties of shortest paths between users. Table 5.3 shows the average path lengths, diameters, and radii[3] for the four social networks. In absolute terms, the path lengths and diameters for all four social networks are remarkably short. Interestingly, despite being comparable in size to the Web graph we considered, the social networks have significantly shorter average path lengths and diameters. This property may again result from the high degree of reciprocity within the social

[3]Due to the computational complexity associated with determining the actual radius and diameter, the numbers presented here are from determining the eccentricity of 10,000 random nodes in each network.

networks. Incidentally, Broder et al. [23] noted that if the Web were treated as an undirected graph, the average path length would drop from 16.12 to 7.

Network	Average Path Length	Radius	Diameter
Web [23]	16.12	475	905
Flickr	5.67	13	27
LiveJournal	5.88	12	20
Orkut	4.25	6	9
YouTube	5.10	13	21

Table 5.3 : Average path length, radius, and diameter of the studied networks. The path length between random nodes is very short in social networks.

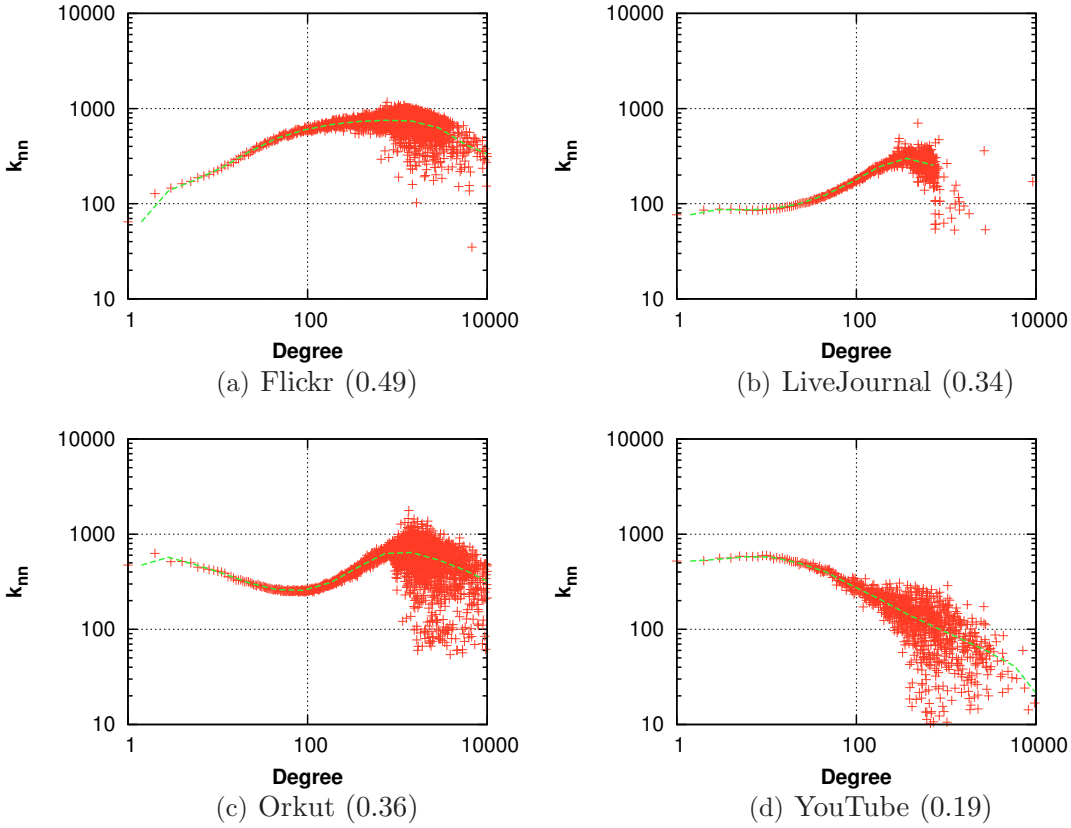
5.6 Link degree correlations

To further explore the difference in network structure between online social networks and previously observed networks, we examine which users tend to connect to each other. In particular, we focus on the *joint degree distribution* (JDD), or how often nodes of different degrees connect to each other.

5.6.1 Joint degree distribution

The JDD is approximated by the degree correlation function k_{nn}, which is a mapping between outdegree and the average indegree of all nodes connected to nodes of that outdegree. Clearly, an increasing k_{nn} indicates a tendency of higher-degree nodes to

connect to other high-degree nodes; a decreasing k_{nn} represents the opposite trend. Figure 5.6 shows a plot of k_{nn} for the four networks we studied.



(a) Flickr (0.49) (b) LiveJournal (0.34)

(c) Orkut (0.36) (d) YouTube (0.19)

Figure 5.6 : Log-log plot of the outdegree versus the average indegree of friends. The scale-free metrics, included in the legend, suggest the presence of a well-connected core.

The trend for high-degree nodes to connect to other high-degree nodes can be observed in all networks except YouTube (the unexpected bump at the head of the Orkut curve is likely due to the undersampling of users). This suggests that the high-degree nodes in social networks tend to connect to other high-degree nodes, forming a "core" of the network. Anecdotally, we believe that the different behavior seen in

YouTube is due its more "celebrity"-driven nature; there are a few extremely popular users on YouTube to whom many unpopular users connect.

To quantitatively explore this phenomenon, we next examine two metrics based on the joint degree distribution: the scale-free metric s and the assortativity r.

5.6.2 Scale-free behavior

The scale-free metric of the networks are shown in the legend of Figure 5.6. All of the networks with the exception of YouTube show a significant s, indicating that high-degree nodes tend to connect to other high-degree nodes, and low-degree nodes tend to connect to low-degree nodes.

5.6.3 Assortativity

The scale-free metric is related to the assortativity coefficient r, which is a measure of the likelihood for nodes to connect to other nodes with similar degrees. Recent work has suggested that the scale-free metric is more suitable for comparing the structure of different graphs [8], as it takes into account the possible configurations of networks with properties including connectedness and no self-loops. However, for completeness, we calculated the assortativity coefficients for each of the networks, and found 0.202 for Flickr, 0.179 for LiveJournal, 0.072 for Orkut, and -0.033 for YouTube.

The assortativity shows yet another difference between the social networks and other previously observed power-law networks. For example, the Web and the Internet have both been shown to have negative assortativity coefficients of -0.067 and

-0.189, respectively [116]. On the other hand, many scientific coauthorship networks, a different type of social network, have been shown to have positive r [116].

Taken together, the significant scale-free metric and the positive assortativity coefficient suggests that there exists a tightly-connected "core" of the high-degree nodes which connect to each other, with the lower-degree nodes on the fringes of the network. In the next few sections, we explore the properties of these two components of the graph in detail.

5.7 Densely connected core

We loosely define a *core* of a network as any (minimal) set of nodes that satisfies two properties. First, the core must be necessary for the connectivity of the network (i.e., removing the core breaks the remainder of the nodes into many small, disconnected clusters). Second, the core must be strongly connected with a relatively small diameter. Thus, a "core" is a small group of well-connected group of nodes that is necessary to keep the remainder of the network connected.

To more closely explore the core of the network, we use an approximation previously used in Web graph analysis [23]. Specifically, we remove increasing numbers of the highest degree nodes and analyze the connectivity of the remaining graph.[4] We calculate the size of the largest remaining SCC, which is the largest set of users who can mutually reach each other.

[4]The large size of the graphs we study makes a cut set analysis computationally infeasible.

As we remove the highest degree nodes, the largest SCC begins to split into smaller-sized SCCs. Figure 5.7 shows the composition of the splits as we remove between 0.01% and 10% of the highest-degree nodes in Flickr. The corresponding graphs for the other social networks look similar, and we omit them for clarity. Once we remove 10% of the highest indegree nodes,[5] the largest SCC partitions into millions of very small SCCs consisting of only a handful of nodes.

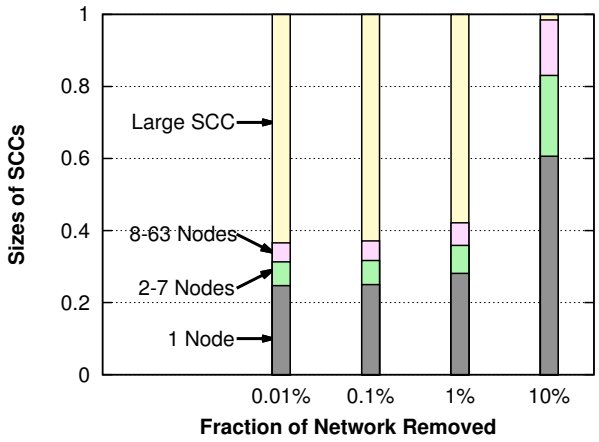


Figure 5.7 : Breakdown of network into SCCs when high-degree nodes are removed, grouped by SCC size.

To understand how much the network core contributes towards the small path lengths, we analyzed the path lengths of subgraphs containing only the highest-degree nodes. Figure 5.8 shows how path lengths increase as we generate larger subgraphs of the core by progressively including nodes ordered inversely by their degree. The average path length increases sub-logarithmically with the size of the core. In Flickr,

[5]We obtained the similar results when using both indegree and outdegree, thus we only present the indegree results here.

for example, the overall average path length is 5.67, of which 3.5 hops involve the 10% of nodes in the core with the highest degrees. This suggests that the high-degree core nodes in these networks are all within roughly four hops of each other, while the rest of the nodes, which constitute the majority of the network, are at most a few hops away from the core nodes.

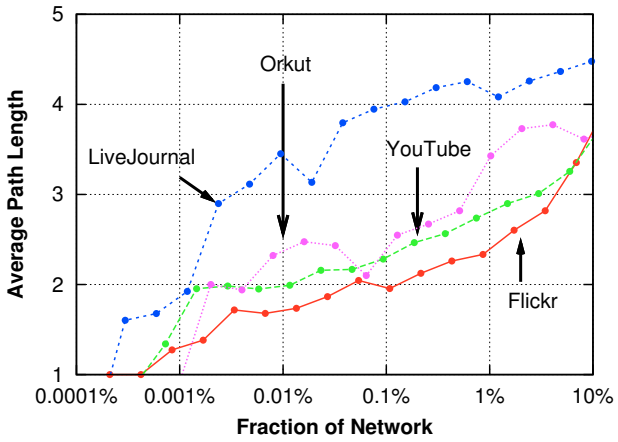


Figure 5.8 : Average path length among the most well-connected nodes. The path length increases sub-logarithmically.

Thus, the graphs we study have a densely connected *core* comprising of between 1% and 10% of the highest degree nodes, such that removing this core completely disconnects the graph.

The structure of social networks, with its high dependence on few highly connected nodes, may have implications for information flow, for trust relationships, and for the vulnerability of these networks to deliberate manipulation. The small diameter and path lengths of social networks are likely to impact the design of techniques for finding paths in such networks, for instance, to check how closely related a given pair of nodes

is in the network. Such techniques have applications, for instance, in social networks used to verify the trustworthiness or relevance of received information [57].

5.8 Tightly clustered fringe

Next, we consider the graph properties at the scale of local neighborhoods outside of the core. We first examine clustering, which quantifies how densely the neighborhood of a node is connected.

		Ratio to Random Graphs	
Network	C	**Erdös-Rényi**	**Power-Law**
Web [3]	0.081	7.71	-
Flickr	0.313	47,200	25.2
LiveJournal	0.330	119,000	17.8
Orkut	0.171	7,240	5.27
YouTube	0.136	36,900	69.4

Table 5.4 : The observed clustering coefficient, and ratio to random Erdös-Réyni graphs as well as random power-law graphs.

Table 5.4 shows the clustering coefficients for all four social networks. For comparison, we show the ratio of the observed clustering coefficient to that of Erdös-Réyni (ER) random graphs [48] and random power-law graphs constructed with preferential attachment [15], with the same number of nodes and links. ER graphs have no link bias towards local nodes. Hence, they provide a point of reference for the degree of local clustering in the social networks. Graphs constructed using preferential attach-

ment also have no locality bias, as preferential attachment is a global process, and they provide a point of reference to the clustering in a graph with a similar degree distribution.

The clustering coefficients of social networks are between three and five orders of magnitude larger than their corresponding random graphs, and about one order of magnitude larger than random power-law graphs. This unusually high clustering coefficient suggests the presence of strong local clustering, and has a natural explanation in social networks: people tend to be introduced to other people via mutual friends, increasing the probability that two friends of a single user are also friends.

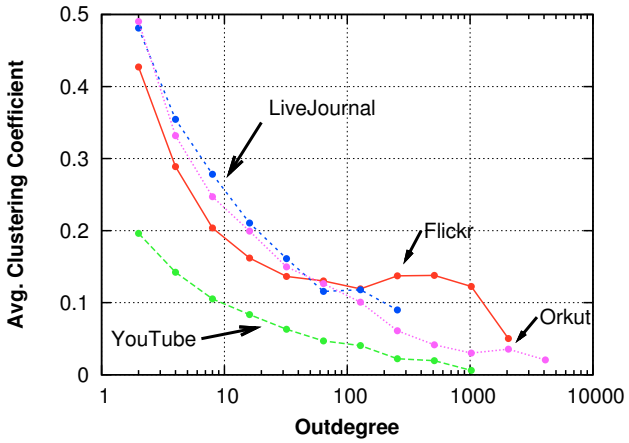


Figure 5.9 : Clustering coefficient of users with different outdegrees. The users with few "friends" are tightly clustered.

Figure 5.9 shows how the clustering coefficients of nodes vary with node outdegree. The clustering coefficient is higher for nodes of low degree, suggesting that there is significant clustering among low-degree nodes. This clustering and the small diameter of these networks qualifies these graphs as small-world networks [159], and further

indicates that the graph has scale-free properties.

5.9 Groups

In many online social networks, users with shared interests may create and join groups. Table 5.5 shows the high-level statistics of user groups in the four networks we study. Participation in user groups varies significantly across the different networks: only 8% of YouTube users but 61% of LiveJournal users declare group affiliations. Once again, the group sizes follow a power-law distribution, in which the vast majority have only a few users each.

Network	**Groups**	**Usage**	**Average Size**	**Average C**
Flickr	103,648	21%	82	0.47
LiveJournal	7,489,073	61%	15	0.81
Orkut	8,730,859	13%	37	0.52
YouTube	30,087	8%	10	0.34

Table 5.5 : Table of the high-level properties of network groups including the fraction of users which use group features, average group size, and average group clustering coefficient.

Note that users in a group need not necessarily link to each other in the social network graph. As it turns out, however, user groups represent tightly clustered communities of users in the social network. This can be seen from the average group clustering coefficients of group members, shown in Table 5.5.[6] These coefficients are

[6]We define the *group clustering coefficient* of a group G as the clustering coefficient of the subgraph

higher than those of the corresponding network graph as a whole (shown in Table 5.4). Further, the members of smaller user groups tend to be more clustered than those of larger groups. Figure 5.10 shows this by plotting the average group clustering coefficient for groups of different sizes in the four observed networks. In fact, many of the small groups in these networks are cliques.

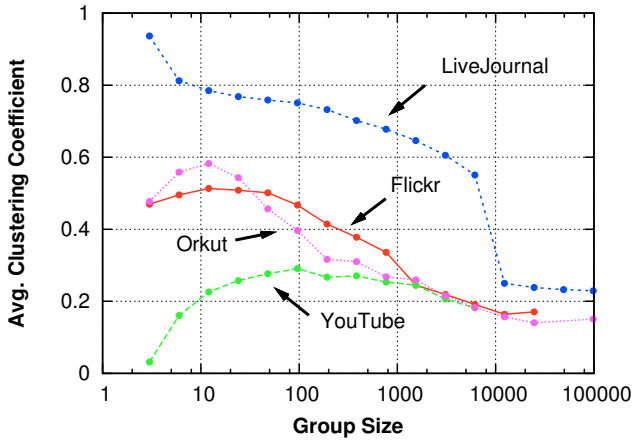


Figure 5.10 : Plot of group size and average group clustering coefficient. Many small groups are almost cliques.

Finally, Figure 5.11 shows how user participation in groups varies with outdegree. Low-degree nodes tend to be part of very few communities, while high-degree nodes tend to be members of multiple groups. This implies a correlation between the link creation activity and the group participation. There is a sharp decline in group participation for Orkut users with over 500 links, which is inconsistent with the behavior of the other networks. This result may be an artifact of our partial crawl of the Orkut network and the resulting biased user sample.

---

of the network consisting of only the users who are members of G.

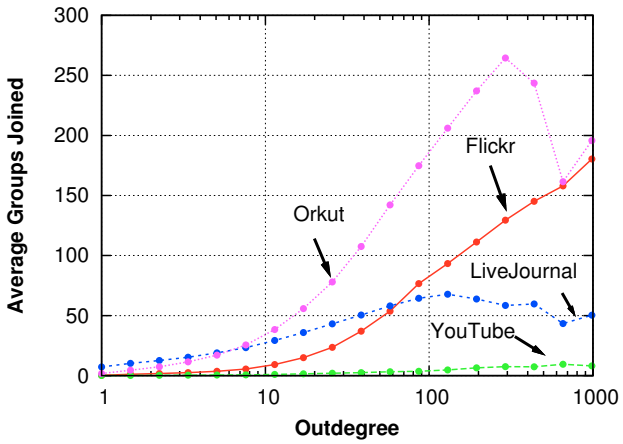


Figure 5.11 : Outdegree versus average number of groups joined by users. Users with more links tend to be members of many groups.

In general, our observations suggest a global social network structure that is comprised of a large number of small, tightly clustered local user communities held together by nodes of high degree. This structure is likely to significantly impact techniques, algorithms and applications of social networks.

5.10 Discussion

We discuss some implications of our findings from this chapter. Our measurements indicate that online social networks have a high degree of reciprocity, a tight core that consists of high-degree nodes, and a strong positive correlation in link degrees for connected users. What do these findings mean for developers? Alternately, how should applications for social networks be designed to take advantage of these properties? Do these properties reveal straightforward attacks on the social structure? Finally, does it make sense to "optimize" algorithms and applications based upon our

findings, since these networks are still growing rapidly and any property we assert now may soon change?

While our findings are likely applicable to many different applications, we concentrate on their effect on information dissemination, search, and trust inference.

5.10.1 Information dissemination and search

Social networks are already used as a means for rapidly disseminating information, as witnessed by the popularity of "hot" videos on YouTube. The existence of a small, well-connected core implies that information seeded via a core node will rapidly spread through the entire network. This is both a strength and a weakness, as spam or viruses could be disseminated this way, as well as important information.

Similarly, searches that proceed along social network links will quickly reach the core. This suggests that simple unstructured search algorithms could be designed if the core users were to store some state about other users. In effect, the users in the core represent "supernodes" in a two-level hierarchy, similar to existing search protocols for unstructured networks, such as Gnutella.

5.10.2 Trust

Social networking sites are the portals of entry into the Internet for many millions of users, and they are being used both for advertisement as well as for the ensuing commerce. Many of these applications, ranging from mail to auctions, implicitly rely on some form of trust. For example, when a user accepts email from an unknown

user, she is trusting the other party not to send spam. When a user selects a winning bidder in an auction, she is trusting the other party to pay the winning amount, and the winning user is trusting the seller to produce the auctioned item.

In a social network, the underlying user graph can potentially be used as a means to infer some level of trust in an unknown user [86], to check the validity of a public key certificate [110], and to classify potential spam [57]. In all of these, trust is computed as a function of the path between the source and target user.

Our findings have interesting implications for trust inference algorithms. The tight core coupled with link reciprocity implies that users in the core appear on a large number of short paths. Thus, if malicious users are able to penetrate the core, they can skew many trust paths (or appear highly trustworthy to a large fraction of the network). However, these two properties also lead to small path lengths and many disjoint paths, so the trust inference algorithms should be adjusted to account for this observation. In particular, given our data, an unknown user should be highly trusted only if multiple short disjoint paths to the user can be discovered.

The correlation in link degrees implies that users in the fringe will not be highly trusted unless they form direct links to other users. The "social" aspect of these networks is self-reinforcing: in order to be trusted, one must make many "friends", and create many links that will slowly pull the user into the core.

5.11 Summary

We end this chapter with a brief summary of important structural properties of social networks which we observed in our data.

- The degree distributions in social networks follow a power-law, and the power-law coefficients for both indegree and outdegree are similar. Nodes with high indegree also tend to have high outdegree.
- Social networks appear to be composed of a large number of highly connected clusters consisting of relatively low-degree nodes. These clusters connect to each other via a relatively small number of high-degree nodes. As a consequence, the clustering coefficient is inversely proportional to node degree.
- The networks each contain a large, densely connected core. Overall, the network is held together by about 10% of the nodes with highest degree. As a result, path lengths are short, but almost all shortest paths of sufficient length traverse the highly connected core.

Chapter 6

Network Growth

To date, most measurement and analysis of online social networks (including the preceding chapter) has focused on the properties of static network snapshots. Despite the different goals and purposes of the various online social networking sites, the underlying social networks have been shown to exhibit a surprising number of common structural features, such as a highly skewed (power-law) degree distribution, a small diameter, and significant local clustering [6, 105]. This intriguing similarity suggests that the same underlying network growth processes may be at play in the different sites.

A proper understanding of these growth processes can provide insights into the observed network structure, allow predictions of future network growth, and enable simulation of systems on social networks of arbitrary size. However, most work on growth processes for large-scale networks has focused on theoretical models, instead of deriving the growth properties from empirical data. For example, two of the popular theoretical growth models are the Barabási-Albert model [15], where users connect to other users in proportion to the destination's popularity, and the random walk model [142, 154], where users connect to other users who are already close in the network.

In this chapter, we use network growth data from two online social networks and two other real-world networks to validate existing models of network growth. In particular, we study how well the empirical data matches the predictions of growth models that have been proposed. As before, we compare our results to those for other, well-understood networks in order to ground our analysis. However, we are well aware that the user graph in social networks is fundamentally different from the interconnection of web pages or the connections between autonomous systems in the Internet.

It is important to note that we can only study how well a particular model predicts the link creation that occurs in the empirical data. We fundamentally do not know why new links were established; we can only observe the source and destination of new links. Thus, we cannot ultimately prove or disprove any particular model; we can only examine the correlation between the observed data and what each model would predict. Nevertheless, knowing how well different models predict link creation in the data can improve our understanding of network evolution, and can provide clues as to the actual underlying processes.

6.1 High-level data characteristics

Table 6.1 shows the high-level statistics of the data we gathered in order to study the growth of large networks at scale. The network sizes vary by over three orders of magnitude. Similarly, other metrics, such as the average number of links per node

and the yearly growth rate also vary greatly between the networks. Despite these differences, as our analysis later shows, the growth of these complex networks shows a number of commonalities.

	Flickr	**Wikipedia**	**YouTube-D**	**YouTube-U**	**Internet**
Network Type	directed	directed	directed	undirected	undirected
Days Observed	104	825	36	165	1,281
Resolution	day	second	day	day	month/week
Symmetric Links	62%	17%	79%	-	-
Initial Nodes	1,620,392	695,353	1,003,975	1,402,949	9,978
Final Nodes	2,570,535	1,892,691	1,137,638	3,218,658	25,526
Nodes Growth	58%	169%	13%	129%	155%
Growth per Year	242%	54%	145%	525%	31%
Initial Links	17,034,807	6,637,456	4,391,336	6,783,917	29,504
Final Links	33,140,018	39,953,145	4,945,382	18,524,095	104,824
Link Growth	63%	500%	12%	173%	255%
Growth per Year	455%	120%	215%	822%	43%

Table 6.1 : High-level statistics of the network growth data.

6.2 Growth dominates network evolution

In all of the networks we examined, we found that link addition was significantly more frequent than link removal. In particular, we found that in Flickr, link additions exceeded link removals in our data sets at a rate of 2.43:1. Similar characteristics were observed in the other networks we studied: in YouTube-U, the ratio of link additions to removals was 3.71:1, and in the Internet, we found that the ratio was 2.06:1. Unfortunately, we did not record removed links for the YouTube-D data set, and we are unable to estimate the fraction of removed links in Wikipedia due to the effects of page vandalism (i.e., vandalized pages often have their entire text, and therefore all of their outgoing links, replaced and then added back).

