UMA ANÁLISE EMPÍRICA DE INTERAÇÕES EM REDES SOCIAIS

FABRÍCIO BENEVENUTO DE SOUZA

UMA ANÁLISE EMPÍRICA DE INTERAÇÕES EM REDES SOCIAIS

Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação.

Orientador: Virgílio Augusto Fernandes Almeida

Belo Horizonte

Março de 2010

FABRÍCIO BENEVENUTO DE SOUZA

AN EMPIRICAL ANALYSIS OF INTERACTIONS IN ONLINE SOCIAL NETWORKS

Thesis presented to the Graduate Program in Ciência da Computação of the Universidade Federal de Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Ciência da Computação.

Advisor: Virgílio Augusto Fernandes Almeida

Belo Horizonte

March 2010

Souza, Fabrício Benevenuto de
S729a An Empirical Analysis of Interactions in Online Social Networks / Fabrício Benevenuto de Souza. — Belo Horizonte, 2010
xxvi, 123 f. : il. ; 29cm

Tese (doutorado) — Universidade Federal de Minas Gerais
Orientador: Virgílio Augusto Fernandes Almeida

1. Online social networks. 2. social interactions. 3. user behavior. 4. clickstream. 5. browsing. 6. silent activity. 7. social media. 8. video response. 9. video spam. I. Título.

CDU 519.6*04

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Uma análise empírica sobre interações em redes sociais

FABRÍCIO BENEVENUTO DE SOUZA

Tese defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. VIRGÍLIO AUGUSTO FERNANDES ALMEIDA - Orientador
Departamento de Ciência da Computação - UFMG

PROFA. MARIA DA GRAÇA CAMPOS PIMENTEL
Departamento de Ciência da Computação - USP

PROFA. SUE MON
KAIST - Coréia

DR. FRANCESCO BONCHI
Yahoo! Research

PROFA. JUSSARA MARQUES DE ALMEIDA
Departamento de Ciência da Computação - UFMG

PROF. MARCOS ANDRÉ GONÇALVES
Departamento de Ciência da Computação - UFMG

PROF. WAGNER MEIRA JÚNIOR
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 12 de março de 2010.

Aos meus pais, José Orlando e Alilás, e ao amor da minha vida, Carol.

"Se de tudo fica um pouco"
mas por que não ficaria um pouco de mim?"
Carlos Drummond de Andrade - "Resíduo"

Acknowledgments

Primeiro e mais importante, eu gostaria de agradecer ao meu orientador, Virgílio Almeida, por tudo que ele me ofereceu durante toda minha carreira acadêmica, que inclui graduação, mestrado, e doutorado. Sem ele nada desse trabalho seria possível. Eu sou profundamente grato a ele por me ensinar a conduzir pesquisa de qualidade, por me permitir co-orientar alunos, por financiar minha pesquisa e viagens e por convidar colaboradores que tanto acrescentaram a esse trabalho. Eu tenho certeza que essa gratidão que tenho a ele só aumentará com o tempo, dado que ele continua me ajudando e aconselhando em diversas situações.

Eu gostaria de agradecer a todos os professores do DCC/UFMG, em particular, a Marcos Gonçalves e Jussara Almeida. Eles colaboraram em vários dos artigos que compõem esta tese e, sem a ajuda deles, este trabalho não seria o mesmo. Gostaria ainda de agradecer a Wagner Meira Jr., Adriano Veloso e Adriano Pereira pelas inúmeras conversas, colaborações e idéias que trocamos. Eu sou muito grato aos funcionários do DCC, por serem tão legais e prestativos comigo.

Eu gostaria de agradecer a todos que fizeram ou fazem parte do e-speed, Camps e VoD. Eu me sinto privilegiado por ter convivido com as várias pessoas que passaram por esse laboratórios tais como Lhg, Barra, Fernando, Fabiano, Marisa, Tatiana, Gabriel, Rauber, Tiago, Cesar, Matheus, Djim, Emanuel, Ítalo, Krusty, e o pessoal da grad001. Eu sou muito grato ao Fernando, por ter sido fundamental na mudança de tema de tese para redes sociais e ao Tiago, que me ajudou em vários artigos dessa tese.