In summary, in the networks in which we were able to record link removals, we observed that link addition significantly exceeded link removal. Thus, in the rest of this chapter, we focus only on how links are added to growing networks, and we leave examining link removal to future work.

All of the networks we observed showed a high growth rate: normalizing for different observation periods across the networks reveals an average growth rate of between 31% and 525% per year in terms of nodes, and a growth rate of between 43% and 822% per year in terms of links. These rapidly growing networks offer us a unique opportunity to observe new link creation.

6.3 Reciprocation

We begin by first examining *reciprocation*, a growth mechanism that exists only in directed graphs. Reciprocation occurs when the creation of a directed link between two nodes causes the reverse link be established. Since undirected graphs are, by definition, symmetric, reciprocation does not make sense in the context of undirected graphs. Reciprocation has been proposed as an independent growth mechanism for large-scale directed graphs [56, 173].

Since we do not know why links were established, we rely on the timing between the creation of the two directed links of a symmetric link to guess whether the creation of the first causally affected the second. Figure 6.1 shows the distribution of the time between the establishment of the two links of a given symmetric link in the three directed graphs (Flickr, YouTube-D, and Wikipedia) that we studied.

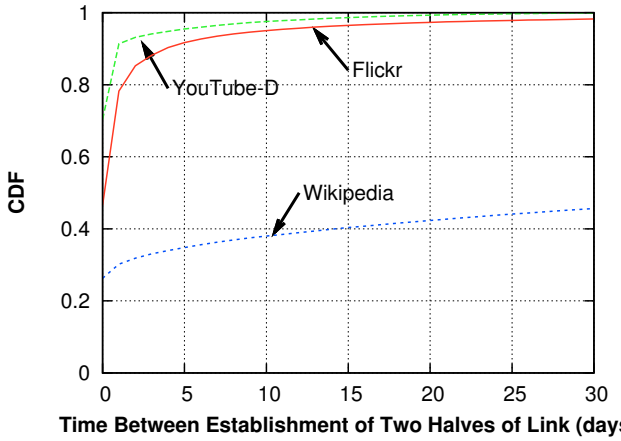


Figure 6.1 : CDF of time between establishment of the two directed links of a symmetric link. In both Flickr and Youtube, links are quickly reciprocated.

From Figure 6.1, it is clear that in the two social networks we observed, users

often respond to incoming links by quickly establishing a reciprocal link back to the source node. In fact, over 83% of all symmetric links we observed in both Flickr and YouTube-D were established within 48 hours after the initial link creation. We hypothesize that this rapid link creation is enabled by the mechanisms on the online social networking sites: most sites email users of new incoming links and provide an easy mechanism for creating a reciprocal link in response.

Thus, our data suggests that users tend to quickly reciprocate links, if they reciprocate at all. It is therefore highly likely that the establishment of the first link in these networks prompted the creation of the reciprocal link. The Wikipedia data, on the other hand, indicates a lower degree of reciprocation; only 30% of the symmetric links in Wikipedia had both halves of the link created within 48 hours of each other.

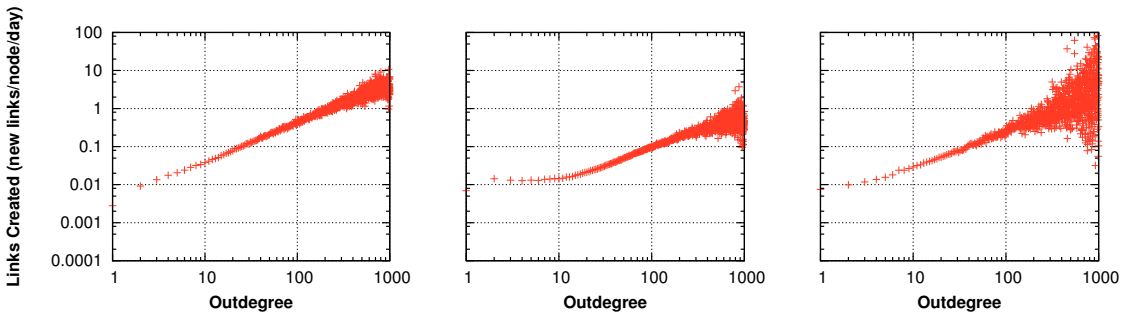
Our data suggests that reciprocation is an independent mechanism shaping the growth of directed networks. The degree of reciprocation is dependent on the network: the two social networks show significant reciprocation, while Wikipedia shows reciprocation, but to a less significant degree.

6.4 Preferential attachment

Preferential attachment [15], colloquially referred to as the "rich get richer" phenomenon, is a growth model in which new links in a network are attached *preferentially* to nodes that already have a large number of links. Under preferential attachment, the probability that a new link attaches to a given node is proportional to the node's

current degree.

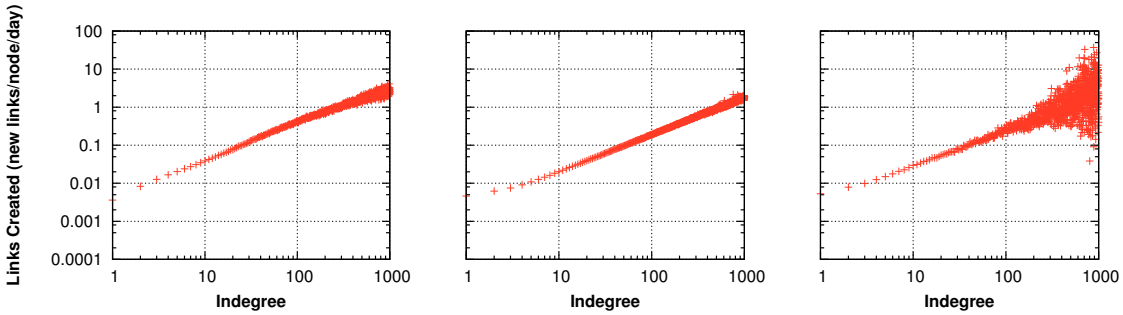
To examine whether preferential attachment predicts the observed growth data, we calculated how the number of new links per day varies with the node degree. If preferential attachment is taking place, we would expect to see a positive correlation between the degree of a node and the number of new links it creates or receives. However, it is important to note that a positive correlation is a necessary but not sufficient condition for the validity of the preferential attachment mechanism, as other mechanisms could also result in such a correlation. For example, the "connecting nearest neighbors" model [154] has been shown to also exhibit such a correlation.



(a) Flickr (b) Wikipedia (c) YouTube-D

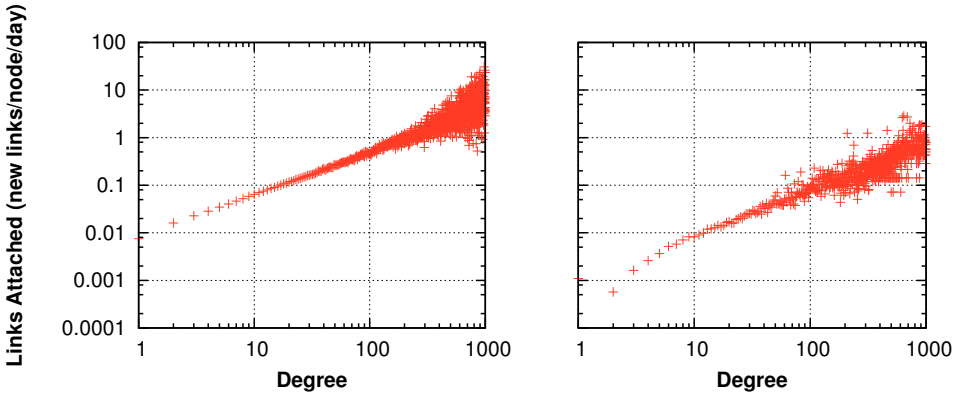
Figure 6.2 : Log-log plot of outdegree versus number of new links per day. All networks show strong evidence of preferential attachment.

Figure 6.4 plots this distribution in log-log scale for each of the five networks we studied. For the three directed graphs, we separately plot the number of new links created and received, with respect to the node's current outdegree and indegree.



(a) Flickr (b) Wikipedia (c) YouTube-D

Figure 6.3 : Log-log plot of indegree versus number of new links per day. All networks show strong evidence of preferential attachment.



(a) YouTube-U (b) Internet

Figure 6.4 : Log-log plot of degree versus number of new links per day. All networks show strong evidence of preferential attachment.

6.4.1 Undirected networks

For the two undirected networks, YouTube-U and the AS-level Internet, we show how the degree of a node correlates with the number of new links per day. We find a strong positive correlation between the current degree and the number of newly created links in both of the networks.

6.4.2 Directed networks

For the three directed networks, we separate the preferential attachment model into two aspects: *preferential creation* and *preferential reception*. Preferential creation describes the mechanism by which nodes *create* new links in proportion to their outdegree, and preferential reception describes the mechanism where nodes *receive* new links in proportion to their indegree. This distinction is consistent with previously proposed models of preferential attachment on directed graphs [25].

It is important to understand why we separate preferential attachment into preferential creation and preferential reception for directed networks. Preferential attachment was originally defined for undirected graphs [15], and therefore does not distinguish between node indegree and outdegree. However, in the directed networks we study, link creation is very different from link reception. Nodes are in complete control over their outgoing links, since they decide who they link to, but they are not in control of their indegree, since it depends upon who they receive links from.

For the three directed networks, Flickr, Wikipedia, and YouTube-D, we separately

examine how the current outdegree and indegree of a node is related to the number of newly created and received links per day. Figure 6.4 shows that the outdegree of nodes is positively linearly correlated with the number of new links created per node per day. This is a necessary, but not sufficient condition for the validity of the preferential creation mechanism. Figure 6.4 also shows, for the three directed networks, that the increase in node indegree is linearly correlated with the current indegree of the node. Similarly, this is a necessary condition for the validity of the preferential reception mechanism.

6.4.3 Discussion

Our data shows that a necessary condition for preferential attachment, a positive correlation between the degree of a node and the number of new links, is present in all five networks. However, this alone is insufficient to claim that any specific mechanism (such as the BA model) is the mechanism that is causing the growth, as a number of different mechanisms could also result in this correlation. In the next section, we more closely examine the growth data to look for further evidence of specific growth mechanisms.

6.5 Proximity bias in link creation

In this section, we take a closer look at our growth data to look for evidence of specific global or local mechanisms that lead to preferential attachment. We look for

evidence of models based on local rules by focusing on the distance between newly-linked users. Specifically, we examine the shortest path distance between the source and destination of newly created links, before a new link is created between them. If, for example, the BA model is the underlying mechanism, then the observed distance distribution between users should match that predicted by the model. Otherwise, if we see a stronger bias towards close users, it may suggest that users follow local, rather than global, rules for selecting the destinations for new links.

Over 50% of the links in all five networks are between nodes that have, a priori, some network path between them (the remainder of the observed new links are between users which are, a priori, disconnected).[1] For these new links among already connected users, Figure 6.5 shows the cumulative distribution of shortest-path hop distances between source and destination nodes. It reveals a striking trend: over 80% of such new links in Flickr connect nodes that were only two hops apart, meaning that the destination node was a friend-of-a-friend of the source node. Similarly, this fraction is over 42% in YouTube-D, over 50% in Wikipedia, over 45% in YouTube-U, and over 57% in the Internet topology.

One might wonder whether in small diameter networks like the ones we observe, this high level of proximity in link establishment is simply a result of preferential attachment. This is plausible, since the high-degree nodes that preferential attachment prefers tend to be close to many nodes. To test this hypothesis, for each newly

[1]For directed networks, we only count directed paths.

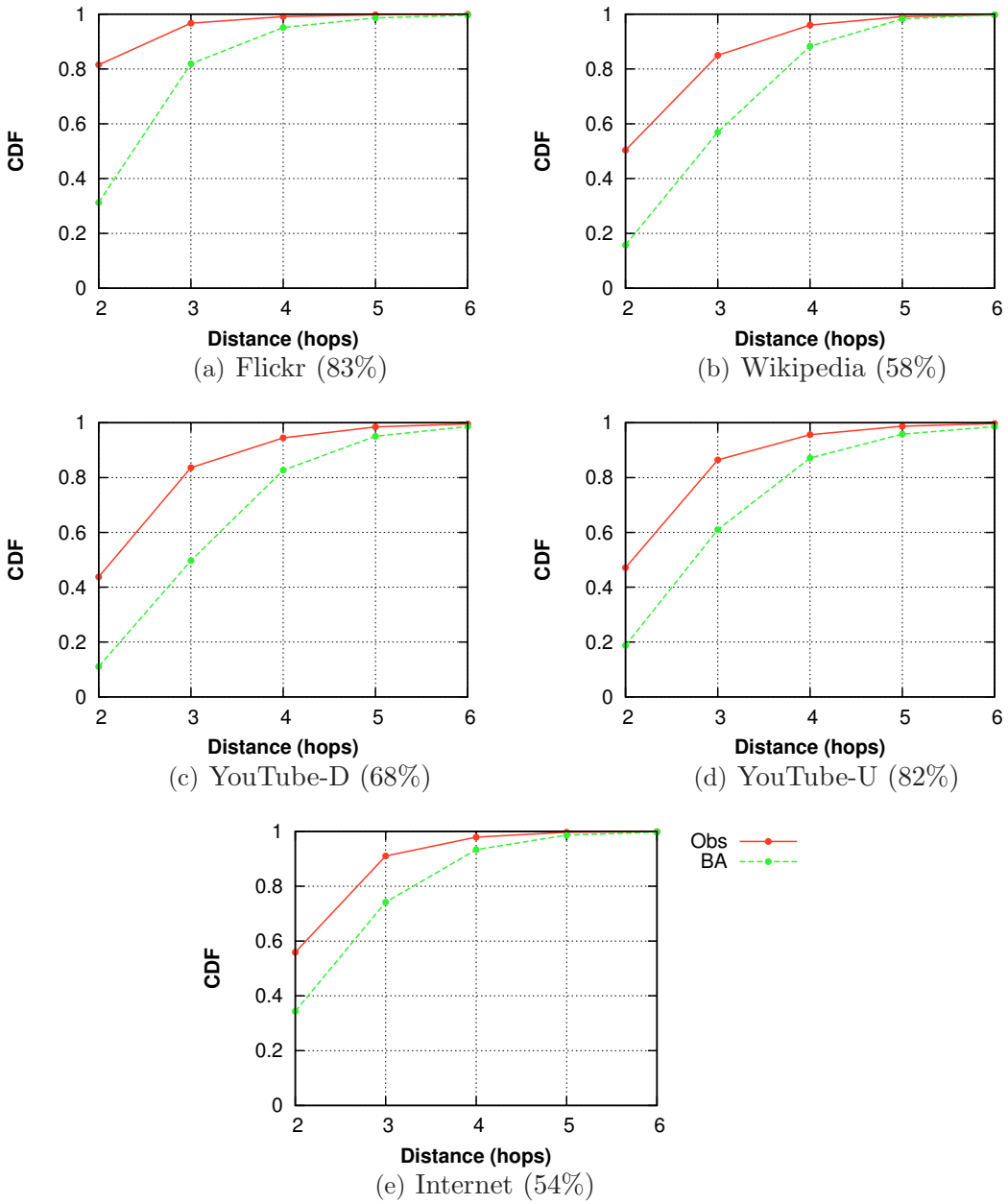
created link, we computed the expected distance from the source to the destination, if the destination is chosen using the BA model. Figure 6.5 also plots this distribution for each network.

In all five networks that we study, the observed distances between the source and destination of links shows a significant bias towards nearby nodes, relative to what the BA model would predict. In fact, in Flickr, Wikipedia, and YouTube-D, we found that the number of new links connecting 2-hop neighbors in the empirical data exceeded that predicted by the BA model by a factor of three.

This result shows that while new link formation in our observed networks follows preferential attachment, the link creation process cannot be explained by the BA model alone. Nodes are far more likely to link to nearby nodes than the model would suggest. This result is consistent with the previous observations on static networks, which showed that the clustering coefficient was significant higher than would be predicted by the BA model. In the next section, we focus on how nodes choose which nearby node to link to.

6.6 Mechanisms causing proximity bias

In the previous section, we showed that newly created links show a strong bias towards nodes which are close together, relative to what the BA model would predict. This suggests that an alternate mechanism is causing the establishment of new links. In this section, we take a closer look at the newly created links, and see if the growth



(a) Flickr (83%) (b) Wikipedia (58%)

(c) YouTube-D (68%) (d) YouTube-U (82%)

(e) Internet (54%)

Figure 6.5 : CDF of distance between source and destination of observed links (Obs). Also shown is the expected CDF from BA model (BA). The numbers in parenthesis are the fraction of all new links connecting nodes that had, a priori, some path between them. All networks show a proximity bias that is not predicted by the BA model.

data matches the expected properties of other proposed mechanisms.

In particular, we examine network growth models that are known to have a stronger bias towards proximity than preferential attachment. To make the analysis tractable, we focus on new links that occur between nodes that are two hops apart. Such links account for over 45% of the links in all networks. We consider the BA model for preferential creation, combined with five different proposed mechanisms for selecting the destination of a newly established link:

- *Random selection (RS)*, where a node choses the destination randomly from its set of two-hop neighbors. This mechanism serves as a baseline for evaluating the other mechanisms.

- *Random two-hop walk (RW)*, where a node performs a random two-hop walk to find the destination [154].

- *Preferential selection (PS)*, where a node choses from its set of two-hop neighbors preferentially according to the nodes' indegrees. This is similar to the BA model, except that a node only considers its two-hop neighbors [93].

- *Common neighbors (CN)*, where a source makes a weighted random choice among its set of two-hop neighbors. The likelihood that a given candidate is chosen is proportional to the number of neighbors the source shares with the candidate [114].

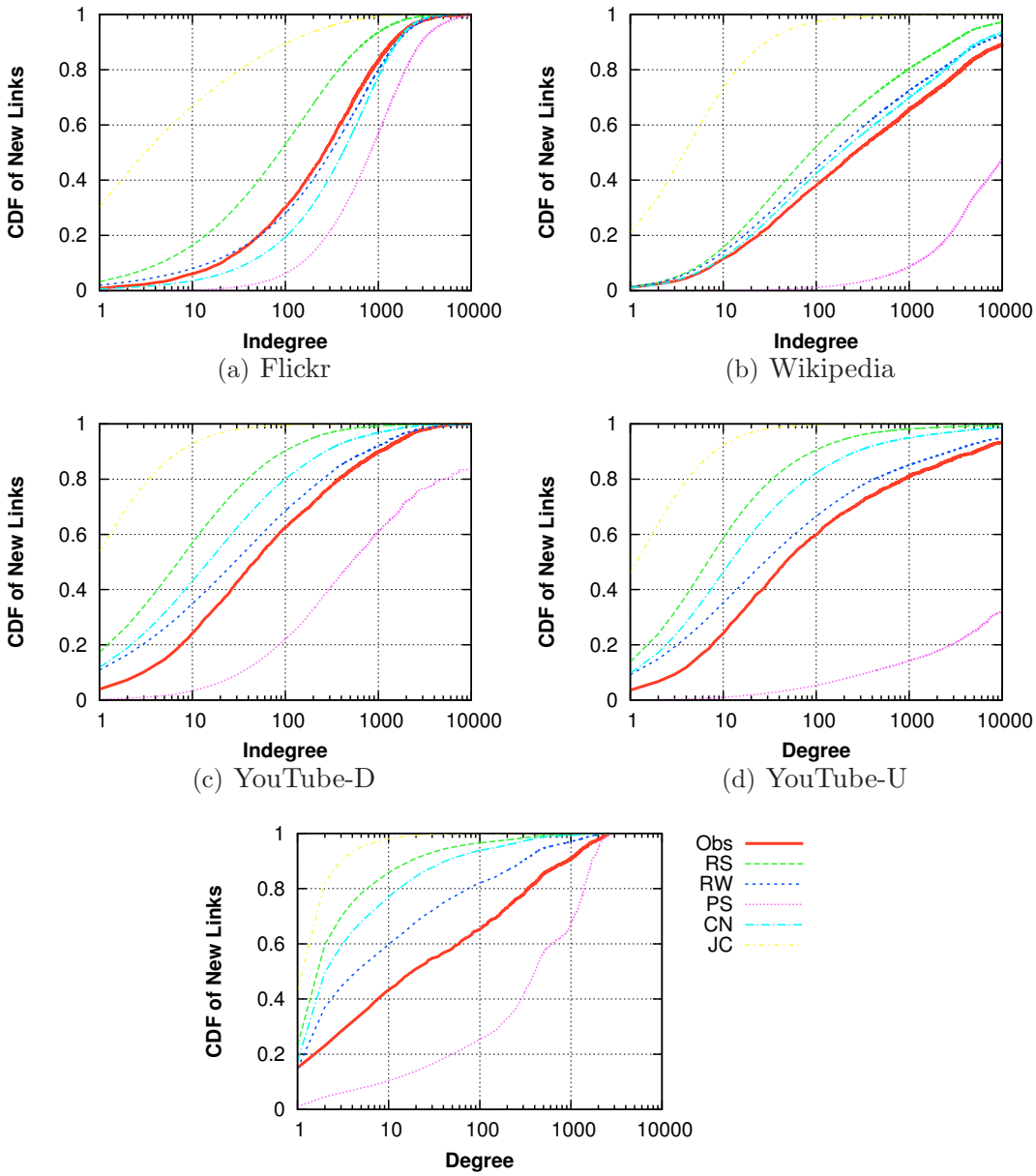
- *Jaccard's coefficient (JC)*, where a source makes a weighted random choice

among its set of two-hop neighbors. Here, the likelihood that a given candidate is chosen is proportional to the number of neighbors the source shares with the candidate divided by the candidate's indegree [93].

We examined newly established links in all networks that connect nodes that were previously two hops apart. We then calculated the expected indegree distribution of nodes that would have been selected using each of the five mechanisms above. We then compared the results to the distribution in the empirical data. Figure 6.6 plots these distributions for each network.

From Figure 6.6, we can see that no one mechanism closely matches the empirical data in all networks. In fact, in two of the networks (Flickr and Wikipedia), the random walk mechanism most closely matches the observed data. However, in the other three networks, the results are less conclusive. To better quantify how well the various mechanisms predict the selected destination of new links, we calculated the accuracy of each mechanism, in the same manner as previous studies [93]. Thus, for each newly created link, we calculated the fraction of time each mechanism correctly predicted the selected destination. The results are shown in Table 6.2, relative to the random selection model.

The accuracy results in Table 6.2 shows that no one model dominates in terms of accuracy across different networks. However, closely examining the results reveals that the two mechanisms that take into account the indegree of the destination (RW and PS) do tend to have higher accuracy. This suggests either that different mecha-



(a) Flickr (b) Wikipedia

(c) YouTube-D (d) YouTube-U

(e) Internet

Figure 6.6 : CDF of nodes receiving new links by indegree. Plots are shown for observed data (Obs), and simulated mechanisms: random selection (RS), random 2-hop walk (RW), preferential selection (PS), common neighbors (CN), and Jaccard's coefficient (JC). The observed data does not match any one mechanism, suggesting that different mechanisms are at play in different networks.

	RS	RW	PS	CN	JC
Flickr	0.17%	2.0	1.1	1.2	1.2
Wikipedia	0.15%	2.9	2.9	1.3	0.7
YouTube-D	0.35%	1.6	1.5	1.1	1.0
YouTube-U	0.59%	1.7	1.3	1.1	1.4
Internet	0.53%	1.9	4.1	1.1	0.5

Table 6.2 : Prediction accuracy of two-hop link creation mechanisms relative to the baseline random selection mechanism. While no one mechanism appears to be the most accurate across all networks, Random Walk and Preferential Selection tend to have higher accuracy.

nisms may be at play in different networks, or that the actual mechanism driving link creation is not among the ones we evaluated, or that the actual mechanism is a complex combination of some of the mechanisms we tested. This result is not surprising, though, as each of the networks represents a different system, and it is unlikely that one single mechanism would describe the link creation behavior in all of them.

6.7 Discussion

In this chapter, we used empirical growth data from multiple large-scale complex networks to test if previously proposed growth models actually are at play in these networks. We have chosen to focus on the well-known BA model model because it is simple and has been suggested as the underlying growth mechanism in different contexts. Clearly, the BA model leads to global degree distributions of the type

observed in many diverse networks, and absent other data, it is an attractive choice for researchers to explain static snapshots of crawled networks.

6.7.1 Is proximity fundamental?

We believe that some notion of proximity is inherent in the link creation processes underlying large networks. As a network grows larger, it is increasingly unlikely that nodes are influenced by knowledge of the global degree ranking when choosing their neighbors. In many networks (in particular, many social networks), it may not even be possible to discover the global degree ranking of nodes, knowledge of which is required for pure preferential attachment. Other mechanisms that rely on global properties are equally unlikely because of technical and policy issues with computing global metrics.

In the networks we have examined, the bias towards proximity can be explained by considering the node discovery mechanisms available to users and the factors that constrain them. In the social networks (YouTube and Flickr), the primary mechanism available to users for exploring the network is to walk their neighborhood. This might explain our observation in Flickr and YouTube that there is a much stronger bias in link creation towards nearby nodes than would be predicted by preferential attachment alone, yet there still is a bias towards high-degree nodes (see Table 6.2). On Wikipedia, semantically closer pages are likely to be proximal in the network, leading to a proximity bias in link creation.

The Internet AS graph is fundamentally different because each AS consists of many different routers and there is a significant cost associated with creating new links. A model for AS link formation is given in [27], and our observations are consistent with the reasoning therein. The AS graph is naturally "tiered" with many small stub ASes interconnected by a few large backbone providers (who also tend to have high connectivity/degree). AS link creation is often constrained by financial, technical, and geographical factors: for most stub ASes, links to far away ASs tend to be costly (especially if the geographic distance is large) and are unlikely to be profitable since the upstream provider already provides transit to reach these ASes. Such links only make sense in specific cases where business relationships mandate a specific inter-AS peering. Thus, stub ASes tend to connect to their nearby backbone AS providers, and the resulting AS graph shows proximity bias coupled with strong preferential selection.

6.7.2 Proximity mechanisms

While our growth data cannot *assert* which mechanism are at play when links are formed, it can be used to *disprove* existing hypotheses. Perhaps unsurprisingly, we find that the simplest mechanisms (such as the BA model) are not sufficient to explain our observations. In particular, we have shown that to explain the empirical growth data, we must include some notion of proximity in the growth models. While proximity has been previously suggested as a factor in link creation in large networks, we

believe we are the first to provide empirical date from multiple large-scale networks to support this conjecture.

The analysis in the previous section revealed some insights into how proximity affects the growth of complex networks. While our results are not conclusive, it appears that growth models that take into account the indegree of the destination (e.g. Preferential Selection and Random Walk) match the data more closely than other models. Moreover, Preferential Selection outperforms Random Walk only for the Internet AS graph.

6.8 Summary

In this chapter, we closely examined network growth data from five different networks and compared the empirical data to the predictions of previously proposed growth mechanisms. We end this chapter with a brief outline of our most important findings.

- We found evidence of reciprocation as a mechanism in directed networks. We found that users tend to often create a reciprocal link in response to an incoming link, explaining the high levels of symmetry observed in social networks with directed links.
- We also found that nodes tend to create and receive links in proportion to the outdegree and indegree, which is consistent with preferential attachment (or preferential creation and preferential reception in directed networks).
- However, we found that the BA model alone did not accurately predict the

proximity bias among nodes connected by new links in any of the empirical data sets. All networks showed a stronger bias towards proximity between new sources and destinations than would have been predicted by the BA model.

- Upon closer examination of the newly created links links, we found than no single proximity model we examined appears to accurately predict this proximity across all networks. However, we did find that models that consider network proximity as a factor in link creation predict the empirical data better then preferential attachment. This suggests that further research into growth mechanisms is necessary.

Chapter 7

Network Communities

The concept of a *community* is central to online, as well as offline, social networks. A community is a subset of the users in a social network that is more tightly interconnected than the overall network [118]. Communities are interesting for a variety of reasons. For example, users in a community tend to interact frequently, often share interests, and trust each other to some extent. Therefore, communities are useful, for instance, to guide information dissemination and acquisition, to recommend or introduce people who would likely benefit from direct interaction, and to express access control policies.

Prior works have proposed algorithms for automatically detecting communities in social networks [13, 31, 58, 99, 118, 131, 153]. However, the algorithms have never been tested on real online social networks at scale. In this chapter, we use fine-grained data from a university online social network to study the effectiveness of existing algorithms for detecting communities, and we propose a new algorithm to overcome the observed limitations of existing approaches.

Specifically, we make three contributions. First, we collect detailed data about a large university social network and analyze the structure of communities in the network. Our data covers almost 4,000 students and alumni of Rice University taken

from the Facebook [49] social network. For each student, we gather attributes like major(s) of study, year of matriculation, and dormitory, to see if communities in the network align with these attributes. We find that users tend to form links to other users who share the same attributes, and that users who share certain attributes define strong communities in the social network.

Second, we examine how well existing techniques can detect communities. We find that existing approaches often perform poorly on our data set, sometimes returning a large part of the network (or the whole network) as a community. We demonstrate that this poor performance is due to the use of community-rating metrics that are biased towards large communities.

Third, we propose and evaluate a new algorithm that can accurately infer memberships of multiple, potentially overlapping communities, when given information about a small subset of the community members. In practice, this means that if even as few as 20% of users provide community information to social networking sites, the remaining members of the community can be determined from the social network alone with high accuracy.

In the following sections, we describe the data we use for our community analysis. We then examine the data set, looking at the correlation between attributes and the links and communities in the network. Finally, we evaluate previously proposed approaches and propose and evaluate new approach for detecting communities in the network

7.1 Data sets used

In this chapter, we use the Facebook data set from Rice University, described in Section 4.5. We partition our data set into a two subsets representing different parts of the Rice University network, which have different properties. The first group we use is the current undergraduates. This subset contains 1,233 users connected with 86,416 links, for an average degree of 70.1. The second group we use is the current graduate students. This subset contains 548 users connected with 6,512 links, for an average degree of 11.8. We examine these two parts of the network separately, since we have different attributes sets for the undergraduates and graduate students and they represent largely distinct parts of the network. In fact, only 1,455 links are present between undergraduates and graduate students.

7.2 Attributes in the network

We first make two observations about how the structure of the social network is correlated with the attributes of users. First, we note that users are significantly more likely to be friends with other users who share their attributes. In some cases, the likelihood is as high as 10-fold more than that would be expected if links were placed randomly. Second, we observe that this affinity for links between similar users leads to communities of users in the network that are centered around attributes. Each of these observations is described in detail below.

7.2.1 Friends with common attributes

Our first observation is that users are statistically much more likely to be friends with other users who share their attributes. In order to show this, we calculated for each attribute a (such as college or matriculation year)

$$S_a = \frac{|\{(i,j) \in E : \text{ s.t. } a_i = a_j\}|}{|E|} \tag{7.1}$$

where a_i represents the value of attribute a for user i, and E represents the set of all links. S_a therefore represents the fraction of links for which users share the same value of attribute a. Finally, we divided this by what would be expected in a graph with a similar distribution of attributes but with the links placed randomly between users. The resulting value, which we call *affinity*, ranges from 0 to ∞ and represents the ratio of the fraction of links between attribute-sharing users, relative to what would be expected in a random graph. Thus, an affinity greater than 1 indicates that links are positively correlated with user attributes.

Table 7.1 shows the affinity of the various attributes for the undergraduates and graduate students at Rice. We observe that for all attributes, a significant affinity is observed, showing that links in the Rice network are correlated with attributes. It is interesting to note that certain attributes are stronger than others: for example, graduate students have a much strong affinity for other students in the same department when compared to other students in the same matriculation year. In some cases, the affinity is as high as 10, implying that users connected by a link are 10 times more likely to share an attribute that would be expected in a random graph. In summary,

Users	Attribute	Affinity
undergrads	college	5.77
	major	2.37
	year	1.93
grads	department	9.98
	school	4.09
	year	1.81

Table 7.1 : Affinity values for various attributes of students at Rice. Links are correlated with numerous user attributes.

we have observed that links in the Rice network are strongly correlated with attribute values, suggesting that communities of users centered around common attributes may be present.

7.2.2 Attribute-based communities

Given that we have observed a correlation between user attributes and links, it is natural to see if the users who share a similar attribute form communities, or dense clusters, in the network. Note that the previous observation is a necessary, but not sufficient, condition for attribute-based communities to exist, since users with common attributes may be linked together but may not form a dense community. In order to investigate whether attribute communities are present in our network, we artificially divide the network into communities based on user attributes, and then quantify the strength of that division into communities using modularity [118].

Undergraduate students

Table 7.2 shows the modularity for the undergraduate population when partitioned according to residential college, major, and matriculation year. Also shown is the modularity of the partitionings that are obtained when multiple attributes are used. The results show a significant modularity for the communities defined by residential college and matriculation year – a relatively high Q of 0.385 is observed when partitioning by residential college, and a Q of 0.259 is seen when dividing by year. However, the modularity of the communities defined by major is almost 0, indicating that no community structure exists based on academic major. Overall, these results indicate that users who share the same college or matriculation year form tightly-knit communities in the social network.

Attributes	**Communities**	**Modularity**
college, major, year	660	0.021
college, major	488	0.025
year, major	270	0.039
major	163	0.046
college, year	36	0.249
year	4	0.259
college	9	0.385

Table 7.2 : Modularity values for communities defined by various attributes of undergraduates at Rice. College and matriculation year reveal strong community structure.

With some knowledge of the actual social network at Rice, the above results are

not unexpected. Undergraduate students are randomly assigned to a residential college upon matriculation, and they remain members of that college for the duration of their undergraduate studies. Thus, it is natural that strong communities form around residential colleges. Additionally, the strong communities among undergraduate students of the same matriculation year are not surprising. Incoming students attend an orientation week together, are mostly assigned to share dormitory rooms with students of their year, and tend to spend time in courses with students of their year. Thus, it is also natural that a community structure exists among undergraduates of the same matriculation year. Finally, the lack of a strong community structure around majors can be explained by the fact that Rice undergraduates obtain liberal arts education (taking courses from many departments), and they often do not choose majors until the end of their sophomore year.

Graduate students

We now turn our focus to the graduate student population. Table 7.3 shows the modularity of the graduate student population when partitioned according to department, academic school, and matriculation year.[1] The results show a significant modularity for the communities based on department – in fact, a Q of 0.586 is observed. A similar modularity is observed when partitioning according to school – this is because each department is a member of exactly one school, and the partitioning

[1]Note that graduate students are not assigned to residential colleges, so that attribute is disregarded here.

according to school ends up being a coarser version of the communities defined by department. Similar to the undergrads, a Q of 0.187 is also seen for the communities defined by matriculation year. This indicates a very strong community structure for the graduate students based on department, and a weak community structure based on matriculation year.

Attributes	**Communities**	**Modularity**
year	11	0.187
department, school, year	139	0.294
department, year	139	0.294
school, year	45	0.304
school	9	0.583
department, school	36	0.586
department	36	0.586

Table 7.3 : Modularity values for communities defined by various attributes for graduate students at Rice. Departments form strong communities.

The results for the graduate student population are also not unexpected. Graduate students are accepted into a specific department at the beginning of their studies, and usually spend their entire tenure in the same department. Thus, the very strong association with the department is not surprising. Moreover, the variable length of graduate programs and the greater tendency of graduate students to interact across seniority levels explains why the partitioning according to matriculation year has a weak community structure.

7.2.3 Summary

In both the Rice undergraduate and graduate student populations, we observe that users with similar attributes tend to be friends in the social network. Moreover, we observe a significant community structure, indicated by a high modularity value, for the communities defined by users who share certain attributes. We also observe that multiple overlapping community structures exist. For the undergraduates, we observe significant modularity when partitioning according to residential college and matriculation year. For the graduate students, we observe significant modularity when partitioning according to department and a weaker modularity according to matriculation year.

7.3 Detecting communities

In the previous section, we observed that the undergraduate network, and to a lesser extent the graduate network, contained communities that were correlated with multiple attributes. For the undergraduates, the partitionings according to college and matriculation year both showed significant correlation with the communities in the network. For the graduates, partitionings according to department and year showed similar behavior. We now consider the use of automatic clustering algorithms to detect a specific community among the multiple communities that exist.

To do so, we split the problem into two parts: first, if partial membership information about all communities in the network is known, we examine the problem

of detecting a specific community partitioning. Second, if partial membership information about only one community is known, we look at the problem of detecting a specific community given a partial membership list.

7.3.1 Global community detection

We assume that some fraction of the user population provides information about which communities they belong to. For example, some users on Facebook list their college and matriculation year in their profile. This information can be used to aid the automatic clustering algorithms.

To evaluate whether this information can aid in identifying multiple community structures, we modified the Clauset [32] algorithm to take in attributes of a subset of the users. Instead of starting with every user in their own cluster, the algorithm pre-assigns users with the same attribute into the same cluster. We then run the algorithm as normal, effectively "seeding" it with the users who reveal their attributes. Finally, we calculate the modularity of the resulting partitioning, and then compare it to the partitioning based on the attributes of all users.

To measure how similar these two community structures are, we use the *normalized mutual information* metric [53]. This metric is calculated as

$$\frac{-2\sum_i\sum_j \mathbf{x}_{ij} log(\frac{\mathbf{x}_{ij}N}{\mathbf{x}_{i.}\mathbf{x}_{.j}})}{\sum_i \mathbf{x}_{i.} log(\frac{\mathbf{x}_{i.}}{N}) + \sum_j X_{.j} log(\frac{\mathbf{x}_{.j}}{N})} \tag{7.2}$$

where $\mathbf{x}$ is a square matrix whose dimension is the number of communities detected. Each element $\mathbf{x}_{ij}$ represents the number of nodes in attribute-defined community i

that appeared in the detected community j. $\mathbf{x}_{.i}$ and $\mathbf{x}_{i.}$ denotes sum over column i, and sum over row i respectively, and N is the number of nodes in the graph. At a high level, the metric ranges between 0 and 1, with 0 representing no correlation between the two community structures, and 1 representing a perfect match.

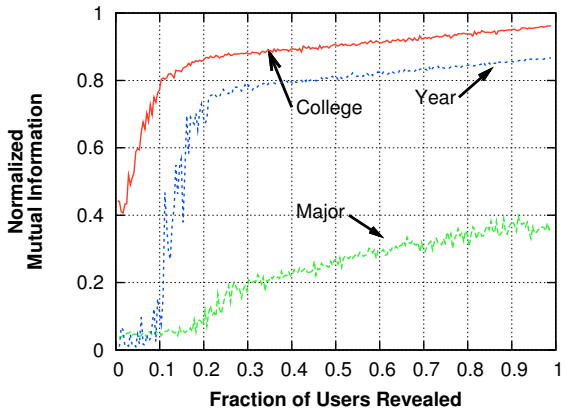


Figure 7.1 : Normalized mutual information versus the fraction of users who reveal their community for Rice undergraduates. Revealing more information naturally leads to partitionings with higher correlations, especially for the college and year attributes. This result shows that different attributes can be accurately inferred with as few as 20% of users revealing their attributes.

Figure 7.1 plots the results of this experiment for the undergraduates, by showing the normalized mutual information for each attribute. Separate lines are plotted for each attribute, and the correlation value is with respect to the attribute that users are revealing. Two trends can be seen in this graph. First, we observe that both college and year quickly lead to community structures with significant correlation. In fact, when just 20% of users reveal their college or year, we can infer the attributes for the remaining users with over 80% accuracy. Second, this is not the case for major of study. However, this result is not surprising, as we observed in the previous section

that communities are not formed around users with common majors. Overall, this experiment shows that multiple attributes can be inferred globally when as few as 20% of the users reveal their attribute information.

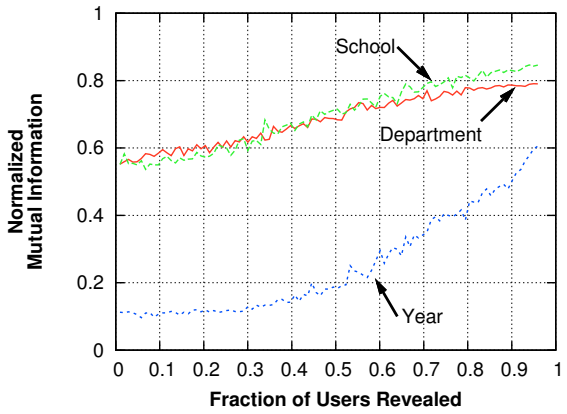


Figure 7.2 : Normalized mutual information versus the fraction of users who reveal their community for Rice graduate students.

Figure 7.2 plots the results of this experiment for the Rice graduate students. Similar to the undergrads, we observe that certain attributes correspond to communities that can be detected with high accuracy. For example, if as few as 5% of the students reveal their department or school, we can infer the department or school for the remaining students with approximately 60% accuracy. However, this is not the case for the matriculation year attribute. We observed in the previous section that matriculation years only correspond to weak communities, so this result is not unexpected.

7.3.2 Local community detection

We now look at detecting communities on a local scale. This is different from the problem in the previous section, where we assumed that partial information about all users in the network is known. Instead, for example, we may know that a subset of five users all live in the same dormitory, and we wish to determine the other users (for which we do not have any information) who also live in that dormitory. To detect these communities, we extend the previously proposed approaches for local community detection to take a seed set of nodes.

While exploring local community detection, we found that previous approaches performed well when detecting certain attributes, but did not perform well on others. For example, we found that the algorithm of Luo [99] could infer the members of a residential college at Rice, but was not able to infer the members of larger communities, such as all students in the same matriculation year. Thus, we propose a new method for detecting a single community, based on the metric of *normalized conductance*. We first describe this new metric below, followed by a description of our algorithm, and finally evaluate the algorithms on our Rice data set.

Normalized conductance

We first define a metric that rates the quality of a single community (as opposed to modularity, which rates the community structure of a partitioning of a graph into a collection of communities). To provide a measure for the quality of a community,

we propose a metric based on the widely adopted metric conductance [71]. Let $G = (V, E)$ denote a graph, let $A \subset V$ be a subset of the vertices that forms a community, and let $B = V \setminus A$. Let us also define e_{AB} to be the number of edges between A and B and e_{AA} as the number of edges within A. The conductance of A is then traditionally defined as e_{AB}/e_{AA}. Therefore, a small value of conductance denotes a strong community, as the community would be tightly linked internally, with very few links to the rest of the graph.

However, this definition of conductance is not a good measure for the "goodness" of a community, as it is biased towards large communities. For example, if we place all the vertices in the graph in a single community, the conductance would be 0, which does not provide any information about the community formed.

Hence, we propose a new metric called *normalized conductance*. To derive normalized conductance, we first define the value K of community A as

$$K = \frac{e_{AA}}{e_{AA} + e_{AB}} \tag{7.3}$$

This value is similar to conductance, except that it ranges between 0 and 1. A measure close to zero indicates very poor community structure, and a measure close to 1 indicates very good community structure with many more links within A than to the outside. However, this metric is still not perfect, as very large communities are naturally biased towards having many more edges within the graph (high e_{AA}). Thus, we define the normalized conductance C for a community A as K minus the expected value of K for a random graph with the same communities A and B.

To calculate the expected value of K for a random graph, we need to calculate the expected values of e_{AA} and e_{AB} for a graph with the same community division and degree distribution, but with the links placed without regard for the communities. We define $e_A = e_{AA} + e_{AB}$ and $e_B = e_{BB} + e_{AB}$, with e_A denoting the number of edges that reach vertices within A, and e_B giving the same quantity for B. In a random graph, we would expect that $e_{XY} = e_X e_Y$. Thus, our normalized conductance metric C can be written as

$$C = \frac{e_{AA}}{e_{AA} + e_{AB}} - \frac{e_A e_A}{e_A e_A + e_A e_B} \tag{7.4}$$

The metric C ranges between -1 and 1. Similar to modularity, a value of 1 indicates a significant community structure in A, a value of 0 indicates no more community structure than a random graph, and a value of -1 indicates less community structure than a random graph. One particularly useful property of this definition of conductance is that it is comparable across communities of different sizes and densities. Previous definitions generally only use the ratio of intra-community links to inter-community links, which is naturally biased towards very large communities.

Algorithm

We now describe our algorithm for detecting a single community, using the normalized conductance metric C. We assume the algorithm is given as input a subset of users S in a community and the social network graph $G = (V, E)$. The algorithm then returns the other members of the community. Similar to the approach that was taken

by Luo [99], we use a greedy approach to maximize the normalized conductance. We initially divide the graph into two components A and B, with $A = S$ initially. At each step, we select a user $v \in V$ in B that upon adding v to A yields the highest increase in the normalized conductance C for A. We repeat this process, adding users to A, until no remaining user would produce an increase in the normalized conductance C for A. At this point, we stop and return the community A as the result.

The primary difference between our method and the previous approaches is the use of a metric that is weighted against a random graph. We found that the metrics used by previous approaches are all biased towards large communities. For example, the metric used by Luo et al. [99] is based on the ratio between the number of intra-community links to the number of inter-community links. As a community grows larger, this value naturally increases; in fact, it becomes infinite if an entire connected component is viewed as a community. Thus, these approaches often have trouble detecting large communities in the network, as they often proceed to detect the entire graph as a community. By weighting our metric against a random graph, we can detect both the small-scale and large-scale communities that exist.

Evaluation

To see how well our algorithm and others perform, we evaluate the performance along two axes. Assume that each algorithm takes as input a subset S of users with attribute H, and the social network graph. The algorithm then returns a set of users

R, representing the other members it believes also have attribute H, based on the community structure in the network. We define the *recall* to be

$$\frac{|R \cap H|}{|H \setminus S|} \tag{7.5}$$

representing the fraction of the remaining community members that the algorithm returns. Similarly, we define the *precision* to be

$$\frac{|R \cap H|}{|R|} \tag{7.6}$$

representing the fraction of the returned users who are actually in the community. Thus, an ideal algorithm would have a recall of 1 (returning all of the remaining users) as well as a precision of 1 (only returning users who are actually in the community).

We now evaluate our algorithm on the Rice data set along with the algorithms of Luo [99], Bagrow [13], and Clauset [31]. First, we examine how well they perform on the undergraduate population by providing the algorithms with varying-size subsets of the students with common attributes such as college, matriculation year, and major. For each attribute (i.e., each college, each major), we select 20 random subsets of users of each size. We then evaluate how well the algorithms perform when given each of these random subsets as input.

For fair comparison with the other algorithms, a few parameters and modifications were required. First, none of the other algorithms accept as input a set of seed nodes; we naturally extended them to start with a set of nodes rather than a single node. Second, the algorithm proposed by Clauset does not specify a stopping condition; instead, it requires the user to specify the number of nodes to be added to

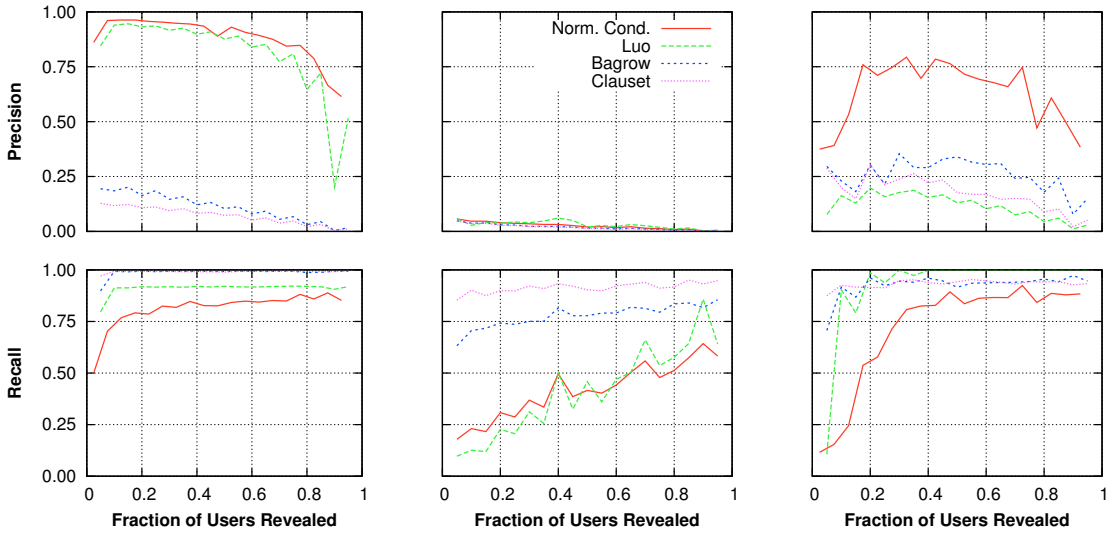
the community. Thus, we utilise the stopping condition proposed by Bagrow [13] for the Clauset algorithm, based on *p-strong* communities.[2] We evaluate the algorithms of Clauset and Bagrow with values of $p = \{0.75, 0.8, 0.85, .., 1.0\}$, as suggested, and select the one with the lowest number of inter-community edges (representing the "best" community). Third, the algorithm of Lou et al. performs iterative adds and deletions, and could therefore remove the original seed nodes from the resulting community. In the case of a single seed node, the authors view the removal of the seed node from the returned community as a failure of the algorithm to detect a community. In order to handle this case for our extended version that accepts a set of seed nodes, we imposed the constraint that we only consider the algorithm of Luo to have found a community if 50% or more of the original seed nodes were present in the resulting community. If not, we do not consider the algorithm to have found a community.