Eu gostaria de agradecer ao Krishna Gummadi e à Meeyoung Cha (Mia) pelo internship no MPI-SWS. Essa experiência foi fantástica e me permitiu conhecer e colaborar com importantes pesquisadores na área de redes sociais. Gostaria também de agradecer a todos que conheci no MPI-SWS por fazerem de Saabruecken um lugar especial para mim.

Eu gostaria de expressar minha imensa gratidão à minha família, especialmente aos meus pais e à minha irmã. Eles são meu porto seguro nos altos e baixos que um doutorado oferece. Obrigado a todos da família da Carol, às minhas tias, tios, primos e

à minha avó por todo apoio. Obrigado também aos meus grandes amigos Ti, Chiquinho e Daniel. Eles não ajudaram em nada tecnicamente, mas os inúmeros churrascos, bares e celaborações que participamos juntos funcionaram como necessárias interrupções em minha pesquisa. Gostaria ainda de agradecer à Ana, à Lívia, ao Ednei, ao Tavinho e à Débora pelo mesmo motivo.

Finalmente, eu não tenho palavras para expressar tudo o que minha esposa fez por mim. Esta tese não é meu triunfo, Carol. É **nosso**.

List of Abbreviations and Acronyms

BFS - Breadth First Search

CCDF - Complementary Cumulative Distribution Function

CDF - Cumulative Distribution Function

CV - Coefficient of Variation

DFS - Depth First Search

FoF - Friend of a Friend

HP - Heavy Promoters

ISP - Internet Service Provider

LP - Light Promoters

OSN - Online Social Network

PDF - Probability Distribution Function

RBF - Radial Basis Function

SCC - Strongly Connected Component

SVM - Support Vector Machines

UGC - User Generated Content

WCC - Weakly Connected Component

Abstract

Online social networks (OSNs) have become extremely popular websites and part of our daily lives. OSNs represent a new kind of information network that allow users to interact online. Every day, huge amounts of content are shared and millions of users converse through online social links. Understanding what activities users do, what content they share, and how they interact when they connect to these sites create not only opportunities for better interface and system design, but is also important for many applications, such as advertising, political campaigning, and detection of opportunistic and malicious behavior. Despite all these potential benefits, little is known in the research community about it.

In this thesis we provide an in-depth study of user interactions in OSNs, covering aspects of user behavior and navigation across social features as well as aspects of unsolicited content exchanged on social interactions. To do that, we gathered data from actual OSN sites, including YouTube and Orkut. We then study how users interact across a number of OSN features, providing a global picture of the range, duration, frequency, and sequence of activities that users do when they connect to these sites. Second, we provide a characterization of the interactions that emerge from YouTube's video responses, a feature that allows users to interact primarily using videos rather than text. Our study unveils typical user behavioral patterns and identifies novel forms of unsolicited content in OSNs. Some users post video response spam (i.e. video responses unrelated to the topic discussed) in order to increase the popularity of a video, or spread advertisements and pornography. We propose a machine learning-based method that is able to accurately identify the majority of these users.

keywords: online social networks, social interactions, user behavior, clickstream, browsing, silent activity, social media, video response, video spam.

Resumo

Redes sociais se tornaram extremamente populares e parte do nosso dia a dia, permitindo usuários interagir de diferentes formas. A cada dia, grandes quantidades de conteúdo são compartilhadas e milhões de usuários conversam através de elos sociais. Entender quais atividades os usuários fazem, que tipo de conteúdo eles compartilham, e como eles interagem quando se conectam a esses sítios pode criar oportunidades para melhorar a interface e o funcionamento dos sistemas existentes e é importante para diversas aplicações relacionadas a propagandas, campanhas políticas e detecção de comportamento malicioso. Apesar de todos esses benefícios, pouco se sabe na comunidade de pesquisa sobre isso.

Essa tese provê um amplo estudo sobre interações em redes sociais, cobrindo aspectos do comportamento e da navegação dos usuários, bem como aspectos relacionados à postagem de conteúdo não solicitado. Para isso, coletamos dados de redes sociais atuais, incluindo o YouTube e o Orkut. Em seguida, estudamos como usuários interagem através de várias ferramentas de redes sociais, provendo uma visão geral da duração, freqüência e seqüência das atividades dos usuários nesses sítios. Feito isso, focamos em caracterizar interações que surgem da ferramenta de vídeo resposta do YouTube, um recurso que permite aos usuários interagir utilizando vídeos ao invés de texto. Nosso estudo revela padrões de comportamento típico e identifica novas formas de conteúdo não solicitado. Alguns usuários postam vídeos resposta contendo spam (vídeos resposta não relacionados ao tópico discutido) na tentativa de aumentar a popularidade de um vídeo, ou espalhar propagandas ou pornografia. Utilizando um método de aprendizagem de máquina conseguimos identificar a maior parte desses usuários.