Detecting undergraduate communities

We now present the results for inferring different attributes for the undergraduate students. For these results, we average over all possible values of each attribute (such as all colleges) into the recall and precision data presented in Figure 7.3. Thus, we feed each algorithm $x\%$ of every college and calculate the recall and precision of the

[2]A community is *p*-strong when a fraction p of nodes within the community satisfy the criteria that they have more neighbors inside the community than outside

result. We repeat this experiment five times for each college and fraction revealed, and then average over all colleges to obtain the data in Figure 7.3 (a).



(a) College (b) Major (c) Year

Figure 7.3 : Recall and precision of single community detection for Rice undergraduates for multiple algorithms. Good performance is observed for our algorithm (Norm. Cond.) for college and year; detecting users with the same major performs poorly due to the low correlation with communities in the network. The algorithm of Luo performs well at inferring college but does not perform well for inferring matriculation year.

As a detailed example, Figure 7.4 presents the recall and precision for each of the matriculation years as different number of users are revealed. A number of interesting observations can be made about the results. First, the performance varies across the different matriculation years; the freshmen and sophomores appear to be the easiest to detect, followed by the juniors and seniors. Second, detecting all of the matriculation years shows good performance once 20% to 30% of the users is revealed. Third, note that the precision naturally deteriorates once very high fractions of the users in each

year are revealed. This is because the precision is defined based on the number of unrevealed users, which becomes much smaller as significant fractions are revealed. We now turn back to Figure 7.3 and discuss each attribute in detail.

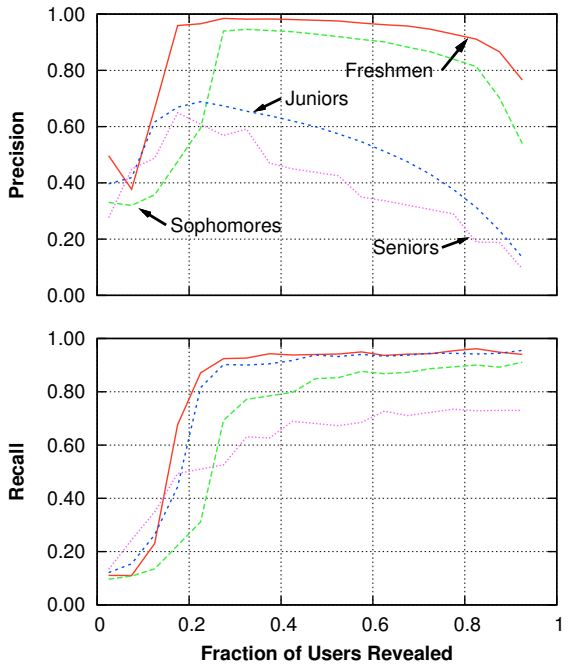


Figure 7.4 : Recall and precision for matriculation year community detection for Rice undergraduates for our algorithm. Individual lines are shown for each matriculation year. Certain values of user attributes are easier to detect than others.

Colleges: The results show that colleges can be inferred with very high recall and precision by both our algorithm and the algorithm of Luo when as few as 10% of the students in the college are known. For example, when 20% of the members of a single college are provided to the algorithms, both our algorithm at that of Luo can infer over 80% of the remaining members of that college with over 95% accuracy. Figure 7.5 shows this in detail for our algorithm, focusing on the performance when

between 1% and 16% of the college members are provided. The algorithms of Clauset and Bagrow both perform rather poorly at detecting colleges: they each often return a large part of the network as belonging to the college, resulting in a very low precision score.

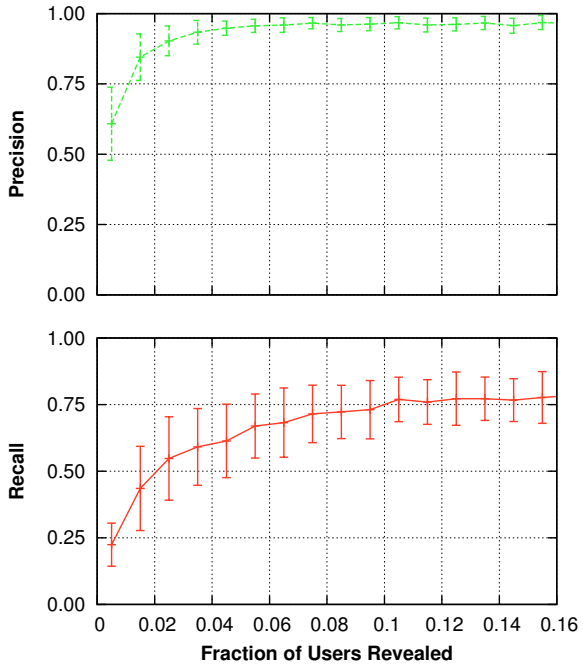


Figure 7.5 : Detail on recall and precision for college inference for Rice undergraduates with our algorithm.

Years: However, for inferring matriculation years, all algorithms have significant recall, but only our algorithm has good precision. In fact, the other algorithms tend to detect the entire graph as a community, which leads to the low precision. Again, we believe that this poor performance is a function of the metrics that the other algorithms use. Since they essentially try to maximize the ratio of intra-community links to inter-community links, they occasionally end up returning the whole graph.

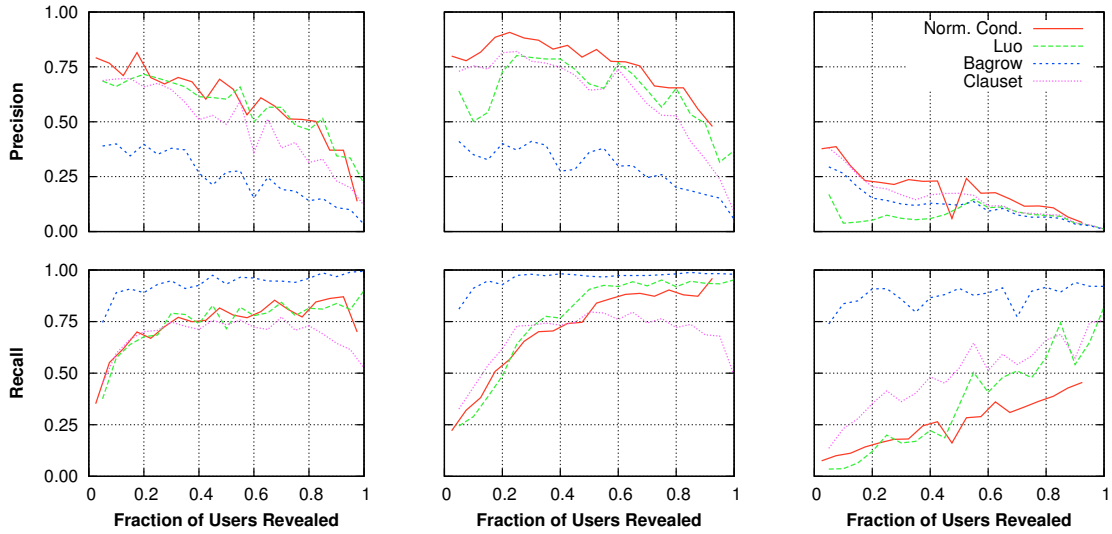
Majors: Finally, we observe that none of the algorithms are able to infer major; all have extremely low precision. This result is expected, though, since we observed in the previous section that majors do not form significant communities in the network.

Detecting graduate student communities

We now turn to evaluate our approach on the graduate student network. Figure 7.6 shows how the recall and precision vary as different fractions of the department, school, and matriculation year of graduate students are provided. Inferring the department and school of students shows good performance for all algorithms except for Bagrow's (as we observed with the undergraduates, the algorithm of Bagrow tended to return a large portion of the network as a community). We find that knowing 20% of the user attributes is sufficient to infer most of the remaining users with high accuracy. However, inferring matriculation year does not perform as well for any algorithm, having low recall and precision. Again, the poor performance at detecting matriculation years can be explained by the data in Section 7.2, which shows that the matriculation years form weak communities in the social network.

7.4 Summary

We began this section by asking whether the multiple overlapping community structures that exist in online social networks can be detected. We demonstrated that existing techniques can be "seeded" with attributes provided by users to detect mul-



(a) Department (b) School (c) Year

Figure 7.6 : Recall and precision for single community detection for Rice graduate students. Good performance is observed for department and school; much weaker performance is seen for year.

tiple partitionings according to different attributes. In fact, we found that with as few as 20% of users with known attributes, the remaining users can be classified with over 80% accuracy. Moreover, we proposed a new algorithm that can detect a community when given as input only a few users in the community. We found that this algorithm is able to detect communities in both the undergraduate and graduate student networks when given as few as 10% of the community. Thus, we found that with partial information about users, we are able to detect the multiple community structures that exist with high accuracy.

Our work has a number of implications and uses. For example, many of the popular online social networks could directly apply our algorithm in order to detect

communities in the network. This would enhance the user experience on the sites, as communities are often used for guiding search results, for suggesting users who may benefit from interaction, and for grouping users.

However, our findings also raise interesting questions about the nature of privacy in online social networks. In particular, almost all privacy mechanisms available to users today are based on access control: users can specify which other users are able to view the content or information they upload. Our results show, though, that even information that is not provided by users can be inferred from the user's location in the network. Thus, a user's privacy is not only a function of their actions, but also the actions of their friends and community members.

Chapter 8

Ostra: Leveraging Relationships

Internet-based communication systems such as email, instant messaging (IM), voice-over-IP (VoIP), online social networks, and content-sharing sites allow communication at near zero marginal cost to users. Any user with an inexpensive Internet connection has the potential to reach millions of users by uploading content to a sharing site or by posting messages to an email list. This property has democratized content publication: anyone can publish content, and anyone interested in the content can obtain it.

Unfortunately, the same property can be abused for the purpose of unsolicited marketing, propaganda, or disruption of legitimate communication. The problem manifests itself in different forms, such as spam messages in email; search engine spam in the Web; inappropriately labeled content on sharing sites such as YouTube; and unwanted invitations in IM, VoIP, and social networking systems.

Unwanted communication wastes human attention, which is one of the most valuable resources in the information age. The noise and annoyance created by unwanted communication reduces the effectiveness of online communication media. Moreover, most current efforts to automatically suppress unwanted communication occasionally discard relevant communication, reducing the reliability of the communication

medium.

Existing approaches to thwarting unwanted communication fall into three broad categories. First, one can target the unwanted communication itself, by automatically identifying such communication based on its content. Second, one can target the originators of unwanted communication, by identifying them and holding them accountable. Third, one can impose an upfront cost on senders for each communication, which may be refunded when the receiver accepts the item as wanted. Each of these approaches has certain advantages and disadvantages, which we discussed in Chapter 3.4.

In this chapter, we describe a method that exploits the difficulty in establishing and maintaining relationships in social networks to impose a cost on the senders of unwanted communication in a way that avoids the limitations of existing solutions. Our system, Ostra, (i) relies on existing social networks to connect senders and receivers via chains of pairwise relationships; (ii) uses a pairwise, link-based credit scheme that imposes a cost on originators of unwanted communications without requiring sender authentication or global identities; and (iii) relies on feedback from receivers to classify unwanted communication. Ostra ensures that unwanted communication strains the originator's relationships, even if the sender has no direct relationship with the ultimate recipient of the communication. A user who continues to send unwanted communication risks isolation and the eventual inability to communicate.

The relationships (or social links) that Ostra uses exist in many applications. The

links can be explicit, as in online social networking sites, or implicit, as in the links formed by a set of email, IM, or VoIP users who include each other in their contact lists. Ostra can use such existing social links as long as acquiring and maintaining a relationship requires some effort. For example, it takes some effort to be included in someone's IM contact list (making that person's acquaintance); and it may take more effort to maintain that status (occasionally producing wanted communication). With respect to Ostra, this property of a social network ensures that an attacker cannot acquire and maintain arbitrarily many relationships or replace lost relationships arbitrarily quickly.

Ostra is broadly applicable. Depending on how it is deployed, it can thwart unwanted email or instant messages; unwanted invitations in IM, VoIP, or online social networks; unwanted entries or comments in blogging systems; or inappropriate and mislabeled contributions to content-sharing sites such as Flickr and YouTube.

8.1 Ostra strawman

In this section, we describe a strawman design of Ostra. The design is appropriate for trusted, centralized communication systems in which users have strong identities (i.e., each individual user has exactly one digital identity). We discuss the basic properties of this design in the context of two-party communication (e.g., email and IM), multi-party communication (e.g., bulletin boards and mailing lists), and content-sharing sites (e.g., YouTube and Flickr). Section 8.2 describes a refined design that removes

the need for strong identities, because such identities are difficult to obtain in practice.

8.1.1 Assumptions

The strawman design is based on three assumptions.

1. Each user of the communication system has exactly one unique digital identity.

2. A trusted entity observes all user actions and associates them with the identity of the user performing the action.

3. Users classify communication they receive as wanted (relevant) or unwanted (irrelevant).

Assumption 1 would require a user background check (e.g., a credit check) as part of the account creation process, to ensure that a user cannot easily create multiple identities; this assumption will be relaxed in Section 8.2. Assumption 2 holds whenever a service is hosted by a trusted Web site or controlled by a trusted tracker component; the trusted component requires users to log in and associates all actions with a user. We sketch a decentralized design that does not depend on this assumption in Section 8.5.

Producing communication can mean sending a email or chat message; adding an entry or comment to a blog; sending an invitation in an IM, VoIP, or social networking system; or contributing content in a content-sharing site. Receiving communication can mean receiving a message or viewing a blog entry, comment, or search result.

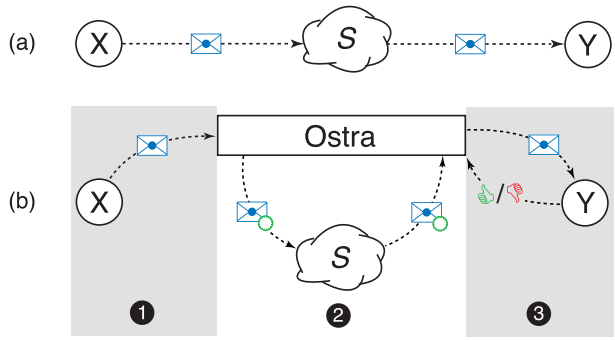


Figure 8.1 : Diagram of (a) the original communication system S, and (b) the communication system with Ostra. The three phases of Ostra — (1) authorization, (2) transmission, and (3) classification — are shown.

Typically, a user considers communication unwanted if she feels the content was not worth the attention. A user considers a blog entry, comment, or content object as unwanted if she considers the object to be inappropriate for the venue (e.g., site, group, or blog space) it was placed in or to have inappropriate search tags, causing the object to appear in response to an unrelated search.

8.1.2 System model

Figure 8.1 shows how Ostra interacts with a given communication system S. Ostra is a separate module that runs alongside the existing communication system. With Ostra, communication consists of three phases.

Authorization

When a sender wishes to produce a communication, she first passes the communication to Ostra. Ostra then issues a token specific to the sender, recipient, and commu-

nication. If the sender has previously sent too much unwanted communication, Ostra refuses to issue such a token and rejects the communication.

Transmission

Ostra attaches the token to the communication and transmits it using the existing communication mechanism. On the receiving side, Ostra accepts the communication if the token is valid. The communication is then provided to the recipient. Note that Ostra is not involved in the actual transmission of the communication.

Classification

The recipient classifies the communication as either wanted or unwanted, according to her personal preferences. This feedback is then provided to Ostra. Finally, Ostra makes this feedback available to the sender.

Note that in message-based communication systems, Ostra would normally be needed only for communication among users who do not regularly communicate with each other. Therefore, as an optimization, it is assumed that users maintain a whitelist of other users from whom they are willing to accept communication without endorsement from Ostra. For convenience, this whitelist could be automatically derived from the list of a user's direct friends in the social network.

8.1.3 User credit

Ostra uses credits to determine whether a token can be issued. Each user is assigned a credit balance, B, with an initial value of 0. Ostra also maintains a per-user balance range $[L, U]$, with $L \leq 0 \leq U$, which limits the range of the user's credit balance (i.e., $L \leq B \leq U$ at all times). We denote the balance and balance range for a single user as B_L^U. For example, if a user's state is 3_{-5}^{+6}, the user's current credit balance is 3, and it can range between –5 and 6.

When a token is issued, Ostra requires the sender to reserve a credit and the receiver to reserve a place holder for this credit in their respective credit balances. To make these reservations, the sender's L is raised by one, and the receiver's U is lowered by one. If these adjustments would cause either the sender's or the receiver's credit balance to exceed the balance range, Ostra refuses to issue the token; otherwise, the token is issued. When the communication is classified by the receiver, the range adjustments are undone. If the communication is marked as unwanted, one credit is transferred from the sender to the receiver.

Let us consider an example in which both the sender's and the receiver's initial balances are 0_{-3}^{+3}. When the token is issued, the sender's balance changes to 0_{-2}^{+3}, and the receiver's balance changes to 0_{-3}^{+2}, representing the credit reservation. Let us assume that the communication is classified as unwanted. In this case, a credit is transferred from the sender to the receiver; the receiver's balance becomes 1_{-3}^{+3}, and the sender's becomes -1_{-3}^{+3}.

This algorithm has several desirable properties. It limits the amount of unwanted communication a sender can produce. At the same time, it allows an arbitrary amount of wanted communication. The algorithm limits the number of tokens a user can acquire before any of the associated communication is classified; thus, it limits the total amount of potentially unwanted communication a user can produce. Finally, the algorithm limits the number of tokens that can be issued for a specific recipient before that recipient classifies any of the associated communication; thus, an inactive or lazy user cannot cause senders to reserve a large number of credits, which would be bound until the communication were classified.

8.1.4 Credit adjustments

Several issues, however, remain with the algorithm described so far. When a user's credit balance reaches one of her credit bounds, she is, in effect, banned from producing (in the case of the lower bound) or receiving (in the case of the upper bound) any further communication. What can cause a legitimate user's credit balance to reach her bounds? Note that on the one hand, a user who receives unwanted communication earns credit. On the other hand, even a well-intentioned user may occasionally send communication to a recipient who considers it unwanted and therefore lose credit. Across all users, these effects balance out. However, unless a user, on average, receives precisely the same amount of unwanted communication as she generates, her credit balance will eventually reach one of her bounds. As a result, legitimate users

can find themselves unable to communicate.

To address this problem, credit balances in Ostra decay towards 0 at a constant rate d with $0 \leq d \leq 1$. For example, Ostra may be configured so that each day, any outstanding credit (whether positive or negative) decays by 10%. This decay allows an imbalance between the credit earned and the credit lost by a user. The choice of d must be high enough to cover the expected imbalance but low enough to prevent considerable amounts of intentional unwanted communication. As we show as part of Ostra's evaluation, a small value of d is sufficient to ensure that most legitimate users never exceed their credit range.

With this refinement, Ostra ensures that each user can produce unwanted communication at a rate of at most

$$d * L + S \tag{8.1}$$

where S is the rate at which the user receives communication that she marks as unwanted.

A denial of service attack is, however, still possible. Colluding malicious users can inundate a victim with large amounts of unwanted communication, causing the victim to acquire too much credit to receive any additional communication. For these users, the rate of decay may be too low to ensure that they do not exceed their credit balances. To prevent such attacks, we introduce a special account, C, that is not owned by any user and has no upper bound. Users with too much credit can transfer credit into C, thereby enabling them to receive further communication. Note that

Operation	Net Change in System Credit
User joins system	0, as user's initial credit balance is 0
Wanted communication sent	0, as no credit is exchanged
Unwanted communication sent	0, as credit is transferred between users
Daily credit decay	0, as total credit was 0 before decay

Table 8.1 : Operations in Ostra, and their effect on the total system credit. No operation alters the sum of credit balances.

the credit transferred into C is subject to the usual credit decay, so the total amount of credit available to active user accounts does not diminish over time. Additionally, users can only deposit credit into C; no withdrawals are allowed.

Finally, there is an issue with communication failures (e.g., dropped messages) and users who are offline for extended periods. Both may cause the sender to reserve a credit indefinitely, because the receiver does not classify the communication. The credit decay does not help in this situation, because the decay affects only the credit balance and not the credit bounds. Therefore, Ostra uses a timeout T, which is typically on the order of days. If a communication has not been classified by the receiver after T, the credit bounds are automatically reset as though the destination had classified the communication as wanted. This feature has the added benefit that it enables receivers to plausibly deny receipt of communication. A receiver can choose not to classify some communication, thus concealing its receipt.

	Action	Cost	Reward
Sending	Send wanted comm.		
	Send unwanted comm.	1 credit	
Classifying	Classify as wanted		Sender likely to send more
	Classify as unwanted		1 credit, throttle sender
	Misclassify as wanted	Encourage sending more	
	Misclassify as unwanted	Discourage sending more	1 credit
Abuse	Don't use token	Ties up credit for T	
	Don't classify	Ties up credit for T	
	Drop incoming comm.	1 credit	

Table 8.2 : Incentives for users of Ostra. Users are incentivized to send only wanted communication, to classify communication correctly, and to classify received communication promptly. Marking an incoming communication as unwanted has the effect of discouraging the sender from sending additional communication, as the sender is informed of this and loses credit. Alternatively, marking an incoming communication as wanted costs the sender nothing, allowing the sender to send future communication with increased confidence.

8.1.5 Properties

Ostra's system of credit balances observes the following invariant:

At all times, the sum of all credit balances is 0

The conservation of credit follows from the fact that (i) users have an initial zero balance when joining the system, (ii) all operations transfer credit among users, and (iii) credit decay affects positive and negative credit at the same rate. Table 8.1 details how each operation leaves the overall credit balance unchanged. Thus, credit can be neither created nor destroyed. Malicious, colluding users can pass credits only between themselves; they cannot create additional credit or destroy credit. The amount of unwanted communication that such users can produce is the same as the sum of what they can produce individually.

We have already shown that each user can produce unwanted communication at a rate of no more than $d * L + S$. We now characterize the amount of unwanted subset of the user population can produce. Let us examine a group of users F. Owing to the conservation of credit, users in this group cannot conspire to create credit; they can only push credit between themselves. Thus, the users in F can send unwanted communication to users not in F at a maximal rate of

$$|F| * d * L + S_F \tag{8.2}$$

where S_F is that rate at which users in F (in aggregate) receive communication from users not in F that they mark as unwanted.

The implication of the above analysis is that we can characterize the total amount of unwanted communication that non-malicious users can receive. Let us partition the user population into two groups: group G are "good" users, who rarely send unwanted communication, and group M are "malicious" users, who frequently send unwanted communication. Now, the maximal rate at which G can receive unwanted communication from M is

$$|M| * d * L + S_M \tag{8.3}$$

which implies that, on average, each user in G can receive unwanted communication at a rate of

$$\frac{|M|}{|G|} * d * L + \frac{S_M}{|G|} \tag{8.4}$$

However, we expect S_M to be small as users in G rarely send unwanted communication. Thus, the rate of receiving unwanted communication is dominated by static system parameters and by the ratio between the number of good and malicious users. Moreover, this analysis holds regardless of the amount of good communication that the malicious users produce.