Palavras-chave: redes sociais, interações sociais, comportamento dos usuários, atividade silenciosa, mídia social, vídeo resposta, vídeo spam.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Since its beginning, the Internet has been a place for a series of new applications including the World Wide Web, Peer-to-Peer applications, and email. Nowadays, the Web is experiencing a new wave of applications associated with the proliferation of social networks and the growth of digital media. Numerous online social networks (OSNs) have emerged, including networks of professionals (e.g., LinkedIn), networks of friends (e.g., MySpace, Facebook, Orkut), and networks for sharing specific types of content such as short messages (e.g., Twitter), diaries and journals (e.g., LiveJournal), photos (e.g., Flickr), and videos (e.g., YouTube).

OSN sites have attracted millions of users. According to Nielsen Online [Report], social media has pulled ahead of email as the most popular online activity. More than two-thirds of the global online population visit and participate in social networks and blogs. As a comparison, if Facebook was a country, its 500 million registered users would rank it as the third largest country of the world in terms of population [3]. According to Alexa.com, Facebook, MySpace, and YouTube are among the top-10 most visited websites of the world, both in terms of users and amount of time spent on the sites.

A common functionality among the OSN sites is to allow users to share content, which may vary from simple text messages communicating the status of the user to multimedia-rich files such as photos and home videos. As a consequence, the statistics about these user generated content sites are also impressive. Facebook contains more than 60 billion photos that occupy more than 1.5 PB of storage space [4]. It has been reported that the amount of content uploaded to YouTube in 60 days is equivalent to the content that would have been broadcast for 60 years, without interruption, by NBC, CBS and ABC altogether [nyt]. Moreover, Indeed, YouTube has reportedly served over 100 million users only on January 2009 [com], with a video upload rate equivalent to

10 hours per minute [8].

The extreme popularity and rapid growth of OSN sites are related to a new perspective that these sites bring to Web users. OSNs not only allow users to create and retrieve content, but also allow them to interact through the created content, working as a communication medium where users can deliver information, spread news, express feelings and emotions, get to know about other individuals, or simply keep in touch with friends.

Despite the increased interest of the research community in studying OSNs, it has been difficult to gain insight into the usage patterns of OSNs and the characteristics of the interactions in these sites. A few recent studies focused on characterizing specific features and interactions through textual messages [51; 80; 132; 138] and third-party applications [69; 109]. While these studies yield insights into how a social network created through textual interactions is structurally different from the social network created by friendship links, little attention has been paid to other kinds of interactions that can emerge from the OSN features (e.g., interactions through multimedia content such as photos or videos). Moreover, these studies have reconstructed user actions from "visible" textual interactions, neglecting "silent" user actions like browsing a profile page or viewing a photo.

In this thesis we aim at filling this gap, providing an in-depth study of user interactions in OSNs, covering aspects of user behavior and navigation across different social features. Next, in Section 1.1, we discuss in details some of the opportunities that have motivated our work. Section 1.2 presents the thesis statement, highlighting the goals and steps taken to conduct this work. Finally, Section 1.3 discusses how the rest of the thesis is organized.

1.1 Motivation

OSNs are still in their infancy, experiencing a number of new emerging trends and features. As this paradigm matures, we expect that social systems will eventually become portals for both personal and commercial online interactions. Below, we outline a few of the many potential applications that could benefit from understanding the current aspects of user activity and interactions in OSNs.

- **Comercial**: With users spending most of their Internet time in OSNs, these sites have become a suitable place for advertise placement. In fact, in 2007, $1.2 billion was spent on advertisement in OSNs worldwide, and this is expected to triple by 2011 [B. Williamson]. Moreover, OSNs are places where users share

and receive information, influencing and being influenced by friends [Cha et al.]. Consequently, OSNs are increasingly becoming important for political campaigning [63] as well as any sort of viral marketing, where users can be encouraged to share a product advertisement with their friends [95].