Finally, Ostra has an incentive structure that discourages bad behavior and rewards good behavior. Table 8.2 shows a list of possible user actions and their costs and rewards.

8.1.6 Multi-party communication

Next, we show how the design can be used to support moderated multi-party communication, including mailing lists and content-sharing sites. The existing design generalizes naturally to small groups in which all members know each other. In this case, communication occurs as a series of pairwise events between the originator and each of the remaining group members.

In moderated groups, which are usually larger, a moderator decides on behalf of the list members if communication submitted to the group is appropriate. In this case, Ostra works exactly as in the two-party case, except that the moderator receives and classifies the communication on behalf of all members of the group.

Thus, only the moderator's attention is wasted by unwanted communication, and the cost of producing unwanted communication is the same as in the two-party case. However, an overloaded moderator may choose to increase the number of credits required to send to the group, to mitigate her load by discouraging inappropriate submissions.

Large content-sharing sites usually have content-rating systems or other methods for flagging content as inappropriate. Ostra could be applied, for instance, to thwart the submission of mislabeled videos in YouTube, by taking advantage of the existing "flag as inappropriate" mechanism. When a user's video is flagged as inappropriate, it is reviewed by a YouTube employee; if it is found to be mislabeled, the submission is classified as unwanted for the purposes of Ostra.

Extending Ostra to work with unmoderated multi-party communication systems is beyond the scope of this thesis.

8.2 Ostra design

The strawman design described in the previous section requires strong user identities: that is, each individual user is guaranteed to have at most one unique digital identity. Such identities are not practical in many applications, as they require a background check as part of the account creation process. Such checks may not be accepted by users, and as far as we know, few services that require such a strong background check have been widely adopted on the Internet.

In this section, we refine the design of Ostra so that it does not require strong user identities. It is assumed that the communication system ensures that each identity is unique, but an individual user may sign up multiple times and use the system under different identities at different times. Our refined design leverages relationships to preserve Ostra's properties despite the lack of strong user identities. We still assume that a trusted entity such as a Web site hosts the communication service and runs Ostra. Later, in Section 8.5, we sketch out how Ostra could be applied to decentralized services.

The refined design of Ostra replaces the per-user credit balances with balances that are instead associated with the links among users in a trust network. We show that this mapping preserves the key properties of the strawman design, even though

Ostra no longer depends on strong identities. We begin by defining a trust network and then describe how Ostra works with weak identities.

8.2.1 Trust networks

A trust network is a graph $G = (V, E)$, where V is the set of user identifiers and E represents undirected links between user identifiers who have a trust relationship. Examples of trust networks are the user graph of an email system (where V is the set of email addresses and E is the set of email contacts) and online social networks (where V is the set of accounts and E is the set of friends). For convenience, we shall refer to two users connected by an edge in the trust network as friends.

For the purposes of Ostra, a trust network must have the property that there is a non-trivial cost for initiating and maintaining links in the network. As a result, users in the trust network cannot acquire new relationships arbitrarily fast and cannot maintain an arbitrarily large number of relationships. We do not make any assumptions about the nature or the degree of trust associated with a relationship.

Finally, the trust network must be connected, meaning that there is a path of trust links between any two user identities in the network. Previous studies [26, 105] have shown that the user graphs in existing social networks tend to be dominated by a single large component, implying that the networks are largely connected.

Ostra assumes that the users of a communication system are connected by a trust network and that Ostra has a complete view of this network.

8.2.2 Link credit

Because a user may have multiple identities, we can no longer associate a separate credit balance with each identity. Otherwise, a malicious user could gain additional credit and send arbitrary amounts of unwanted communication simply by creating more identities. Instead, Ostra leverages the cost of forming new links in trust networks to enforce a bound on each user.

Specifically, each link in the trust network is assigned a link credit balance B, with an initial value of 0, and a link balance range $[L, U]$, with $L \leq 0 \leq U$ and $L \leq B \leq U$. These are analogous to the user credit balance and range in the original design. We denote the balance and balance range for a link $X \leftrightarrow Y$ from X's perspective as $\overset{X \to Y}{B_L^U}$. For example, if the link has the state $\overset{X \to Y}{3_{-5}^{+6}}$, then X is currently owed 3 credits by Y, and the balance can range between –5 and 6.

The link balance represents the credit state between the user identities connected by the link. Ostra uses this balance to decide whether to issue tokens. It is important to note that the credit balance is symmetric. For example, if the link balance on the $X \leftrightarrow Y$ link is $\overset{X \to Y}{1_{-2}^{+3}}$, then X is owed one credit by Y, or, from Y's perspective, Y owes X one credit (the latter can be denoted $\overset{Y \to X}{-1_{-3}^{+2}}$).

We map the user credit balance in the strawman design to a set of link credit balances on the user's adjacent links in the trust network. For example, as shown in Figure 8.2, if a user has two links in the trust network, the user's original credit balance is replaced with two separate credit balances, one on each link. However, we

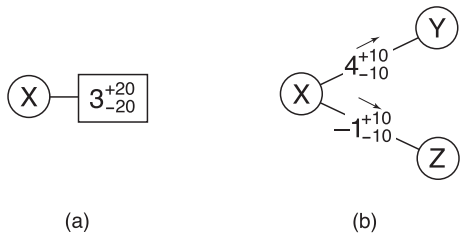


Figure 8.2 : Mapping from (a) per-user credits to (b) per-link credits.

cannot compute a user balance by taking the sum of the link balances – in fact, the concept of a user balance is no longer useful because a user can create many identities and establish links between them. Instead, we introduce a new mechanism for credit transfer that uses link balances, rather than user balances, to bound the amount of unwanted communication that users can send.

We now describe this mechanism for transferring credits. For simplicity, we first describe the case of communication between users who are friends in the trust network. We then generalize the credit transfer mechanism to the case in which two arbitrary users wish to communicate.

Communication among friends

As in the strawman design, a user who wishes to send communication needs to obtain a token during the authorization phase. For example, a user X may request to send communication to another user Y, a friend of X's. Ostra determines whether transferring this credit would violate the link balance range on the $X \leftrightarrow Y$ link, and if not, it issues a signed token. The token is then included in X's communication to

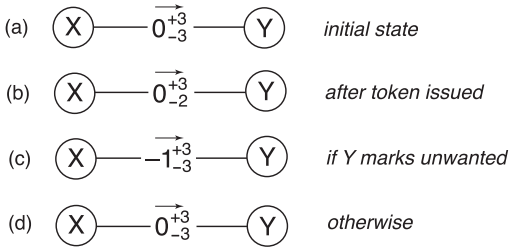


Figure 8.3 : Link state when X sends communication to friend Y. The state of the link balance and range is shown (a) before the token is issued, (b) after the token is issued, (c) if Y marks the communication as unwanted, and (d) if Y marks the communication as wanted or if the timeout occurs.

user Y.

As in the strawman design, Ostra allows users to have multiple outstanding tokens by reserving credits for each potential transfer. In the example in the previous paragraph, Ostra raises the lower bound for the $X \leftrightarrow Y$ link by one. This single adjustment has the effect of raising X's lower bound and lowering Y's upper bound, because the lower bound on the $X \leftrightarrow Y$ link can be viewed as the upper bound on the $Y \leftrightarrow X$ link. Figure 8.3 shows the state of the $X \leftrightarrow Y$ link during each stage of the transaction. By adjusting the balance this way for outstanding tokens, Ostra ensures that the link balance remains within its range regardless of how the pending communication events are classified.

Later, in the classification stage, user Y provides Ostra with the token and the decision whether X's communication was wanted. The balance range adjustment that was performed in the authorization phase is then undone. Moreover, if Y reports that the communication was unwanted, Ostra adjusts the balance on the $X \leftrightarrow Y$ link by subtracting one, thereby transferring a credit from X to Y. Thus, if the previous

state of the link was $\overset{X\to Y}{0^{+3}_{-3}}$, the final state would be $\overset{X\to Y}{-1^{+3}_{-3}}$, meaning X owes Y one credit. Finally, Ostra automatically cancels the token after a specified timeout T.

Communication among non-friends

So far, we have considered the case of sending communication between two friends in the trust network. In this section, we describe how Ostra can be used for communication between any pair of users.

When a user X wishes to send communication to a non-friend Z, Ostra finds a path consisting of trust links between X and Z. For example, such a path might be $X \leftrightarrow Y \leftrightarrow Z$, where X and Y are friends in the trust network, and Y and Z are also friends. When this path is found, the range bounds are adjusted as before, but this occurs on every link in the path. For example, if X wishes to send communication to Z, Ostra would raise the lower bound of both the $X \leftrightarrow Y$ and the $Y \leftrightarrow Z$ links by one. Figure 8.4 shows a diagram of this procedure. If this adjustment can be done without violating any link ranges, Ostra issues a token to X.

Similar to the transfer between friends, the token is then attached to X's communication to Z. Later, in the classification stage, Z provides Ostra with the token and the decision whether the communication was wanted. Now, the range adjustments on all the links along the path are undone. If the communication was unwanted, the credit is transferred along every link of the path; Figure 8.4 (c) shows the result of this transfer.

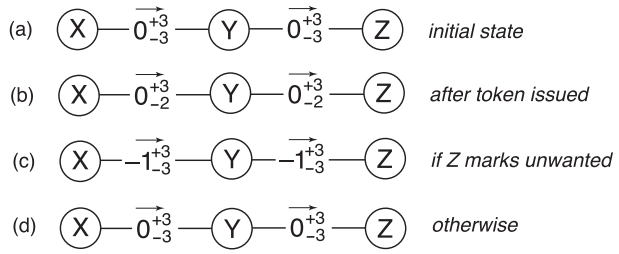


Figure 8.4 : Link state when X sends communication to non-friend Z is shown (a) before the token is issued, (b) after the token is issued, (c) if Z marks the communication as unwanted, and (d) if Z marks the communication as wanted or if the timeout occurs.

It is worth noting that the intermediate users along the path are largely indifferent to the outcome of the transfer, as any credit transfer will leave them with no net change. For example, consider the scenarios shown in Figure 8.4 (c) and (d). In either case, the total amount of credit that intermediate user Y has with all her friends is the same regardless of the outcome. If Z marks the communication as unwanted, as shown in Figure 8.4(c), Y owes a credit to Z, but X now owes a credit to Y. Ostra allows users to transfer credits along trust paths such that intermediate users along the path are indifferent to the outcome.

Generalization of Ostra strawman

One can show that Ostra generalizes the strawman design from the previous section. Recall the account C that is owned by the trusted site. Now, we construct a trust network in which each user has a single link to C, with the link balance and balance range equal to their user balance and balance range in the strawman design. Ostra with such a trust network has the same properties as the strawman design. To see this,

note that sending communication from X to Y requires raising the lower bound on the $X \leftrightarrow C$ link and lowering the upper bound on the $Y \leftrightarrow C$ link, which is equivalent to adjusting X's and Y's user balance ranges in the same manner. Figure 8.5 (b) shows an example of this generalization for the specific set of user accounts in Figure 8.5 (a).

More importantly, Ostra preserves the conservation of credit that was present in the strawman system. This can be derived from the fact that credit is associated with links instead of users. Any credit in Ostra is naturally paired with a corresponding debt: for example, if the state of a link is $-1^{X \to Y}_{-2}$, then X owes Y one credit, but Y is owed a credit by X. Thus, all outstanding credit is balanced by outstanding debt, implying that credit cannot be created or destroyed.

The conservation of credit holds for each link independently, and is therefore independent of the trust network topology (Figure 8.5 (c) shows an example of a trust network with a different topology). As a result, the analysis of the strawman system in Section 8.1.5 applies to the full version of Ostra. For example, malicious, colluding users cannot conspire to manufacture credit; the amount of unwanted communication that such users can produce together is the sum of what they can produce independently.

8.2.3 Security properties

We now discuss the security properties of Ostra's refined design in detail. Ostra's threat model assumes that malicious users have two goals: sending large amounts

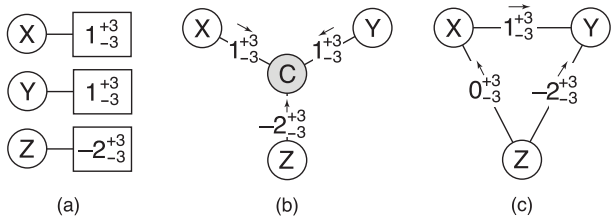


Figure 8.5 : Generalization of per-user credit accounting to per-link credit accounting. Ostra with per-user credit (shown in (a)) can be expressed as per-link credit over a star topology (shown in (b)), with the central site C as the hub. The addition of links (shown in (c)) does not change the properties.

of unwanted communication, and preventing other users from being able to send communication successfully. Strategies for trying to send additional unwanted communication include signing up for multiple accounts and creating links between these accounts, as well as conspiring with other malicious users. Strategies for trying to prevent other users from communicating include targeting a specific user by sending large amounts of unwanted communication and attempting to exhaust credit on specific links in the trust network. In this section, we describe how Ostra handles these threats.

Multiple identities

One concern is whether users who create multiple identities (known as Sybils [44]) can send additional unwanted communication. Ostra naturally prevents such users from gaining additional credit.

To send unwanted communication to another user, a user must eventually use one of her "real" links to a different user, which has the same effect as if the user only

had a single identity. To see this, assume a user with a set of multiple identities $M = \{M_1, M_2, ..., M_n\}$ is sending to a different user U. Now, regardless of how the links between the identities in M are allocated, any path between M_i and U must contain a link $M_j \leftrightarrow V$, where $V \notin M$. If this property does not hold, then $U \in M$, which is a contradiction.

Thus, using per-link balances has the effect that the total credit available to a user no longer depends on the number of identities a user has. Instead, the credit available depends on the number of links the user has to other users. Figure 8.6 shows a diagram of how Ostra prevents users with multiple identities from sending additional unwanted communication.

Ostra allows users to create as many identities as they wish but ensures that they cannot send additional unwanted communication by doing so. Malicious users may attempt to use multiple Sybil identities to create multiple links to a single user. Although they may succeed occasionally, these links require effort to maintain and the malicious user, therefore, cannot create an unbounded number of them.

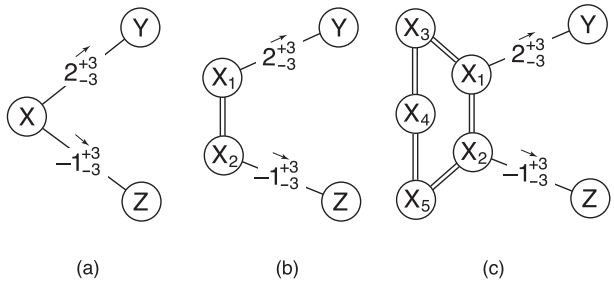


Figure 8.6 : Diagram of how Ostra handles various attacks: (a) a normal user, (b) multiple identities, and (c) a network of Sybils. The total amount of credit available to the user is the same.

Targeting users

Another concern is whether malicious users could collectively send a large amount of unwanted communication to a user, thus providing this victim with too much credit to receive any additional messages. This attack is possible when the attacking users collectively have more links to legitimate users than the victim, as exhausting the credit on one of the victim's links requires the malicious users exhaust the credit on one of their own links.

However, the victim has a simple way out by forgiving some of the debt on one of her links. If a user finds that she has too much credit on all of her links, she can forgive a small amount of debt from one of her friends. This is the same mechanism as transferring credit to the overflow account (C) described in Section 8.1. To see this equivalence, consider the star-topology trust network constructed in Section 8.2.2. In that case, a user transferring credit to the overflow account is essentially forgiving debt on their only link (to C). This mechanism does not allow malicious users to send additional unwanted communication to the victim, as the victim only forgives debt to her direct friend (i.e., the victim's friend does not repeat the process).

Targeting links

One final concern is whether malicious users could prevent large numbers of innocent users from communicating with each other by exhausting the credit on certain links in the trust network. If successful, such an attack could prevent a group of users from

sending to the rest of the user population.

To exhaust the credit on specific links, attacking users would need both knowledge of the trust network topology and some control over trust path selection. Because the path selection is performed by the trusted site, the attacking users have the choice of only the destination and not the path. Even if we assume a powerful attacker who has control over the path selection, the trust network would need to have a topology that is susceptible to such an attack. For example, a barbell topology would be susceptible, as the link connecting the two halves of the network could be exhausted.

Analysis of current online social networks (which are typical trust networks) shows that these have a very dense core [105]. We show in Section 8.4 that the structure of these networks makes it unlikely that such an attack would succeed on a large scale.

8.3 Discussion

In this section, we discuss some issues associated with deploying Ostra.

8.3.1 Joining Ostra

Fundamentally, Ostra exploits the trust relationships in an existing social network of users to thwart unwanted communication. As a result, users are expected to acquire and maintain a certain number of social links to be able to communicate.

To join Ostra, a new user must be introduced to the system by an existing Ostra user. Requiring this form of introduction ensures that the trust network among users

is connected and that each new user has at least one trust link. Thus, Ostra can be used only in conjunction with a "invitation-only" social network.

Users with few links in the trust network are more susceptible to credit exhaustion (whether accidental or malicious). Thus, there is an incentive for users to obtain and maintain a sufficient number of trust links. Establishing additional links can be done via the communication system after the user has joined Ostra. Link invitations are treated as normal messages, so users who attempt to send unwanted link invitations are blocked in the same manner as users who send other forms of unwanted communication.

8.3.2 Content classification

Ostra requires that recipients classify incoming communication as either wanted or unwanted. Providing explicit feedback is a slight burden on the user, but it may be a small price to pay for a system that responds to each user's preferences and is free of the misclassifications that are common in current content-based filtering systems [5]. Moreover, the feedback can often be derived implicitly from a user's actions; for instance, deleting a message probably indicates that the message was unwanted, whereas archiving or replying to the message strongly indicates that it was wanted.

As an optimization in message-based communication systems, a user could maintain a whitelist indicating users from whom communication is immediately and un-

conditionally classified as wanted. In this case, Ostra would need to operate only among users who are not on each other's whitelists.

8.3.3 Parameter settings

Ostra limits the amount of pending communication that a user can have, where a pending item of communication is one that was generated by the user but not yet classified by the receiver. In Section 8.4, we show that Ostra's design parameters (L, U, and d) can be chosen such that most legitimate users are not affected by the rate limit, while the amount of unwanted communication is still kept very low.

The L parameter controls the number of unclassified items of communication a user can have at any one time. A large L allows many outstanding messages but also admits the possibility that a considerable amount of this outstanding communication would be unwanted. In contrast, an L close to 0 ensures that very little unwanted communication is received, at the cost of potentially rate-limiting legitimate senders. The d parameter represents the rate at which users who have sent unwanted communication in the past are "forgiven". Setting d too high allows additional unwanted communication, whereas setting it too low may unduly punish senders who have inadvertently sent unwanted communication in the past. In the Section 8.4, we show that the conservative settings of L=–3 and d=10% per day provide a good trade-off in practice.

8.3.4 Compromised user accounts

If a user's account password is compromised, the attacker can cause the user to run out of credit by sending unwanted communication. However, the amount of unwanted communication is still subject to the same limits that apply to any individual user. Moreover, a user would quickly detect that her account has been compromised, because she would find herself unable to generate communication.

8.4 Evaluation

In this section, we present an experimental evaluation of our Ostra prototype. Using data from a real online social network and an email trace from our institute, we show how Ostra can effectively block users from sending large amounts of unwanted communication.

8.4.1 Experimental trust network

To evaluate Ostra, we used a large, measured subset [105] of the social network found in the video-sharing Web site YouTube [167]. We extracted the largest strongly connected component consisting of symmetric links from the YouTube graph, which resulted in a network with 446,181 users and 1,728,938 symmetric links.

Strictly speaking, the YouTube social network does not meet Ostra's requirements, because there is no significant cost for creating and maintaining a link. Unfortunately, trust-based social networks that do meet Ostra's requirements cannot be easily ob-

tained due to privacy restrictions. For instance, in the LinkedIn [95] professional networking site, users "vouch" for each other; link formation requires the consent of both parties and users tend to refuse to accept invitations from people they do not know and trust. But, unlike YouTube, it is not possible to crawl the LinkedIn network.

However, we were able to obtain the degree distribution of users in the LinkedIn network. We found that both YouTube and LinkedIn degree distributions follow the power-law with similar coefficients. We used maximum-likelihood testing to calculate the coefficients of the YouTube and LinkedIn graphs, and found them to be 1.66 and 1.58 (the resultant Kolmogorov-Smirnov goodness-of-fit metrics were 0.12 and 0.05, suggesting a good fit). This result, along with the previously observed similarity in online social networks' structure [105], leads us to expect that the overall structure of the YouTube network is similar to trust-based social networks like LinkedIn.

Despite their structural similarity, the YouTube social network differs from the LinkedIn trust network in one important aspect: some users in YouTube collect many links (one user had a degree of over 20,000!). The maximum degree of users in actual trust-based social networks tends to be much smaller. Anthropological studies [47] have shown that the average number of relationships a human can actively maintain in the real world is about 150 to 200. Because the amount of unwanted communication a user can send in Ostra is proportional to her degree in the trust network, the results of our YouTube-based evaluation may understate the performance of Ostra on a real

trust-based network.

8.4.2 Experimental traffic workload

We were unable to obtain a communication trace of the same scale as the social network we use. Therefore, we had to make some assumptions about the likely communication pattern within the social network. We expect that users communicate with nearby users much more often than they communicate with users who are far away in the social network. To validate this hypothesis, we collected an email trace from the Max Planck Institutes for Informations and Software Systems, consisting of two academic research institutes with approximately 200 researchers. Our anonymized email trace contains all messages sent and received by the mail servers for 100 days, and the anonymized addresses in the trace are flagged as internal or external addresses.

Similar to previous studies [28, 143], we extracted a social network from the email data by examining the messages sent between internal users. Specifically, we created a symmetric link between users who sent at least three emails to each other. We filtered out accounts that were not owned by actual users (e.g., helpdesk tickets and mailing lists), resulting in a large strongly connected component containing 150 users and covering 13,978 emails.

We then examined the social network distance between sender and receiver for all messages sent between these 150 users. Figure 8.7 compares the resulting distance distribution with one that would result had the senders selected random destinations.

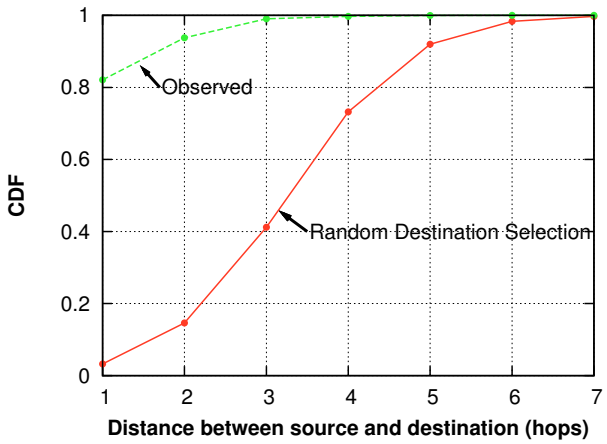


Figure 8.7 : Cumulative distribution (CDF) of distance between sender and receiver for our email trace. The observed data show a strong bias toward proximity when compared to randomly selected destinations.

We found that the selection of senders had a very strong proximity bias: over 93% of all messages were sent to either a friend or a friend of a friend, compared to the expected 14% if the senders were chosen randomly. Thus, we expect that in practice, most communication in Ostra is directed to nearby users, significantly reducing the average path lengths in the trust network.

8.4.3 Setting parameters

We also used the email trace to determine the appropriate settings for the Ostra parameters L and U. To do this, we examined the rate at which users sent and received messages. The trace contains 50,864 transmitted messages (an average of 3.39 messages sent per user per day) and 1,003,819 received messages (an average of 66.9 messages received per user per day). The system administrators estimated that the incoming messages in the email trace consisted of approximately 95% junk mail.

Clearly, most of these receptions would not occur in an actual Ostra deployment. However, we could not access the spam filter's per-message junk mail tags, so we randomly removed 95% of the incoming messages as junk.