Consequently, understanding how users interact through social links, what are the features used to interact and how content is diffused is increasingly important in commercial advertising, in political campaigning, and ultimately to society. For instance, viral marketers might want to exploit patterns of user interaction to spread their content or promotions quickly and widely [95].

- **Sociological**: In a recent past, social network analysis was a domain of sociologists and anthropologists, when typical tools to collect social network data were surveys and interviews [133]. Online social networks represent an opportunity to understand social aspects of individuals and characteristics of user behavior with much larger datasets. Furthermore, OSNs are bringing new forms of communication to the society and changing aspects of our daily's lives. OSNs help people to interact more, allow individuals to get to know about others, and allow individuals to express themselves and be heard by a local or global audience. Thus, understanding the dynamics of OSN interactions not only can unveil characteristics of human nature and social behavior but it can also help us to understand evolutionary aspects of communication and its consequences to society.

- **System enhancements**: OSN sites are vulnerable to new trends and, as several other Web systems, they are susceptible to see their users quickly move to another system, with no advance warning. For instance, MySpace experienced an exponential growth on the number of users followed by a slowdown after April 2008 due to an increase in the popularity of Facebook [129]. In its beginning, Orkut was growing fast in several countries, but such popularity was concretized only in a few countries, out of which Brazil is the most popular [5].

  Among a number of reasons to explain these phenomena, we can point out the design of website interface, performance issues, cultural and social characteristics of users, etc. A characterization of the use of these systems and their features could be used in the design and evaluation of current systems and could also shed light towards better designs of future OSN systems.

  Another important aspect related to system enhancements has to do with traffic generated by OSNs sites. Intuitively, there is a crucial difference between publishing content on the traditional web and sharing content over OSNs. When people

publish content on the Web, they typically do so to make the content accessible to Internet users everywhere. On the other hand, when users publish content on OSNs they often have an intended audience; usually, their friends. Sometimes the audience is explicitly defined by the user or the site's policy. Thus, OSNs represent a unique class of applications with potential to reshape Internet traffic with its increasing growth. Understanding how users use OSNs to interact can be valuable in designing the next-generation Internet infrastructure and content distribution systems [90; Rodriguez].

- **Unwanted content**: OSNs are increasingly becoming a target opportunistic users that send unsolicited marketing, propaganda, or other disruptions of legitimate content. The problem manifests itself in different forms, such as posts on top lists containing spam, inappropriate labeled content on sharing sites such as YouTube, and unwanted invitations.

  Unsolicited content wastes human attention, which is one of the most valuable resources in the information era. The noise and annoyance created by some users reduces the effectiveness of online communication media. Understanding properties of online interactions and characterizing typical user behavior can be useful to reduce the problem.

1.2 Thesis statement

Despite all the potential benefits discussed in Section 1.1, little is known about interactions in OSN sites. How do users behave and what are they focusing on when they connect to these sites? What are the features they use to interact and what are the properties of these interactions? To what extent do they interact with others, compared to creating and consuming content? What fraction of all friends do users interact with in the social graph? Do users interact with non-friends?

Aiming at answering these questions, this thesis has a two-fold goal. First we aim at understanding the properties of user navigation and interactions in OSNs as well as exploring interactions that emerge from non-textual features. Second, we want to exploit characteristics of users and their interactions to identify opportunistic users and propose techniques able to reduce unwanted interactions in OSNs.

More specifically, we take the following steps.

- Initially, we aim at providing a global picture of the types, duration, and frequency of activities that users do when they connect to OSNs. We aim at

understanding how users navigate from one activity to another and studying interactions that emerge from these activities.

- Our second goal consists of providing an in-depth understanding of a unique type of interaction that emerges from multimedia content. Most part of the studies on social interactions relies on textual interactions [51; 80; 132; 138] or interactions through third part applications [69; 109]. Here we want to explore a new form of interaction that uses a much richer content than text. We choose to characterize video-based interactions that emerge from YouTube's video responses, a feature that allows users to discuss themes and to provide reviews for products or places using videos rather than text. We aim at uncovering typical user behavioral patterns in video-based environments as well as aspects of video-based interactions.

- Finally, we go a step further, leveraging our understanding of video-based interactions to identify unsolicited content in the YouTube video response feature. Video responses can be a *video response spam* when they are not related to the topic being discussed. Users can create this form of opportunistic interaction for several reasons, including increasing the popularity of a video, marketing advertisements, distributing pornography, or simply polluting the system. Our ultimate goal is to provide mechanisms to detect the users that post video response spam.