To determine how often a given setting of L and U would affect Ostra, we simulated how messages in the email trace would be delayed due to the credit bounds. We ran two experiments with different assumptions about the average delay between the time when a message arrives and the time when the receiving user classifies the message. We first simulated casual email users who classify messages after six hours, and we then simulated heavy email users who classify messages after two hours.

Table 8.3 presents the results of these two experiments with L=–3 and U=3. We found that messages are rarely delayed (less than 1.5% of the time in all cases), and that the average delay is on the order of a few hours. We also found that the delays for receiving messages are more significant than the delays for sending messages. We believe this is an artifact of our methodology. Over 98% of the delayed messages were received by just 3 users. In practice, it is likely that these users (who receive a high volume of relevant email) check and classify their email very frequently. This effect would reduce the frequency and magnitude of delays, but our simulation does not account for it.

	Average classification delay (h)	Fraction delayed	Delay (h)		
			Avg.	Med.	Max.
Sending	2	0.38%	2.2	1.9	7.6
	6	0.57%	6.1	5.3	23.6
Receiving	2	1.3%	4.1	3.2	13.2
	6	1.3%	16.6	14.7	48.6

Table 8.3 : Message delays in sending and receiving with L=–3 and U=3. The delays are shown for heavy email users (2 hour average classification delay) and casual email users (6 hour average classification delay).

8.4.4 Effectiveness of Ostra

In this section, we simulate deployments of Ostra in a message-based system (such as the messaging service on Flickr) and in a content-sharing system (such as YouTube). We evaluate Ostra under three traffic workloads: *Random*, where users select destinations randomly; *Proximity*, where users select destinations with the distribution that was observed in Section 8.4.2; and *YouTube*, where users send to a single YouTube account in the network. We show that in all cases, Ostra effectively bounds the rate at which malicious users can send unwanted communication while not impeding wanted communication.

Expected performance

Ostra limits the amount of unwanted communication that can be sent. A single user user can send unwanted communication at a rate of at most $d * L * D + S$, where D

is the degree of the user. Thus, the rate at which a malicious user can send unwanted communication is in direct proportion to her degree. As the d or L parameters are increased, we expect the rate of unwanted communication to increase accordingly. Additionally, as the proportion of malicious users in the network increases, we expect the overall rate of unwanted messages to increase.

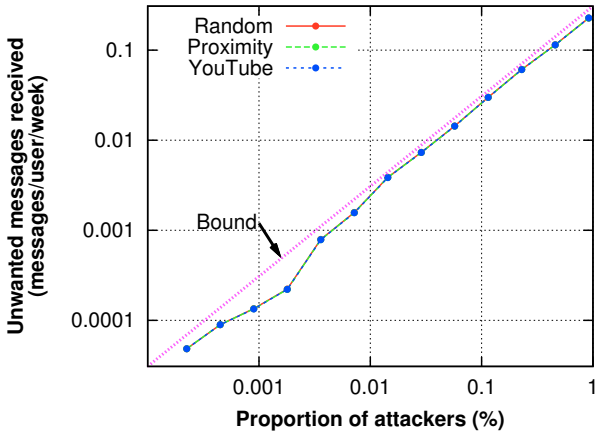


Figure 8.8 : Amount of unwanted communication received by good users as the number of attackers is varied. As the number of attackers is increased, the number of unwanted messages delivered scales linearly.

Preventing unwanted communication

In this section we verify experimentally that Ostra performs as described in Section 8.4.4. Unless otherwise noted, the experiments were run with 512 randomly chosen attackers (approximately 0.1% of the population), L=–3, U=3, and d=10% per day. Each good user sent 2 messages and each attacker sent 500 messages.

To evaluate Ostra in the context of a content-sharing site, we modeled Ostra working in conjunction with YouTube. For these experiments, we configured the

network so that uploading a video involves sending a message via Ostra to a single 'YouTube' account in the network. An existing, well-connected user (1,376 links) in the core of the network was selected to represent this account.

We first show that the rate at which users receive unwanted communication varies with the number of attacking users. In Figure 8.8, we present the results of experiments in which we vary the number of attackers in the network between 1 and 4,096 users (0.0002% to 1% of the network). We examine the rate at which unwanted messages were received by non-attacking users, along with the expected bound derived from the equations in Section 8.4.4.

As can be seen in Figure 8.8, Ostra effectively bounds the number of unwanted messages in proportion to the fraction of users who send unwanted communication. Even with 1% of the network sending unwanted messages, each legitimate user receives only 0.22 unwanted messages per week, translating to approximately 12 unwanted messages per year.

Next, we explore Ostra's sensitivity to system parameter settings and other conditions. Important parameters in Ostra are the credit bounds L and U for each link. If these bounds are set too high, attackers can send many messages before being cut off. However, if these bounds are set too low, a legitimate user could be temporarily prevented from sending messages. Figure 8.9 shows how the rate of unwanted message delivery is affected by the maximal credit imbalance across a link. As the maximum allowed imbalance increases, the amount of unwanted communication received by

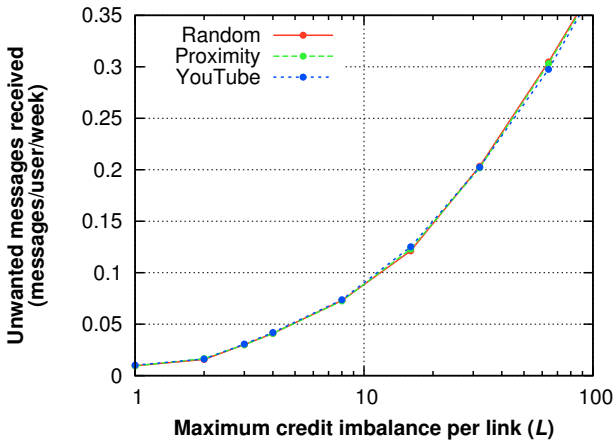


Figure 8.9 : Amount of unwanted communication received by good users as the maximum credit imbalance per link is varied.

good users increases, as expected.

Finally, we examine the sensitivity of Ostra to the false positive rate of legitimate users' message classification. In other words, if users incorrectly mark other good users' messages as unwanted, how often are users blocked from sending message? We show how this probability of false classification affects the proportion of messages that cannot be sent in Figure 8.10. As can be seen, even a high false positive rate of 30% results in only a few blocked messages. This resiliency results from the rich connectivity of the social network (i.e., if one link becomes blocked, the users can route through other friends), and the fact that the false positive rate affects all users equally.

In the case of the content-sharing site, because all paths intersect, good users are blocked more quickly as the amount of content that is marked as unwanted increases. For example, when the false classification rate is 64%, about 40% of messages cannot

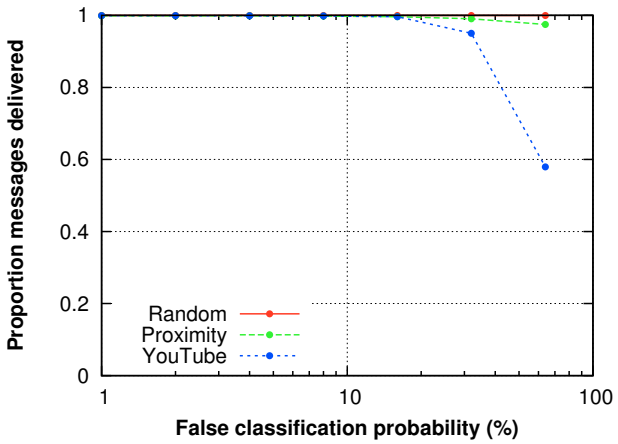


Figure 8.10 : Proportion of messages delivered versus false classification probability for wanted messages.

be sent. However, it seems very unlikely that the moderator of a sharing site would misclassify content at such a high rate.

Resilience to link attacks

In a potential security attack discussed in Section 8.2, malicious users attempt to exhaust credit on a set of links inside the trust network, i.e., links other than the attackers' adjacent links. If successful, this attack could disrupt communication for innocent users. To evaluate whether a real-world social network is susceptible to this attack, we performed a min-cut analysis of the YouTube social network.

Assuming uniform link weights of one, we calculated the min-cuts[1] between 3,000 randomly selected pairs of users. We then looked for cases in which the set of links

[1]A min-cut is a minimal set of links that, if removed, partitions two users; note that several such cuts can occur between two users.

involved in a min-cut for a given pair of users differed from the set of links adjacent to either one of the two users. Such a min-cut could be the target of an attack, because the attackers could exhaust credit on this set of links before they exhaust the credit on their own links.

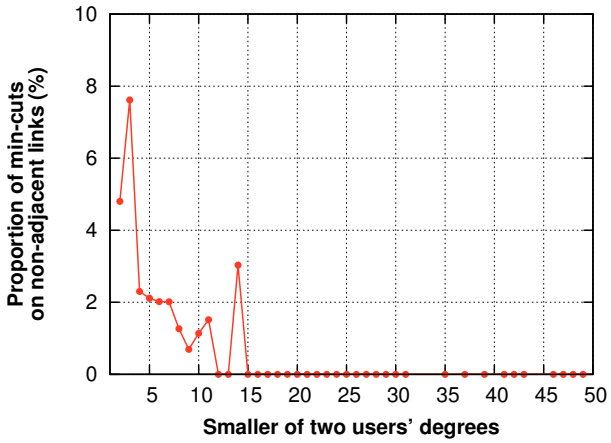


Figure 8.11 : Proportion of 3,000 random user pairs for which the min-cut was not adjacent to one of the users, as a function of the lower of the two users' degrees. The fraction decreases as the users become well-connected, suggesting that a trust network with well-connected users is not vulnerable to link attacks.

Figure 8.11 plots the proportion of user pairs for which the min-cut was not adjacent to one of the users, as a function of the lower of the two users' degrees. The results suggest that vulnerable links inside the network occur rarely, and that their frequency decreases with the degree of user connectivity. Therefore, the better connected users are in the trust network, the more robust the network is to link attacks. Because users in Ostra already have an incentive to maintain a certain number of links for other reasons, one would expect that a real Ostra trust network would not be vulnerable to link attacks.

8.5 Decentralizing Ostra

The design of Ostra we have described so far assumes the existence of a trusted, centralized component that maintains the trust network and credit state. This design is suitable for centralized communication systems, such as those hosted by a Web site. Peer-to-peer communication systems with a centralized "tracker" component can also use this design. However, completely decentralized systems like SMTP-based email cannot use it. In this section, we briefly sketch out a design of Ostra that works without any trusted, centralized components.

8.5.1 Overview

In the absence of a trusted, centralized entity, both the trust network and the credit state must be distributed. We assume that each participating user runs an Ostra software agent on her own computer. This Ostra agent stores the user's key material and maintains secure network connections to the Ostra agents of the user's trusted friends. The two Ostra agents adjacent to a trust link each store a copy of the link's balance and bounds.

Ostra authorization requires a route computation in the trust network. Because user trust networks can be very large (many online social networks have hundreds of millions of users), the path computation must be scalable. Moreover, it is assumed that users wish to keep their trust relationships private. In a centralized design, such privacy can be ensured easily. In the decentralized design, this concern complicates

the distributed route computation, as no user has a global view of the trust network.

In the sections below, we sketch out distributed designs for the route computation, for maintaining link balances and for ensuring that users follow the Ostra protocol.

8.5.2 Routing

Routing in large networks is a well-studied problem. We use a combination of existing techniques for distributed route discovery in large trust networks.

We divide the problem into two cases. To find routes within the local neighborhood of a user (e.g., all users within three hops), we use an efficient bloom filter-based [21] mechanism. To discover longer paths, we use landmark routing [152] to route to the destination's neighborhood and then use bloom filters to reach the destination. Each user creates and publishes a bloom filter (representing her local neighborhood) and a landmark coordinate (representing her location in the global network).

A user's bloom filter represents the set of users within the two-hop neighborhood of the user's trust network. Thus, given a destination's bloom filter, a user can determine whether any of her friends are within the destination's two-hop neighborhood. If so, the user has found the next hop toward the destination. The solution works on arbitrary connected graphs. However, the approach is most efficient in sparse graphs in which the three-hop neighborhood accounts for a small percentage of the total network. Many real-world trust networks, such as social networks, have this property [105].

For long paths, we use landmark routing to reach the destination's neighborhood. A small subset of the user population is chosen as landmarks, and every user in the network determines her hop distance and the next hop to each of these landmarks. The landmarks are selected such that every user is within three hops of at least one landmark. Then, the resultant coordinate system can be used to route to within three hops of any destination user, and the bloom filters to reach the destination. Thus, given a destination user's coordinate, a user can first route to a landmark user who is "near" the destination, and this landmark user can then use bloom filter routing for the last few hops.

We describe these in terms of an interval I, which is the frequency with which the bloom filters and coordinates are recomputed and updated. Typical values of I are on the order of a few days.

8.5.3 Bloom filter routing

Bloom filters are a space-efficient probabilistic data structure for representing set membership. When testing whether an element is in the set, bloom filters have no false negatives, but have a configurable false positive rate [21]. In Ostra, each user U makes available two separate bloom filters: a one-hop bloom filter F^1 and a two-hop bloom filter F^2. The one-hop bloom filter contains all of the direct friends of U, and the two-hop bloom filter contains all of U's friends-of-friends.

Construction

If U is friends with A, B, and C, then F^1 for U would be a bloom filter with the following contents: $F(S_A, S_B, S_C)$ where $F(\cdot)$ represents a bloom filter. However, U enters a friend X in the bloom filter not using X's public identity, but using an alias for X (which we denote S_X) that is only know to X's friends. This ensures the privacy of U's set of friends, since it is impossible to enumerate U's friends given only the bloom filter and the public identifiers of nodes. Specifically, a user M, given U's bloom filter, can determine if U is friends with another user only if M is also friends with the other user herself. Moreover, since each user chooses unique parameters, it is impossible to estimate, given two user's bloom filters, the size of the intersection among the users' friends.

Users construct their two-hop bloom filters by requesting one-hop bloom filters with a specified set of parameters from all of their friends. To construct a two-hop bloom filter, a user then simply perform a bit-wise OR of all of their friends' responses. Additionally, whenever a user creates or removes links, the user resends its one-hop bloom filter to each of her friends, so that they can update their two-hop bloom filters.

Use

When a user A wishes to discover a path to user B, A obtains B's one-hop and two-hop bloom filters using the lookup mechanism of the underlying communication system. A first checks to see if any of her friends appear in B's one-hop bloom filter.

If so, this implies that these friends are direct friends with B. Thus, A has found a path to B.

If none of A's friends appear in B's one-hop bloom filter, then A can be sure that *no* two-hop path exists between herself and B. A then checks for three-hop paths by testing to see if any of her friends appear in B's two-hop bloom filter. If so, then A knows that these friends are friends-of-friends of B. In this case, A has found the first hop on a three hop path to B.

If none of A's friends appear in B's one-hop bloom filter or two-hop bloom filter, this implies that *no* path shorter than three hops exists between A and B. In this case, A uses the coordinates described next to find a path between herself and B.

False positives in bloom filters have the effect of artificially inflating path lengths. A user may, due to a false positive, forward to another user who is no closer to the destination. As we demonstrate in the evaluation, this case is rare and does not affect the eventual success of the route computation.

8.5.4 Landmark routing

To find long paths, users advertise their *coordinates*, which indicate their location in the trust network. A coordinate is a vector of distances, in hops, from a set of *landmark users* in the trust network. A node U's coordinate might be $\{3_M, 7_N\}$, meaning U is 3 hops from M and 7 hops from N. We describe in Section 8.5.4 below how landmarks are selected.

Users determine their coordinate using their friend's coordinates. For each landmark user, the distance from that landmark is the minimum of all of their friends' hop distances plus 1. For example, if U is friends with A and B, and A's coordinate is $\{2_M, 4_N\}$ and B's coordinate is $\{4_M, 7_N\}$, then U's coordinate is $\{3_M, 5_N\}$. Additionally, U records her next-hop for each coordinate. (In the example, U would record that A is the next hop towards M, and either A or B are the next hop towards N).

Friends periodically exchange their coordinates and repeat the same calculation. Given a stable set of landmarks, the calculation converges to a stable set of coordinates. In order to reduce the overhead of coordinate updates, new coordinates are only published by users once per interval I.

Routing

When a user A computes a path to a user B, A obtains B's coordinate through the underlying communication system's lookup service. A then looks for landmarks that appear in both B's and A's coordinate, and are within three hops of B. If such a landmark L exists, then A has found a path to B. This is because A knows how to get to L (by routing via the next hop), and L is able to use B's bloom filter to find a path to B (since B is within 3 hops of L). Additionally, each user along the path can check to see if B can be reached using bloom filter routing, attempting to detect a shorter path.

If A is unable to find a shared landmark that is within 3 hops of B, then A is

unable to route to B. This situation may arise if B is very weakly connected to the network and is not within three hops of any landmark. Thus, users who are not within three hops of a landmark are unreachable by other users via the coordinate mechanism. However, they can still originate communication to other reachable users, as well as receive communication from users within their three-hop radius.

Landmark selection

Next, we discuss how to select landmarks. Too few landmarks limits the reachability of users in the network. Too many landmarks impacts the efficiency of the system, as the sizes of the coordinates grow with each additional landmark. Ideally, one would like to pick the minimal set of users to be landmarks, such that every user in the network is within 3 hops of at least one landmark.

We use a simple distributed landmark selection scheme. Each user periodically checks to see if she is within 3 hops of a landmark. If not, and the user is sufficiently well connected to the network (i.e., she has at least L_{min} friends), the user becomes a landmark herself. This scheme guarantees that all sufficiently well connected users are reachable, but it does not guarantee a minimally-sized set of landmarks.

8.5.5 Decentralized credit update

When the path in the trust network between the sender and receiver has been determined, the credit balances and bounds are updated in a decentralized manner during authorization, classification, and token expiration.

During authorization, the sender sends a signed authorization request message along the path. This request includes a unique identifier, the public key of the destination, and the destination's bloom filter and coordinate. Each user along the path (i) forwards the message, (ii) updates the balances and bounds of the message's incoming and outgoing links according to the rules stated below, (iii) records the destination, request identifier, previous hop, next hop, and expiration time of the request, and (iv) sets a timer for the expiration time. When the destination receives the request, it issues a signed token and sends it directly to the sender.

The link bounds are updated as follows. Each user along the path increments the lower bound L for the next hop, as was done in the centralized Ostra described in Section 8.2. Thus, the state of the network after a token is issued is exactly as shown in Figure 8.4 (b).

During classification, the destination sends a signed classification message along the path in the reverse direction. Each user checks if she has a record of a matching authorization request. If so, the adjustments of the link bounds performed during the authorization are undone, and the link balances are adjusted as described below. The message is then forwarded, and the record is deleted. Otherwise, if no matching record exists, the message is ignored.

The link balances are adjusted as was done in the centralized case. If the message was classified as wanted, the link balances are not changed, as shown in Figure 8.4 (d). However, if the message was classified as unwanted, each user raises the credit balance

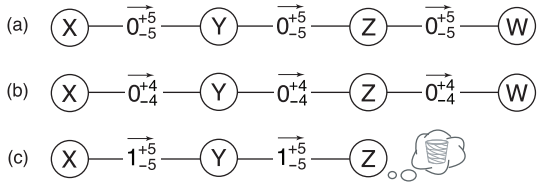


Figure 8.12 : Diagram of how credit exchange occurs when X sends to W, with the penalty for dropping being one credit. The state of the link credits is shown (a) before the message is sent, (b) before the message is classified, and (c) after the timeout T if Z drops the message.

of the next hop in the path (the user to whom the original request was forwarded) and lowers the credit balance of the previous hop (the user from whom the original request was received). In this case, the resultant state of the network is shown in Figure 8.4 (c).

When the timer associated with an authorization request expires, then the user undoes the adjustments made to the link states during the authorization phase and deletes the request record.

Because authorization and classification messages are forwarded by the Ostra agents of users in the trust network, one concern is whether malicious users can simply drop such incoming messages. To protected against this, we provide users with an incentive to forward authorization requests and responses: users penalize the next hop along the path by lowering the next hop's credit if the message does not reach its destination.

Each user along the path adjusts the next hop's upper bound U by a penalty amount during the authorization phase. When the message is classified by the des-

tination, the bound is restored. Otherwise, if a user drops the message, each of the users penalizes the next hop after the timeout T. An example is shown in Figure 8.12: while the message is pending classification (b), both the upper bound U and the lower bound L are changed to account for all possible outcomes. In the case in which Z drops the message (c), X penalizes Y, and Y penalizes Z. Thus, Z is penalized for dropping the message, whereas Y, who properly forwarded the message, has a neutral outcome.

8.5.6 Security and privacy

One concern is whether malicious users can abuse the false positives in the bloom filters to attract request to be routed through them. In order to maintain an acceptable false positive rate, users select the number of hash functions and the bloom filter length so that the number of bits set to 1 is less than a specified values B_{max}. Any bloom filters with more than B_{max} bits set is ignored. Otherwise, malicious users could simply create bloom filters consisting of all 1s, implying that they would be a good choice when routing to any other user. Appropriate settings can be independently determined based on the size of the their two hop neighborhood.

In Ostra, it is not possible to explore the trust network beyond two hops, due to the use of private aliases. By requesting the bloom filters of many nodes, it is however possible to determine the identities of friends of a friend with high probability. However, many deployed social networks, such as FaceBook [49] and LinkedIn [95]

do in fact already allow two-hop browsing, so there is no loss of privacy in many applications.

8.6 Summary

In this chapter, we have presented Ostra, a system that leverages the difficulty in creating and maintaining links in social networks to prevent unwanted communication. Ostra ensures that unwanted communication strains the originator's trust relationships, even if the sender has no direct relationship with the ultimate recipient of the communication. A user who continues to send unwanted communication risks isolation and the eventual inability to communicate. Finally, we demonstrated that Ostra can effectively prevent unwanted communication upon a social network from a real-world site.

Chapter 9

PeerSpective: Leveraging Shared Interest

Over the last decade, the World Wide Web and Web search engines have fundamentally transformed the way people find and share information. Recently, a new form of publishing and locating information, known as online social networking, has become very popular. While numerous studies have focussed on the hyperlinked structure of the Web and have exploited it for searching content, few studies, if any, have examined the information exchange in online social networks.

In the Web, explicit links called hyperlinks between content (typically pages) are the primary tool for structuring information. Hyperlinks are used by authors to embed a page in the Web of related information, by human users to manually browse the Web, and by search engines to crawl the Web to index content, as well as to rank or estimate the relevance of content for a search query.

In contrast to the Web, no explicit links exist between the content (typically photos, videos, and blog postings) stored in social networks. Instead, explicit links between users, who generate or publish the content, serve as the primary structuring tool. For example, in social networking sites like MySpace [111], Orkut [121], and Flickr [52], a link from user A to user B usually indicates that A finds the information published by B interesting or relevant, or A implicitly endorses B's content due to an

established social relationship. Such social links enable users to manually browse for information that is likely of interest to them, and could be used by search tools to index and locate information.

In this chapter, we seek to understand whether the shared interest that these social links represent can be exploited by systems. To answer this question, we design, build, and deploy PeerSpective, a prototype system that leverages the shared interest between users in a social network to produce more relevant Web search results.

Overall, we make three contributions. First, we compare the mechanisms for content publication and location in the Web and online social networks. We argue that search techniques could benefit from integrating the different mechanisms used to find relevant content in the Web and social networks. Second, we present the design of PeerSpective, a Web search system that leverages the shared interest from a social network to improve Web search. Third, we present results from a deployment of PeerSpective that support our contention that shared interest in social networks can be leveraged in systems.

9.1 The Web versus social networks

We begin with a comparison of the Web and social networking systems, with respect to their mechanisms for *publishing* and *locating* content.[1] Publishing refers to the mechanism by which content creators make information available to other users; it

[1]We ignore the mechanisms for distributing content between users as they are similar in both the Web and many current online social networks. In both systems, the content is transferred using

includes the way users relate their content to other content found in the system. Locating refers to the mechanism by which users find information relevant to them; it includes the ways users browse or search the content in the system.

9.1.1 The Web

In the Web, the content typically consists of Web pages written in HTML.