1.3 Thesis organization

The remaining chapters of this thesis are organized as follows.

- **Chapter 2** presents definitions, describes common OSN features and mechanisms of interaction. Moreover, this chapter also presents a brief background on social network theory, discussing metrics and properties of social systems.

- **Chapter 3** presents a literature review about interactions in OSNs, also covering aspects of opportunistic behavior.

- **Chapter 4** presents the measurement methodology used in this thesis. We detailed how we obtained the datasets used in our empirical studies, describing not only the procedures used to collect these datasets, but also providing characteristics of the data and limitations of our methodology.

- **Chapter 5** presents an in-depth study about user navigation and interaction in OSNs, providing a global picture of the range, duration, and frequency of activities that users do when they connect in OSNs.
- **Chapter 6** characterizes video-based interactions that emerge from the video response feature provided by YouTube, unveiling forms of unwanted communication in these interactions.
- **Chapter 7** builds on Chapter 6, proposing a mechanism to prevent unwanted communication in video-based interactions that comes from the YouTube video response feature.
- **Chapter 8** concludes the thesis, discussing our contribution and offering directions for future research.

Part of this work has appeared in a number of conferences and journals. Particularly, references [21; 22; 23; 24; 25; 26; 27; 28; 29; 30; 56; 120; 121] contains pieces of work that are partially or fully presented in this thesis.

Chapter 2

Basics of online social networks

In this chapter, we provide an overview of online social networks, describing their characteristics and the types of user interactions they allow. First, Section 2.1 presents basic definitions of online social networks. Then, we present a description of common features available in OSNs in Section 2.2. Section 2.3 presents the definition of OSN interactions and some types of interactions that emerge from OSN features. Finally, Section 2.4 provides a brief background on social network theory and presents common metrics used to describe properties of complex systems.

2.1 Definition

The term online social network is usually used to describe any group of people that primarily interact via communication media. Thus, based on this definition, OSNs can be considered to exist since the beginning of the Internet. However, in this work we use a more restricted definition, also adopted in previous studies [34; 103]. We define an online social network as a web-based service that allows individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system.

Based on this definition, there is a number of OSN systems available on the Web, which vary according to their purposes. For instance, there are OSNs of professionals (e.g., LinkedIn), friends (e.g., MySpace, Facebook, Orkut), short messages (e.g., Twitter), diaries and blogs (e.g., LiveJournal), bookmarks (e.g., Digg and Delicious), photos (e.g., Flickr), and videos (e.g., YouTube and Metacafe).

Table 2.1 summarizes the purposes of some popular OSN sites. An updated and exhaustive list of existent OSNs, which accounts with more than 150 different sites,

Name	Purpose	URL
Orkut	Friendship environment	http://www.orkut.com
Facebook	Friendship environment	http://www.facebook.com
MySpace	Friendship environment	http://www.myspace.com
Hi5	Friendship environment	http://www.hi5.com
LinkedIn	Professionals	http://www.linkedin.com
YouTube	Video sharing system	http://www.youtube.com
Flickr	Photo sharing system	http://www.flickr.com
LiveJournal	Blog, journals, and diaries	http://www.livejournal.com
Digg	Recomendation and bookmarking sharing	http://digg.com
Twitter	Short messages and user status	http://twitter.com

Table 2.1: Some popular online social networks

can be found in [wik].

2.2 Common features of OSNs

The existing types of user interactions that OSNs promote are directly related to some of the features provided by these websites. Next, we discuss some of the most important OSN features. Our purpose here is not to provide a complete and exhaustive list of features, but to describe the most relevant ones.

- **User profiles**: OSNs have all functionalities around a user profile, a form of individual (or, less frequently, group of individuals) home page, which offers a description of a member. Profiles can be used not only to identify the individual as a user in the system, but also to identify people with common interests and articulate new relationships. Typically, profiles contain demographic details (age, sex, location, etc.), interests (hobbies, favorite bands, etc.), and a photo. In addition to text, images, and other objects created by the user, the social network site profile also contains messages from other members and lists of the people identified as friends within the network. Profiles are usually publicly accessible to anyone with an account in the OSN or they can be private, according to user defined privacy policies.