Publishing

Users publish content by placing documents on a Web server. An author places hyperlinks into her page that refer to related pages. She may also ask other authors to include links to her page in their pages. Often, such links are placed deliberately to ensure the page is indexed and ranked highly by search engines.

Locating

Today, the predominant way of locating information on the Web is via a search engine. Modern Web search engines employ sophisticated information retrieval techniques and impressive systems engineering to achieve high-quality search results at massive scale.

The key idea behind search engines like Google is to exploit the hyperlink structure of the Web to determine both the corpus of information they index and the relevance of a Web page relative to a given query [122]. This approach has proven highly

HTTP over TCP, and the users navigate the systems using their Web browser.

effective, because the incident links to a page are strong indicators of the importance or relevance of the page's content in the eyes of other users.

However, hyperlink-based search has some well known limitations. First, while Web search is very effective for relatively static information, it may under-rate or miss recently published content. For a new page to be noticed and appropriately ranked by a search engine, (a) it must be discovered and indexed by the search engine, (b) hyperlinks to the new page must be included in subsequently published or edited pages, and (c) all such links must then be discovered by the search engine.

Second, as search engines determine the relevance of a page by its incident hyperlinks, their rating reflects the interests and biases of the Web community at large. For instance, a search for "Michael Jackson" yields mostly pages with information about the pop star. Computer scientists, however, may find the Web page of a professor with the same name more relevant. Refining the search to find that page is possible but can be tricky, particularly if one does not recall the professor's current affiliation or field of specialization.

Third, the hyperlink structure influences whether a page is included in a search engine's index. Unlinked pages and non-publicly accessible pages are not indexed. Many other pages are not indexed because the search engine deems them insufficiently relevant, due to their location in the hyperlink structure. As a result, obscure, special-interest content is less likely to be accessible via Web search.

9.1.2 Social Networks

Online social networking Web sites have recently exploded in popularity. Sites offer services for finding friends like MySpace [111], Orkut [121], and Friendster [55], for sharing photos like Flickr [52], for sharing videos like YouTube [167] and Google Video [62], and for writing blogs like LiveJournal [97] and BlogSpot [20]. These sites are extremely popular with users: MySpace claims to have over 246 million users, while Facebook and Orkut boast 124 million and 67 million users, respectively. MySpace recently has even been observed to receive more page hits than Google [112].

Examples of online social networking, though, have existed for much longer. For instance, the common practice of placing content on the Web and sending its URL to friends or colleagues is essentially an instance of social networking. Typically, the author has no intention of linking the content; thus, the content remains invisible to users other than the explicit recipients of the URL. The content is advertised not via hyperlinks, but via links between users.

Publishing

Users publish content by posting it on a social networking site. Content is associated with the user who introduced it, and with users who explicitly recommend the content. Explicit links do not generally exist between content instances, and the content can be of any type. Often, the content is temporal in nature (e.g., blog postings), non-textual (e.g., photos and video clips), and may be of interest only to a small audience.

Independent of the content, users maintain links to other users, which indicate trust or shared interest.

Locating

The predominant method of finding information in online social networks is to navigate through the social network, browsing content introduced or recommended by other users. Some sites also provide keyword-based search for textual or tagged content. Additionally, other sites have 'top-10' lists showing the most popular content, where the popularity is determined according to how often users have accessed the content or based on explicit recommendations provided by users.

Moreover, social networks enable users to find timely, relevant and reliable information. This is because users can browse adjacent regions of their social network, which likely consist of users with shared interests or mutual trust. Since the content can be non-textual, obscure, or short-lived, it may be hard to find by the way of Web search. For example, blog posts are generally of short-term interest, videos and photos are non-textual, and all three types of content tend to be of interest to a limited audience.

Content in social networks can also be rated rapidly, based on implicit and explicit feedback of a large community of content consumers. In contrast, Web search relies on the slower process of discovering hyperlinks in the Web, which are created by a relatively smaller number of content authors. Since content rating in social networks

is performed by the content consumers, rather than the producers, content introduced into the network can by rated almost immediately.

9.1.3 Leveraging shared interest in Web search

Today, the information stored in different social networks and in the Web is mostly disjoint. Each system has its own method of searching information. While search companies have started to address this issue with specialized search tools for RSS-based news feeds and for blogs, there is no unified search tool that locates information across different systems. Social network-based search methods are not generally used in the Web, though services like Google Scholar support search facilities tailored to a specific community. Given that end users access both the Web and the social networks from the same browsers, it seems natural to unify the methods to find information as well.

In this chapter, we explore the idea of integrating Web search with search in social networks, with the goal of leveraging the shared interest that exists between users. We believe that such an approach could combine the strengths of both types of systems: simultaneously exploiting the information contained in hyperlinks, and information from implicit and explicit user feedback; leveraging the huge investment in conventional Web search, while also ranking search results relative to the interests of a social network; and locating timely, short-lived, non-textual or special-interest information alongside the vast amounts of long-lived and textual information on the

Web.

9.2 PeerSpective

Our discussion above suggests that (a) a growing body of Internet content cannot be retrieved by traditional Web search as it is not well-connected to the hyperlinked Web, and that (b) social network links can be leveraged to improve the quality of search results. To explore this potential, we designed, built, and deployed the PeerSpective system. In this section, we describe the design of PeerSpective and discuss our experimental results.

9.2.1 Design

PeerSpective is designed as a lightweight HTTP proxy. Thus, each PeerSpective user configures their Web browser to use PeerSpective as a HTTP proxy, which allows PeerSpective to observe the content of pages that the client browses to. PeerSpective decodes these pages, parses out the text for know document formats (currently HTML and PDF), and then indexes the documents with the enclosed text.

When the user performs a Google search, the proxy transparently forwards the query to both Google, as normal, as well as the PeerSpective proxies of other users in the social network. Each proxy (including the user's local proxy) executes the query on the local index and returns the result to the sender.

The results are then collated and presented alongside the Google results as shown

in Figure 9.1. To do so, PeerSpective modifies the returned HTML from Google in order to include the PeerSpective results. Thus, to use PeerSpective, the user does not have to do any work, beyond the initial setup of PeerSpective. The PeerSpective index is populated as the user browses the Web normally, and Web search results are automatically inserted into the Google results page.

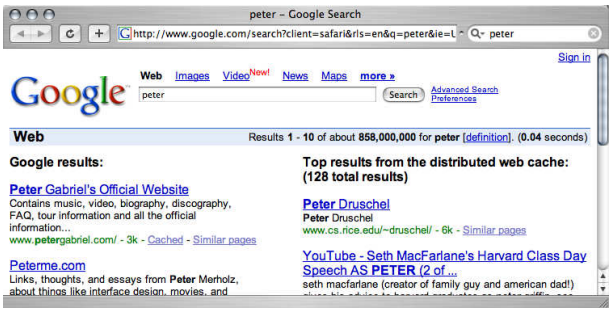


Figure 9.1 : Screenshot of our PeerSpective search interface. Results from the distributed cache appear alongside the normal Google results.

Our PeerSpective implementation is built using the Lucene [98] text search engine and the FreePastry [54] peer-to-peer overlay. We configured Lucene to follow Google's query language, so that search qualifiers such as '+', '-', and quotes would be consistent across both systems. We also configured Lucene to automatically remove pages older than 30 days in order to prevent the index from getting stale.

We ranked the results obtained from PeerSpective by multiplying the Lucene score of a search result by the Google PageRank of that result and adding the scores from all users who previously viewed the result. Thus, PeerSpective's ranking takes advantage of both the hyperlinks of the Web (via Google's PageRank) and the social links of the user community.

9.2.2 Privacy

One potential concern with the PeerSpective architecture is the privacy of the users. PeerSpective indexes browsed Web pages, which sometimes contain sensitive information. In order to mitigate the privacy impact of PeerSpective on users, we configured PeerSpective to only serve HTTP, and specifically not HTTPS, traffic. As many privacy-sensitive services, such as online banking and email, use HTTPS, this prevents such sites from being indexed by PeerSpective. We also configured PeerSpective to respect the "Cache-Control" header returned by servers by only indexing pages that were labeled "Cache-Control: public".

When using PeerSpective, users only see aggregated results, and are not aware of which user returned which result. Thus, users in PeerSpective have k-anonymity, where k is the size of the group running PeerSpective. Finally, we configured a simple control panel for PeerSpective that allowed users to browse their local index, and to remove any pages that they did not wish to be included.

9.2.3 Experimental methodology

We recruited a group of 10 graduate students and researchers the Max Planck Institute for Software Systems to run PeerSpective. We present measurements and experiences from a one month long experimental deployment. During this time, the 10 users issued 439,384 HTTP requests covering 198,492 distinct URLs. Only 25.9% of the HTTP requests were of content type `text/html` or `application/pdf`, meaning they could

be indexed by our proxy. The remaining requests consisted of images, javascript, and other miscellaneous types.

Given that our user base is small, includes the authors, and represents a single community with highly specialized interests, we cannot claim that our results would be representative of a deployment with a larger, diverse user base. However, we believe our results indicate the potential of social network-based Web search.

9.2.4 Limits of hyperlink-based search

Even the best Web search engines do not index content that is not well linked to the general Web or content that is not publicly available. So, our first goal is to understand and quantify the Internet content that is viewed by users, but is not captured by the search engines. We would also like to know how much of this content is already indexed by another user in PeerSpective.

To estimate the limits of hyperlink-based search, we check what fraction of the URLs actually visited by the users are not indexed by Google. There are a number of reasons why a page may not be indexed by Google: (a) the page could be *too new*, such a as blog posting or news article; (b) the page could be in the *deep web* and not well-connected enough for Google to choose to crawl it; or (c) the page could be in the *dark web*, where it is not publicly available or is not referred to by any other page.

For each HTTP request, we checked whether Google's index contains the URL, and if some peer in PeerSpective has previously viewed the URL. Since search engines

only index static HTML content, we considered only URLs of indexable content types that did not have any GET or POST parameters and ended in either `.html` or `.htm`. Further, we discarded URLs with an auto-refresh feature (such as the scoreboard sites for sports), as they would artificially bias the results against Google. This left us with 6,679 requests for 3,987 URLs.

Our analysis shows that Google's index covers only 62.5% of the requests, representing 68.1% of the distinct URLs. This implies that about one third of all URLs requested by our users cannot be retrieved by searching Google! Our analysis also showed that the union of the PeerSpective peer indexes covers about 30.4% of the requested URLs. While PeerSpective achieves only half of the coverage of Google's index, it does this with a much smaller size: at the end of the experiment, the PeerSpective indexes contained 51,410 URLs, compared to Google's index of over 8 billion URLs.

Additionally, we found that 13.3% of the URLs viewed were contained in PeerSpective but not in Google's index. These documents were not available via Google's search engine but had been requested before by someone in the peer network. This increase in coverage amounts to a 19.5% improvement by PeerSpective compared to normal Google search. It is worth noting that, for our small social network of computer science researchers, this improvement in coverage was possible by adding just a few thousand URLs to a Google index containing billions or URLs.

Our results naturally raise the question, what are these documents that are of a of

URL	Too new	Deep web	Dark web
jwz.livejournal.com/413222.html	✓	✓	
www.mpi-sws.mpg.de/ ... /pres0031.html		✓	
sandiego.craigslist.org/w4m/179184549.html	✓	✓	
edition.cnn.com/ ... /italy.nesta/index.html	✓		
72. ... .163/status.asp			✓
www.itv.com/news/ ... a8e4b6ea.html	✓		
www.stat.rice.edu/∼riedi/ ... /target21.html		✓	
amarok.kde.org/forum/index.php/board,9.20.html	✓	✓	

Table 9.1 : Sample URLs that were not indexed by Google. We manually inspected the URLS to determine the likely reason for not being in Google's index, as discussed in Section 9.2.4.

interest to our users, but are not indexed by Google? We manually analyzed a number of such URLs and show a random sample of them in Table 9.1. We additionally list the likely reasons why each URL does not appear in Google's index.

9.2.5 Benefits of social network-based search

Another challenge facing search engines is ranking all the indexed documents in the order of their relevance to a user's query. Ranking is crucial for search, as most users rarely go beyond the first few query results [146]. Our goal here is to study how often users click on query results from PeerSpective as opposed to Google. As shown in Figure 9.1, our users are presented with results from both Google and PeerSpective for every Google query.

During the course of the month, we observed 1,730 Google searches. While Google's first result page contained an average of 9.45 results, our smaller PeerSpective index resulted in an average of 5.17 results on the first page. Of the 1,730 queries, 1,079 (62.3%) resulted in clicks on one or more search result links, 307 (17.7%) were followed by a refined query, and after the remaining 344 (19.8%), the user gave up. We found that 933 (86.5%) of the clicked results were returned only by Google, 83 (7.7%) of the clicked results were returned only by PeerSpective, and 63 (5.7%) of the clicked results were returned by both. This amounts to a 9% improvement in search result clicks over Google alone, as 83 of the search result clicks would not have been possible without PeerSpective.

It should be kept in mind that this 9% improvement over Google, considered by many to be the gold standard for Web search engineering, was achieved by a simple, very small, social network-based system quickly put together by three systems researchers over a period of a few days. Based on our early experience, we feel that these results suggest inherent advantages of using social links for search, which could be exploited better with more careful engineering.

9.3 Discussion

To better understand the cases when PeerSpective search results outperform Google results, we manually analyzed the corresponding queries and result clicks. We show a random sample of the data we analyzed in Table 9.2. We observed that the reasons

for clicks on PeerSpective results fall into three categories, described below.

9.3.1 Disambiguation

Some search terms have multiple meanings depending on the context. Search engines generally assume the most popular term definition. Social networks can take advantage of the fact that communities tend to share definitions or interpretation of such terms. An example for disambiguation is shown in Table 9.2, where a user's query for "bus" yielded the local bus schedule, as it is the page with this keyword that is most visited by local users in the network.

Query	**Page clicked on**	**D**	**R**	**S**
`bus`	Saarbrücken bus schedule	✓	✓	
`stefan`	FIFA World Cup site			✓
`peter`	Peter Druschel's home page	✓		
`serbian currency`	XE.com exchange rates		✓	
`coolstreaming`	CoolStreaming INFOCOM paper		✓	
`moose`	Northwest Airlines' contract of carriage			✓
`münchen`	Peter Druschel's homepage			✓

Table 9.2 : Sample search queries for which PeerSpective returned results not in Google. The results are categorized into the three different scenarios of disambiguation (D), ranking (R), and serendipity (S) discussed in Section 9.3.

9.3.2 Ranking

Search engines rank all relevant documents and return the top of the resulting list. Social networks can inform and bias the ranking algorithm, since nearby users in the network often find similar sets of pages relevant. An example we observed is a search with the term "coolstreaming". A Google search ranks most highly popular sites (such as Wikipedia) discussing the CoolStreaming technique for P2P streaming of multimedia content. PeerSpective ranked the INFOCOM paper describing CoolStreaming at the top, as it is most relevant to our researchers.

9.3.3 Serendipity

While browsing the Web, users often discover interesting information by accident, clicking on links that they had not intended to query for. This process, termed serendipity, is an integral part of the Web browsing experience. Search results from PeerSpective provide ample opportunity for such discoveries. For example, while looking for information about "München" (Munich), one of our users discovered that a fellow researcher attended school in München, thus finding a convenient source of information about the city.

9.4 Summary

Online social networking enables new forms of information exchange in the Internet. First, end users can very easily and conveniently publish information, without

necessarily linking it to the wider Web. Second, social networks make it possible to locate and access information that was previously exchanged by "word of mouth", that is, by explicit communication between individuals. Third, unlike Web search engines, which organize the world of information according to popular opinion, social networks can organize the world of information according to the shared interest of smaller groups of individuals.

In this chapter, we explored the potential of the integration of the Web and social network search technologies. In a small-scale experiment, we found that a significant fraction of URLs requested by our users cannot be retrieved by today's most popular search engine, as the URLs are too new, of interest to only a small population, or not publicly available. However, we found that by including pages browsed to by friends in a social network, the index coverage could be increased significantly. Moreover, we found that by including these pages in search results, a noticeable improvement in click-rate was observed, underscoring the potential for leveraging the shared interest in social networks.

Chapter 10

Conclusion

Originally conceived to solve computational problems in science, defense, and business, computer systems now augment many human activities, including communication and social interaction. Today, millions of people use information technology to work, play, read, learn, socialize, connect, and express themselves. This broad range of new applications inspired the work in this thesis, where we have measured and analyzed the properties of online social networks, and designed, deployed, and evaluated new information systems that exploit the properties of these networks. In the following sections, we describe the high-level contributions of this thesis and discuss potential future research directions.

10.1 Summary

Recently, online social networks have exploded in popularity. MySpace (over 246 million users) and Facebook (over 124 million users) are examples of wildly popular networks that are used to find and organize contacts; numerous other sites are used to share photos, videos, blogs, and news items. Despite the massive popularity of online social networks, surprisingly little is known about how people are using them to connect and share content. To better understand the structure of online social

networks, we conducted a large-scale measurement study that collected data on the social networks of four popular sites, covering over 11 million users and 328 million links. Surprisingly, the results showed that the social network graphs contained in different sites shared a number of graph-theoretic properties, even though the sites have very different goals, mechanisms, and policies. These results have a number of implications for system designers. For example, all the networks contain a dense core of popular users that holds the network together; any information flowing through the network must traverse this core, implying that these users will naturally have significant influence on the spread of information. This was the first study to collect data at large scale and the first study to collect data on multiple social networks. Moreover, we were the first to make the collected data available to the research community; the data is currently in use by more than 100 research groups.

The static graph structure present in online social networks reflects the process by which users create links; to understand this process, it is necessary to observe how the networks change and grow over time. Thus, we conducted a second measurement study that collected data from multiple online social networks by crawling the network daily for more than three months. This was the first study to collect network growth data at significant scale and at fine temporal resolution. The analysis of this growth data provides intuitive explanations for a number of the observed structural properties. The results of our study can be used as the basis for constructing synthetic networks that reflect both global and local characteristics of online social

networks, leading to better structural and growth models. We also made the data in this study available to the research community and it is currently in use by more than 25 research groups.

We have also examined the community structure of online social networks. We found that individual users are often members of multiple overlapping communities, but that existing algorithms for detecting communities do not perform well on real data from an online social network. We addressed this limitation by devising a new algorithm that can accurately detect multiple overlapping communities when given information about a small subset of the community members. In practice, even if only 10% of users provide community information to social networking sites, the remaining community members can be determined by this algorithm with high accuracy. We demonstrated that this approach can identify communities at a range of scales on a university network: small communities such as sports teams, larger communities such as dormitories, and even very large communities such as every student matriculating in the same year.

While valuable, the measurement studies described above are not an end *per se*, rather, they are a first-order concern when trying to build better systems. For example, links between users in online social networks can represent trust (e.g., users who know each other in the offline world) and shared interest (e.g., users who belong to the same community). We have built two new information systems that exploit each of these properties to solve open problems – these are briefly described below.

First, we demonstrated how to use social networks to address the problem of unwanted communication. Internet-based communication systems such as email, IM, VoIP, online social networks, and content-sharing sites allow communication at near zero marginal cost to users. Unfortunately, this property can be abused for the purpose of spam, unsolicited marketing, propaganda, or disruption of legitimate communication.

Using insights on trust in online social networks, we presented Ostra, a novel mechanism that exploits trust relationships among users to block unwanted communication. Ostra uses an existing network, such as a social network, to connect senders and receivers via chains of pairwise relationships, keeping a credit score associated with each link in the network. A user's links are penalized when she sends unwanted communication, and users whose links have all run out of credit must wait to send messages. Ostra is novel in that it is the first system that can, without assuming strong user identities, effectively ensure that having multiple identities does not benefit the attacker. Ostra is sufficiently general that it can be used not only on social networks, but on any network in which links require some effort to form and maintain.

Second, we showed how social networks can be used to mitigate the privacy and access challenges that arise when the amount of shared content is growing at an exponential rate. In particular, the growing amount of shared content on online social networks is leading to two pressing challenges. First, since users are sharing increasingly personal information, the issue of privacy and access control is becoming more

important. Second, since the volume of shared content is growing at an exponential rate, finding relevant information is becoming more difficult.

Using insights from our measurement studies, we proposed using communities to address these growing dilemmas. Communities can aid in both access control (since they represent a natural middle ground between a user's immediate friends and the rest of the world) and in information retrieval (since they often represent sets of users with shared interests). To demonstrate this approach in a deployed system, we presented the design, implementation, and deployment of PeerSpective, a system that uses a social network to provide better search results than socially-oblivious search engines like Google. A preliminary version of PeerSpective showed a 7% improvement in click-rate over existing Web search technologies, underscoring the potential of leveraging communities in social networks.

10.2 Future work

So far, we have studied how users interact on today's online social networks and observed how the trust and shared interest that links represent can be used to solve systems problems. The recent explosion in popularity of online social networks underscores the continuing integration of computing in our daily lives, a trend that provides a number of interesting research challenges. In the following paragraphs, we outline a few of these challenges.

In this thesis, we have focused exclusively on the user graph of social networking

sites; many of these sites allow users to host content, which in turn can be linked to other users and content. Establishing the structure and dynamics of the content graph is an open problem, the solution to which will enable us to understand how content is introduced in these systems, how data gains popularity, how users interact with popular versus personal data, and whether these trends can be hardened to prevent deliberate manipulation. Similarly, the data we collected on the growth of online social networks can be used to test previously proposed growth models to see how well they match the observations, as well as to guide the development of new models based on empirical data.

Users often share very personal information on today's online social networks, with little regard for who will be allowed to view the content. Anecdotal evidence suggests that this often results in unintended consequences, ranging from public embarrassment to job loss. The underlying problem is, essentially, ensuring privacy for users while allowing them to share information and knowledge freely. This problem has aspects that span the areas of security, systems, and interface design. Thus, one challenge is to design mechanisms that enable the wide-spread sharing that users desire while ensuring that users understand who else is able to access their content. One potential first step to solving this problem is to use communities as an abstraction for expressing privacy policies, allowing users to share content with more than just their friends but not necessarily with the entire world.

Another problem concerns ensuring the relevance of information obtained. In prior

information sharing networks, ranking systems have proved invaluable for finding relevant information – the most well-known example is PageRank for Web documents. However, the content shared in emerging systems like online social networks is different from previous systems: the content items rarely have links to other content items; rather, the links connect the users themselves. Thus, a new approach to finding relevant information is needed that can compute the reliability of a given piece of information based on the combined reputation of the users who created or endorsed it. Whereas a Web page linked to by `nytimes.com` is likely important (because many important pages link to `nytimes.com`), it is unclear whether this same transitive importance will apply to links between users. Additionally, since this computation is based on a social network of users, the links between whom may represent shared interest, it may be possible to easily compute customized rankings for each user's interests. The PeerSpective system represents a first step in this direction, however, the general problem of finding relevant content remains an open challenge.

Bibliography

[1] A9 Search. `http://www.a9.com`.

[2] Martín Abadi, Andrew D. Birrell, Mike Burrows, Frank Dabek, and Ted Wobber. Bankable postage for network services. In *Proceedings of the Asian Computing Science Conference (ASIAN'03)*, Mumbai, India, December 2003.

[3] Lada A. Adamic. The Small World Web. In *Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries (ECDL'99)*, Paris, France, September 1999.

[4] Lada A. Adamic, Orkut Buyukkokten, and Eytan Adar. A social network caught in the Web. *First Monday*, 8(6), 2003.

[5] Sharad Agarwal, Venkata N. Padmanabhan, and Dilip A. Joseph. Addressing email loss with suremail: Measurement, design, and evaluation. In *Proceedings of the 2007 Usenix Annual Technical Conference (USENIX'07)*, June 2007.

[6] Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong. Analysis of Topological Characteristics of Huge Online Social Networking Services. In *Proceedings of the 16th international conference on World Wide Web (WWW'07)*, Banff, Canada, May 2007.

[7] Réka Albert, Hawoong Jeong, and Albert-László Barabási. The Diameter of the World Wide Web. *Nature*, 401:130, 1999.

[8] David Alderson and Lun Li. Diversity of graphs with highly variable connectivity. *Physics Review E*, 75, 2007.