  Recently, Boyd *et al.* [33] showed that for most part of OSN users, there is a strong relationship between the real individual identity and her user profile.

- **User updates**: User updates are an effective way of helping users discover content. In order to encourage information sharing, OSNs make user's updates immediately visible to her friends in the social network. User updates are known

to motivate contribution by new users to the system. Burke *et al.* [39] recently conducted a study using the logs of 140,000 new users on Facebook to determine that activity streams (Called 'News Feed' in Facebook terminology) can be vital for motivating new users to contribute to the site. Since updates can be followed by commentaries from other users, they are a special form of communication among users in OSNs.

- **Comments**: Most social network sites let users to make comments about the content shared. Some sites also provide the ability of commenting on user profiles. Comments are the primary conversational medium on OSN sites, which also express social relationships [15; 51]. As example, videos can receive comments on YouTube, photos can be commented on Facebook, Flickr, and Orkut, LiveJournal users can post comments on blogs, etc.

- **Ratings**: OSN content is usually evaluated by users through a mechanism called ratings. Ratings appear in OSNs with different levels of granularity and forms. In Facebook, users are only allowed to click in an "I like this" link. With more than 400 million users registered, Facebook users evaluate 9 pieces of content each month in average [3]. On YouTube, videos can be evaluated with ratings that vary from one to five stars, similarly to the evaluation commonly employed to categorize hotels. YouTube also provides a binary evaluation (positive or negative) for comments on videos as an attempt to filter comments that other users dislike.

  Content rating is useful in many forms. It is important in systems such as YouTube to help users to identify relevant content. These ratings can also help system administrators to identify poor quality or even inappropriate content, which they can inspect and possibly remove. Additionally, ratings can be used as input for a number of other features on the website, such as featured content, top-lists, recommendation systems, etc.

  An OSN site that places user ratings at the center of the system is Digg. It allows users to rate submitted links and stories and use these votes to expose only the most popular content [94].

- **Favorite lists**: Several social applications utilize favorite lists to allow users to select and organize content. Favorite lists help users to manage their own content and can also be useful for social recommendations. For instance, users can keep lists of favorite videos on YouTube and favorite photos on Flickr. Users can browse someone else's favorite list in order to search for new content [44].

Thus, such favorites also work as a means for content discovery and information propagation. Systems such as Orkut and Twitter also provide list of favorite users (fans).

- **Top lists**: Typically, OSNs that place media content at the center of the system, such as YouTube, provide lists of top content or users. Usually these lists rely on evaluations and other statistics about the content (e.g., number of views, ratings, number of comments) or about the users (e.g., number of subscribers).

- **Content metadata**: One of the new trends of the Web 2.0 is to allow users to freely create and associate metadata to content [52]. In OSN sites like YouTube and Flickr, users typically annotate metadata to shared content with a title, a description, and tags. Metadata information is usually the key element for content retrieval approaches in OSNs.

- **Applications**: In a successful attempt to enhance user experience in the OSN site, Facebook made a key innovation: they opened their platform to third-party developers [2]. With this innovation, developers are able to create applications that promotes interactions among users. With the success of Facebook applications, other OSN sites such as Orkut and MySpace also adopted this strategy. The type and number of applications became unbounded with this new open-API model deployed. Facebook alone has over 81,000 third-party applications [dev]. Companies like Zynga, specialized in development of these applications, accounts with more than 60 million Facebook users registered in their applications [dev]. For a large list of Facebook applications we refer the reader to reference [1].

2.3 Types of user interaction

We define interactions in OSNs as any action involving two or more users that occurs via the medium of an OSN site. Interactions are usually used to strengthen and maintain current relationships or even forge new ones. As opposed to communication, an interaction is also related to "something done", and not to only "something said". Thus, actions such as watch videos or browse pictures can be interpreted as interactions between the involved users. Next, we define and present examples of common forms of interactions that occur in OSN sites.

- **Textual interactions** are the interactions that occur primarily through text. They can appear as a form of one-to-one communication via private messages,

or scrapbook messages. Textual interactions can also appear as one-to-many, such as in a Forum. Textual interactions can also occur based on the concept of comments around some object, such as comments in a blog, comments of a video, comments on a photo, etc. Another interesting emerging form of textual interaction is through short messages containing the individual status or her personal take on news and events. Short messages became popular with Twitter and were also adopted by OSN sites such as Facebook and Orkut.

- **Photo interactions**: This form of interaction occurs primarily through photo objects. For instance, a photo interaction occurs when a user is tagged (or annotated) in a friend's photo. Systems that allow users to be tagged, such as Facebook and Orkut, dedicate a special section in a user's photo album to the photos the user was tagged. Additionally, according to the user settings, users can receive a notification by email or in their updates that they were tagged, working as a mechanism in which users can interact with each other.

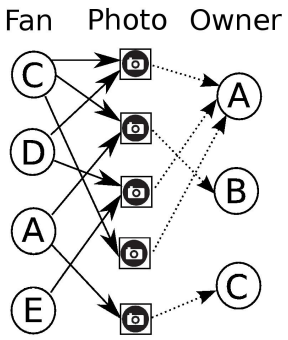


Figure 2.1: Example of a graph formed by photo interactions in Flickr

Another interesting form of photo interaction was recently presented by Valafar *et al.* [131]. They define a photo interaction in Flickr as an interaction that occurs when a user become fan (i.e. favorite mark) of a photo. The photo interaction occurs between the fan and the owner of the photo. As example, Figure 2.1 depicts a representation of fan-owner interactions. An edge from fan C to photo p indicates that p is one of C's favorite photos and thus represents an indirect interaction between fan C and the owner of photo p.

- **Video interactions**: The video interactions are interactions that occur through video objects. OSN sites are not only allowing users to share videos, but they are also offering video-based functionalities as alternative to text-based ones [125].

As example, the YouTube video response feature allows users to interact and converse through video, by creating a video sequence that begins with an opening video followed by video responses from other users. Video responses allow users to provide review for products or places, can be used as a tool for political debates, and allow users to exchange opinions about certain themes (such as in a video forum) using a much richer media than simple text. YouTube provides a number of mechanisms to ease video interactions, allowing users to post video responses by directly uploading it from the user's webcam, choosing a video from her pre-existing YouTube contributions, or uploading a video from the user's disk drive.

Figure 2.2: Illustrative snapshot of the YouTube video response feature

In order to illustrate one of the uses of this tool in YouTube, Figure 2.2 presents a snapshot of a video posted by a user with questions to candidates of the U.S. presidential election in 2008. Some YouTube videos were chosen to be answered lively by the candidates in a debate promoted by CNN. Those questions could be responded with video responses on YouTube by any user [you].

- **Application interactions**: Third-party applications offer an unlimited number of forms to users to interact, which can vary according to the application. As example, Facebook promotes interactions involving applications by notifying its members about the applications their friends have installed and used. Some applications themselves prompt the user to invite friends to install them, provoking direct messages between users. Other applications consist of games in which the

users directly play with each other. We call all forms of interactions related to applications by application interactions.

- **Silent interactions**: Unlike existing social interactions that expect some form of reciprocation, OSN users can read daily chatters of their friends without having to participate in a conversation. As example, consider two fictitious OSN users, Carol and Paul. Carol can freely browse the profile and the pictures of her friend Paul, without leaving any message. Although these users did not directly communicate, Carol got to know that Paul is now married through a new status in Paul's profile information and also through Paul's wedding pictures available in the OSN site. We can think of such interactions as being "silent", as opposed to interactions like leaving a comment to a friend that can be read by at least one other person.

- **Opportunistic interactions**: The wide range of features and new forms interactions available in OSN sites open the doors to new forms of opportunistic user actions. We call opportunistic interaction any interaction that comes as an unsolicited or unwanted interaction. We can list a number of scenarios in which the problem manifests, including unwanted invitations in OSNs and the disruption of legitimate communication with unsolicited marketing or propaganda posted as comments to forums, blogs, and videos.

2.4 Complex network theory

The study of complex networks spawns a number of areas and its theory has been used as a tool for understanding a number of phenomena, including disease epidemics [107], information cascading [135], Web search [38], and consequences of computer network attacks [13]. There are several statistical properties that are used to analyze and classify networks. We discuss these metrics next, in Section 2.4.1. Sections 2.4.2 and 2.4.3 discuss the properties of small-world and power-law networks. For a detailed review of complex network theory and metrics we refer the reader to reference [113].

2.4.1 Network metrics

In this section we present the main metrics used to characterize networks.