[9] Luís A. Nunes Amaral, Antonio Scala, Marc Barthélémy, and H. Eugene Stanley. Classes of small-world networks. *Proceedings of the National Academy of Sciences (PNAS)*, 97:11149–11152, 2000.

[10] Chris Anderson. *The Long Tail*. Hyperion, New York, NY, USA, 2006.

[11] Asad Awan, Ronaldo A. Ferreira, Suresh Jagannathan, and Ananth Grama. Distributed uniform sampling in real-world networks. Technical Report CSD-TR-04-029, Purdue University, 2004.

[12] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group Formation in Large Social Networks: Membership, Growth, and Evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*, Philadelphia, PA, August 2006.

[13] James P. Bagrow. Evaluating local community methods in networks. *Journal of Statistical Mechanics*, 2008(5), 2008.

[14] James P. Bagrow and Erik M. Bollt. Local method for detecting communities. *Physical Review E*, 72(4), 2005.

[15] Albert-László Barabási and Réka Albert. Emergence of Scaling in Random Networks. *Science*, 286:509–512, 1999.

[16] Jeffrey Baumes, Mark Goldberg, Mukkai Krishnamoorthy, Malik Magdon-Ismail, and Nathan Preston. Finding communities by clustering a graph into overlapping subgraphs. In *Proceedings of the International Conference on Applied Computing (IADIS'05)*, pages 27–36, Algarve, Portugal, February 2005.

[17] Jeffrey Baumes, Mark Goldberg, and Malik Magdon-Ismail. Efficient identification of overlapping communities. In *Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI'05)*, pages 27–36, Atlanta, GA, May 2005.

[18] Luca Becchetti, Carlos Castillo, Debora Donato, and Adriano Fazzone. A Comparison of Sampling Techniques for Web Graph Characterization. In *Proceedings of the Workshop on Link Analysis (LinkKDD'06)*, Philadelphia, PA, August 2006.

[19] Matthias Bender, Sebastian Michel, Peter Triantafillou, Gerhard Weikum, and Christian Zimmer. MINERVA: Collaborative P2P search. In *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB'05)*, Trondheim, Norway, August 2005.

[20] BlogSpot. `http://www.blogspot.com`.

[21] Burton H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, 1970.

[22] Valentino Braitenberg and Almut Schz. *Anatomy of the Cortex: Statistics and Geometry*. Springer-Verlag, Berlin, Germany, 1991.

[23] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph Structure in the Web: Experiments and Models. In *Proceedings of the 9th International Conference on the World Wide Web (WWW'00)*, Amsterdam, May 2000.

[24] CAIDA AS Topology. `http://as-rank.caida.org/data/`.

[25] Andrea Capocci, Vito D. P. Servedio, Francesca Colaiori, Luciana S. Buriol, Debora Donato, Stefano Leonardi, and Guido Caldarelli. Preferential attachment in the growth of social networks: The Internet encyclopedia Wikipedia. *Physics Review E*, 74, 2006.

[26] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC'07)*, San Diego, CA, October 2007.

[27] Hyunseok Chang, Sugih Jamin, and Walter Willinger. To peer or not to peer: Modeling the evolution of the internet's as-level topology. In *Proceedings of the*

25th Conference on Computer Communications (INFOCOM'06), Barcelona, Spain, April 2006.

[28] Anurat Chapanond, Mukkai S. Krishnamoorthy, and Bulent Yener. Graph Theoretic and Spectral Analysis of Enron Email Data. *Computational & Mathematical Organization Theory*, 11(3), October 2005.

[29] Hyunwoo Chun, Haewoon Kwak, Young-Ho Eom, Yong-Yeol Ahn, Sue Moon, and Hawoong Jeong. Comparison of Online Social Relations in Terms of Volume vs. Interaction: A Case Study of Cyworld. In *Proceedings of the 6th ACM/Usenix Internet Measurement Conference (IMC'08)*, Vouliagmeni, Greece, October 2008.

[30] Classmates.com. `http://www.classmates.com`.

[31] Aaron Clauset. Finding local community structure in networks. *Physical Review E*, 72, 2005.

[32] Aaron Clauset, Mark E. J. Newman, and Christopher Moore. Finding community structure in very large networks. *Physical Review E*, 70(6), 2004.

[33] Aaron Clauset, Cosma R. Shalizi, and Mark E. J. Newman. Power-law distributions in empirical data, June 2007. `http://arxiv.org/abs/0706.1062v1`.

[34] CyWorld. `http://www.cyworld.com`.

[35] danah boyd. Friendster and publicly articulated social networks. In *Proceedings of the Conference on Human Factors and Computing Systems (CHI'04)*, Vienna, Austria, April 2004.

[36] danah boyd. Friends, Friendsters, and Top 8: Writing community into being on social network sites. *First Monday*, 11(12), 2006.

[37] danah boyd. *Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life*. MIT Press, Cambridge, MA, 2007.

[38] danah boyd and Nicole B. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 2007.

[39] danah boyd and Jeffery Heer. Profiles as conversation: Networked identity performance on friendster. In *Proceedings of the Hawai'i International Conference on System Sciences (HICSS-39)*, Kauai, HI, January 2006.

[40] del.icio.us. `http://del.icio.us`.

[41] Digg. `http://www.digg.com`.

[42] Peter Sheridan Dodds and Duncan J. Watts. A Generalized Model of Social and Biological Contagion. *Journal of Theoretical Biology*, 102(32):11157–11162, 2005.

[43] Pedro Domingos and Matt Richardson. Mining the Network Value of Customers. In *Proceedings of the 7th ACM SIGKDD International Conference on*

Knowledge Discovery and Data Mining (KDD'01), San Francisco, CA, August 2001.

[44] John Douceur. The Sybil Attack. In *Proceedings of the 1st International Workshop on Peer-To-Peer Systems (IPTPS'02)*, Cambridge, MA, March 2002.

[45] dSPAM. `http://dspam.nuclearelephant.com`.

[46] Nan Du, Bin Wu, Xin Pei, Bai Wang, and Liutong Xu. Community detection in large-scale social networks. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (WebKDD/SNA-KDD'07)*, San Jose, California, 2007.

[47] Robert I. M. Dunbar. Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*, 16(4):681–735, 1993.

[48] Paul Erdös and Alfréd Rényi. On Random Graphs I. *Publicationes Mathematicae Debrecen*, 5:290–297, 1959.

[49] FaceBook. `http://www.facebook.com`.

[50] Scott E. Fahlman. Selling interrupt rights: A way to control unwanted e-mail and telephone calls. *IBM Systems Journal*, 41(4):759–766, 2002.

[51] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On Power-Law Relationships of the Internet Topology. In *Proceedings of the Annual Conference*

of the ACM Special Interest Group on Data Communication (SIGCOMM'99), Cambridge, MA, August 1999.

[52] Flickr. `http://www.flickr.com`.

[53] Ana L. N. Fred and Anil K. Jain. Robust data clustering. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)*, Madison, WI, June 2003.

[54] FreePastry Project. `http://www.freepastry.org`.

[55] Friendster. `http://www.friendster.com`.

[56] Diego Garlaschelli and Maria Loffredo. Patterns of link reciprocity in directed networks. *Physics Review Letters*, 93, 2004.

[57] Scott Garriss, Michael Kaminsky, Michael J. Freedman, Brad Karp, David Mazières, and Haifeng Yu. RE: Reliable Email. In *Proceedings of the 3rd Symposium on Networked Systems Design and Implementation (NSDI'06)*, San Jose, CA, May 2006.

[58] Michelle Girvan and Mark E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences (PNAS)*, 99(12):7821–7826, June 2002.

[59] Goodmail Systems. `http://www.goodmailsystems.com`.

[60] Google Co-op. `http://www.google.com/coop/`.

[61] Google Personalized Search. `http://www.google.com/psearch`.

[62] Google Video. `http://video.google.com`.

[63] Mark Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, 78(6), 1973.

[64] Saul Hansell. Internet Is Losing Ground in Battle Against Spam. *The New York Times*, April 22, 2003.

[65] Logan G. Harbaugh. Spam-proof your in-box. *PCWorld*, May 2004.

[66] Jason Hartline, Vahab S. Mirrokni, and Mukund Sundararajan. Optimal Marketing Strategies over Social Networks. In *Proceedings of the 17th international conference on World Wide Web (WWW'08)*, Beijing, China, April 2008.

[67] Petter Holme and Beom Jun Kim. Growing scale-free networks with tunable clustering. *Physical Review E*, 65, 2002.

[68] Glen Jeh and Jennifer Widom. Scaling personalized web search. In *Proceedings of the 12th International World Wide Web Conference (WWW'03)*, Budapest, Hungary, May 2003.

[69] Hawoong Jeong, Zoltan Neda, and Albert-László Barabási. Measuring preferential attachment for evolving networks. *Europhysics Letters*, 61, 2003.

[70] Emily M. Jin, Michelle Grivan, and M.E.J. Newman. The structure of growing social networks. *Phys. Rev. E*, 64, 2001.

[71] Ravi Kannan, Santosh Vempala, and Adrian Vetta. On clusterings: Good, bad and spectral. *Journal of the ACM*, 51(3):497–515, May 2004.

[72] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the Spread of Influence through a Social Network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03)*, Washington, DC, August 2003.

[73] Brian W. Kernighan and Shen Lin. An efficient heuristic procedure for partitioning graphs. *Bell System Technical Journal*, 49:291–307, 1970.

[74] Jon Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46:604–632, 1999.

[75] Jon Kleinberg. Navigation in a Small World. *Nature*, 406:845–845, 2000.

[76] Jon Kleinberg. The Small-World Phenomenon: An Algorithmic Perspective. In *Proceedings of the 32nd ACM Symposium on Theory of Computing (STOC'00)*, Portland, OR, May 2000.

[77] Jon Kleinberg and Steve Lawrence. The Structure of the Web. *Science*, 294:1849–1850, 2001.

[78] Jon Kleinberg and Ronitt Rubinfeld. Short paths in expander graphs. In *IEEE Symposium on Foundations of Computer Science (FOCS'96)*, Burlington, VT, October 1996.

[79] Konstantin Klemm and Victor M. Eguiluz. Highly clustered scale-free networks. *Physical Review E*, 65, 2002.

[80] Gueorgi Kossinets and Duncan J. Watts. Empirical Analysis of an Evolving Social Network. *Science*, 311:88–90, 2006.

[81] Ravi Kumar, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins. On the Bursty Evolution of Blogspace. In *Proceedings of the 12th International Conference on World Wide Web (WWW'03)*, Budapest, Hungary, May 2003.

[82] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and Evolution of Online Social Networks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'06)*, Philadelphia, PA, August 2006.

[83] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the Web for Emerging Cyber-Communities. *Computer Networks*, 31:1481–1493, 1999.

[84] Sailesh Kumar, Sarang Dharmapurikar, Fang Yu, Patrick Crowley, and Jonathan Turner. Algorithms to accelerate multiple regular expressions matching for deep packet inspection. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM'06)*, Pisa, Italy, August 2006.

[85] Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong. Statistical properties of sampled networks. *Physical Review E*, 73, 2006.

[86] Seungjoon Lee, Rob Sherwood, and Bobby Bhattacharjee. Cooperative peer groups in NICE. In *Proceedings of the Conference on Computer Communications (INFOCOM'03)*, San Francisco, CA, March 2003.

[87] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (ACM TKDD)*, 1, 2007.

[88] John R. Levine. An overview of e-postage. 2003. `http://www.taugh.com/epostage.pdf`.

[89] Jinyang Li and Frank Dabek. F2F: Reliable storage in open networks. In *Proceedings of the 5th International Workshop on Peer-to-Peer Systems (IPTPS'06)*, Santa Barbara, CA, May 2006.

[90] Jinyang Li, Boon Thau Loo, Joe Hellerstein, Frans Kaashoek, David R. Karger, and Robert Morris. On the feasibility of peer-to-peer web indexing and search. In *Proceedings of the 2nd International Workshop on Peer-to-Peer Systems (IPTPS'03)*, Berkeley, CA, February 2003.

[91] Lun Li, David Alderson, John C. Doyle, and Walter Willinger. Towards a Theory of Scale-Free Graphs: Definitions, Properties, and Implications. *Internet Mathematics*, 2(4):431–523, 2006.

[92] Xin Li, Bing Liu, and Philip S. Yu. Discovering overlapping communities of named entities. In *Knowledge Discovery in Databases: PKDD 2006 (LNCS 4213*, pages 593–600. Springer, 2006.

[93] David Liben-Nowell and Jon Kleinberg. The Link Prediction Problem for Social Networks. In *Proceedings of the 2003 ACM International Conference on Information and Knowledge Management (CIKM'03)*, New Orleans, LA, November 2003.

[94] David Liben-Nowell, Jasmine Novak, Ravi Kumar, and Prabhakar Raghavanand Andrew Tomkins. Geographic Routing in Social Networks. *Proceedings of the National Academy of Sciences (PNAS)*, 102:11623–1162, 2005.

[95] LinkedIn. `http://www.linkedin.com`.

[96] List of Social Networking Sites. `http://en.wikipedia.org/wiki/List_of_social_networking_websites`.

[97] LiveJournal. `http://www.livejournal.com`.

[98] Lucene Search Engine. `http://lucene.apache.org`.

[99] Feng Luo, James Z. Wang, and Eric Promislow. Exploring local community structures in large networks. *Web Intelligent and Agent Systems*, 6(4):387–400, 2008.

[100] MAAY Search Engine. `http://maay.netofpeers.net`.

[101] Priya Mahadevan, Dmitri Krioukov, Kevin Fall, and Amin Vahdat. Systematic Topology Analysis and Generation Using Degree Correlations. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM'06)*, Pisa, Italy, August 2006.

[102] Stanley Milgram. The small world problem. *Psychology Today*, 2(60), 1967.

[103] Gilad Mishne, David Carmel, and Ronny Lempel. Blocking blog spam with language model disagreement. In *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb'05)*, Chiba, Japan, May 2005.

[104] Alan Mislove, Krishna P. Gummadi, and Peter Druschel. Exploiting social networks for Internet search. In *Proceedings of the 5th Workshop on Hot Topics in Networks (HotNets-V)*, Irvine, CA, November 2006.

[105] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC'07)*, San Diego, CA, October 2007.

[106] Michael Mitzenmacher. A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics*, 1(2):226–251, 2004.

[107] Michael Mitzenmacher. Editorial: The Future of Power Law Research. *Internet Mathematics*, 2(4):525–534, 2006.

[108] Mike Molloy and Bruce Reed. A critical point for random graphs with a given degree distribution. *Random Structures and Algorithms*, 6:161–179, 1995.

[109] Mike Molloy and Bruce Reed. The size of the giant component of a random graph with a given degree sequence. *Combinatorics, Probability and Computing*, 7:295–305, 1998.

[110] Ruggero Morselli, Bobby Bhattacharjee, Jonathan Katz, and Michael A. Marsh. Keychains: A Decentralized Public-Key Infrastructure. Technical Report CS-TR-4788, University of Maryland, 2006.

[111] MySpace. `http://www.myspace.com`.

[112] MySpace is the number one website in the U.S. according to Hitwise. HitWise Press Release, July, 11, 2006. `http://www.hitwise.com/press-center/hitwiseHS2004/social-networking-june-2006.php`.

[113] Atif Nazir, Saqib Raza, and Chen-Nee Chuah. Unveiling Facebook: A Measurement Study of Social Network Based Applications. In *Proceedings of the 6th ACM/Usenix Internet Measurement Conference (IMC'08)*, Vouliagmeni, Greece, October 2008.

[114] Mark E. J. Newman. Clustering and preferential attachment in growing networks. *Physics Review E*, 64, 2001.

[115] Mark E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences (PNAS)*, 98:409–415, 2001.

[116] Mark E. J. Newman. Mixing patterns in networks. *Physics Review E*, 67, 2003.

[117] Mark E. J. Newman. The structure and function of complex networks. *Society for Industrial and Applied Mathematics (SIAM) Review*, 45:167–256, 2003.

[118] Mark E. J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6), 2004.

[119] Mark E. J. Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:026113, 2004.

[120] Andrew M. Odlyzko. The case against micropayments. In *Proceedings of Financial Cryptography: 7th International Conference*, Jan 2003.

[121] Orkut. `http://www.orkut.com`.

[122] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford University, 1998.

[123] Gergely Palla, Imre Derényi, Illés Farkas, and Tamás Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, June 2005.

[124] Josiane Xavier Parreira, Debora Donato, Sebastian Michel, and Gerhard Weikum. Efficient and decentralized PageRank approximation in a peer-to-peer web search network. In *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB'06)*, Seoul, South Korea, September 2006.

[125] PayPerPost. `http://www.payperpost.com`.

[126] Matti Peltomäki and Mikko Alava. Correlations in bipartite collaboration networks. *Journal of Statistical Mechanics*, P01010, 2006.

[127] Bryan Pfaffenberger. *The USENET Book: Finding, Using, and Surviving Newsgroups on the Internet*. Addison Wesley, New York, NY, USA, 2004.

[128] Arun G. Phadke and James S. Thorp. *Computer relaying for power systems*. John Wiley & Sons, Inc., New York, NY, USA, 1988.

[129] Ithiel Pool and Manfred Kochen. Contacts and influence. *Social Networks*, 1:1–48, 1978.

[130] Alex Pothen, Horst D. Simon, and Kan-Pu Liou. Partitioning sparse matrices with eigenvectors of graphs. *Society for Industrial and Applied Mathematics (SIAM) Journal on Matrix Analysis and Applications*, 11(3):430–452, July 1990.

[131] Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences (PNAS)*, 101(9):2658–2663, March

2004.

[132] Reddit. http://www.reddit.com.

[133] David Rezner. The Power and Politics of Weblogs. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW'04)*, Chicago, IL, November 2004.

[134] Rice Culture. http://www.professor.rice.edu/professor/Rice_Culture.asp?SnID=1654701514.

[135] Rice University Alumni Directory. https://online.alumni.rice.edu/directory/detailsearch.asp.

[136] Rice University Student Directory. http://www.rice.edu/search/query.php?advanced=1&tab=people.

[137] Matthew Richardson and Pedro Domingos. Mining Knowledge-Sharing Sites for Viral Marketing. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02)*, Edmonton, Canada, July 2002.

[138] François-René Rideau. Stamps vs spam: Postage as a method to eliminate unsolicited commercial email. 2002. http://fare.tunes.org/articles/stamps_vs_spam.html.

[139] Ronald L. Rivest, Adi Shamir, and David A. Wagner. Time-lock puzzles and timed-release crypto. Technical report, Cambridge, MA, USA, 1996.

[140] Ryze. `http://www.ryze.com`.

[141] Karthikeyan Sankaralingam, Simha Sethumadhavan, and Jams C. Browne. Distributed PageRank for P2P systems. In *Proceedings of the 12th IEEE International Symposium on High Performance Distributed Computing (HPDC-12)*, Seattle, WA, June 2003.

[142] Jari Saramaki and Kimmo Kaski. Scale-free networks generated by random walkers. *Physica A*, 341:80, 2004.

[143] Jitesh Shetty and Jafar Adibi. The Enron Email Dataset Database Schema and Brief Statistical Report. Technical report, University of Southern California Information Sciences Institute, 2004.

[144] Georgos Siganos, Sudhir L. Tauro, and Michalis Faloutsos. Jellyfish: A Conceptual Model for the AS Internet Topology. *Journal of Communications and Networks*, 8(3):339–350, 2006.

[145] SixDegrees.com. `http://www.sixdegrees.com`.

[146] Barry Smyth, Evelyn Balfe, Oisin Boydell, Keith Bradley, Peter Briggs, Maurice Coyle, and Jill Freyne. A live-user evaluation of collaborative web search. In

Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05), Edinburgh, Scotland, July 2005.

[147] Social Network Marketing: Ad Spending and Usage. `http://www.emarketer.com/Report.aspx?code=emarketer_2000478`.

[148] SpamAssassin. `http://spamassassin.apache.org`.

[149] Stanford WebBase Project. `http://www-diglib.stanford.edu/~testbed/doc2/WebBase`.

[150] Jaime Teevan, Susan T. Dumais, and Eric Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th Annual International ACM SIGIR Conference (SIGIR'05)*, Salvador, Brazil, August 2005.

[151] Nguyen Tran, Bonan Min, Jinyang Li, and Lakshmi Submaranian. Sybil-resilient online content voting. In *Proceedings of the 6th Symposium on Networked Systems Design and Implementation (NSDI'09)*, Boston, MA, April 2009.

[152] Paul F. Tsuchiya. The Landmark Hierarchy: A New Hierarchy for Routing in Very Large Networks. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM'88)*, Stanford, CA, August 1988.

[153] Joshua R. Tyler, Dennis M. Wilkinson, and Bernardo A. Huberman. Email as spectroscopy: Automated discovery of community structure within organizations. In *Proceedings of the First International Conference on Communities and Technologies (ICCT'03)*, Dordrecht, The Netherlands, 2003.

[154] Alexei Vásquez. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physics Review E*, 67, 2003.

[155] Ken Wakita and Toshiyuki Tsurumi. Finding community structure in mega-scale social networks (extended abstract). In *Proceedings of the 16th International Conference on the World Wide Web (WWW'07)*, Banff, Canada, 2007.

[156] Michael Walfish, J.D. Zamfirescu, Hari Balakrishnan, David Karger, and Scott Shenker. Distributed quota enforcement for spam control. In *Proceedings of the 3rd Symposium on Networked Systems Design and Implementation (NSDI'06)*, San Jose, CA, May 2006.

[157] Stanley Wasserman and Katherine Faust. *Social Networks Analysis: Methods and Applications*. Cambridge University Press, Cambridge, UK, 1994.

[158] Duncan J. Watts and Jonah Peretti. Viral Marketing for the Real World. *Harvard Business Review*, May 2007.

[159] Duncan J. Watts and Steven Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[160] Fang Wei, Chen Wang, Li Ma, and Aoying Zhou. Detecting overlapping community structures in networks with global partition and local expansion. In *Proceedings of the Asia-Pacific Web Conference (APWeb'08)*, Shenyang, China, April 2008.

[161] Dennis M. Wilkinson and Bernardo A. Huberman. A method for finding communities of related genes. *Proceedings of the National Academy of Sciences (PNAS)*, 101:5241–5248, April 2004.

[162] Walter Willinger, David Alderson, and Lun Li. A pragmatic approach to dealing with high-variability in network measurements. In *Proceedings of the 2nd ACM/Usenix Internet Measurement Conference (IMC'04)*, Taormina, Italy, October 2004.

[163] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P. N. Puttaswamy, and Ben Y. Zhao. Understanding user interactions in facebook. In *Proceedings of the 4th Conference of the European Professional Society for Systems (EuroSys'09)*, Nuremburg, Germany, April 2009.

[164] YaCy Search Engine. `http://www.yacy.net`.

[165] Yahoo! MyWeb. `http://myweb2.search.yahoo.com`.

[166] Jeff Yan and Ahmad Salah El Ahmad. Usability of captchas or usability issues in captcha design. In *Proceedings of the 4th symposium on Usable Privacy and Security (SOUPS'08)*, Pittsburgh, PA, July 2008.

[167] YouTube. `http://www.youtube.com`.

[168] Haifeng Yu, Phillip B. Gibbons, Michael Kaminsky, and Feng Xiao. SybilLimit: a near-optimal social network defense against Sybil attacks. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy (SP'08)*, Oakland, CA, May 2008.

[169] Haifeng Yu, Michael Kaminsky, Phillip B. Gibbons, and Abraham Flaxman. SybilGuard: Defending against Sybil attacks via social networks. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM'06)*, Pisa, Italy, August 2006.

[170] George Udny Yule. A Mathematical Theory of Evolution, based on the Conclusions of Dr. J.C. Willis, F.R.S. *Philosophical Transactions of the Royal Society*, 213:21–87, 1924.

[171] Wayne W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33:452–473, 1977.

[172] Philip R. Zimmermann. *The Official PGP User's Guide*. MIT Press, Cambridge, MA, USA, 1994.

[173] Vinko Zlatić, Miran Božičević, Hrvoje Štefančić, and Mladen Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physics Review E*, 74, 2006.

[174] Zoomr. http://www.zoomr.com.