2.4.1.1 Node degree

A key characteristic of the structure of a network is its degree distribution. Several networks have degree distributions that follow a power law. Thus, a common metric used to compare these networks is the scaling exponent used to fit the power law distribution. Typical scaling exponents for degree distribution of social and communication networks lie in the range of 1.0 and 3.5 [59]. For direct networks, it is common to analyze both in-degree and out-degree distributions. Additionally, a typical network metric is the average degree (in-degree and out-degree for direct graphs).

2.4.1.2 Clustering coefficient

The clustering coefficient of node i, $cc(i)$, is the ratio of the number of existing edges between i's neighbors over the number of all possible edges between i's neighbors. Thus, it works as a measure of the density of established links between the neighbors of a node. The clustering coefficient of a network, CC, is the mean clustering coefficient of all nodes.

2.4.1.3 Components

A component in a graph is a set of nodes where each node in the set has a path to every other node in the set. We say that a component is a strongly connected component (SCC) when the path between the nodes in the set is direct. On the other hand, we say the component is a weakly connect component (WCC) if the path is undirected.

Broder *et al.* studies the Web link structure [38] and proposes a model to represent the Web component known as *bow tie* structure, since the components organization of this model remembers a bow-tie shape. Such model consists of the largest SCC and other groups of nodes that can either reach the largest SCC or can be reached from the largest SCC. This model has been used to compare the components organization for direct graphs [144].

2.4.1.4 Average distance and diameter

The average distance of a graph is the average number of steps along the shortest paths for all possible pairs of network nodes. It is usually computed on the largest WCC (undirected graphs) or on the SCC (direct graphs), since there is no path between nodes located on different components. Another common metric of a graph is its diameter. The diameter of a graph is defined as the distance of the longest shortest path on the graph (also usually computed considering only the SCC or the WCC).

2.4.1.5 Assortative mixing

A network is said to exhibit assortative mixing if the nodes with many connections tend to be connected to other nodes with many connections. Social networks usually show assortative mixing. We define as knn(k) as the average degree of all neighbors of nodes with degree k. The assortativity of a network is usually evaluated by plotting knn(k) as a function of k.

Additionally, for direct graphs, the assortative (or dissortative) mixing can be evaluated by the Pearson coefficient r which is calculated as follows [112]:

$$r = \frac{\sum_i j_i k_i - M^{-1} \sum_i j_i \sum_{i'} k_{i'}}{\sqrt{[\sum_i j_i^2 - M^{-1}(\sum_i j_i)^2][\sum_i k_i^2 - M^{-1}(\sum_i k_i)^2]}}, \tag{2.1}$$

where j_i and k_i are the excess in-degree and out-degree of the nodes that the ith edge leads into and out of, respectively, and M is the total number of edges in the graph.

2.4.1.6 Betweenness centrality

Network centrality can be understood as a measure of the importance of an edge (in terms of location) in a graph. Some measures of centrality are widely used in network analysis: degree centrality, betweenness, closeness, and eigenvector centrality. Particularly, the betweenness centrality B of an edge is defined as the number of shortest paths between all pairs of nodes in the graph that cross the edge [115]. If a pair of nodes has 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

$$B(e) = \sum_{u \in V, v \in V} \frac{\sigma_e(u, v)}{\sigma(u, v)} \tag{2.2}$$

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 indicates 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.

Similarly, the betweenness centrality can be computed for a node rather than an edge. In this case, betweenness measures the number of shortest paths that passes through a node. Nodes that occur on many shortest paths between other nodes have higher betweenness than those that do not.

2.4.1.7 Reciprocity

An interesting metric to observe for direct graphs is the reciprocity of each node. The reciprocity measures the probability of a node having an incoming edge from each node it points. In other words, the reciprocity ($R(x)$) is given by:

$$R(x) = \frac{|O(x) \cap I(x)|}{|O(x)|} \tag{2.3}$$

where $O(x)$ is the set of nodes that receive an edge from user x and $I(x)$ is the set of nodes that reach x through direct edges.

Another interesting metric to observe is the link reciprocity coefficient ρ, a metric which captures the reciprocity of the interactions in the entire network [64]. The reciprocity coefficient ρ is defined as the correlation coefficient between the entries of the adjacency matrix of a directed graph ($a_{ij} = 1$ if a link from i to j is there, and $a_{ij} = 0$ if not):

$$\rho = \frac{\sum_{i \neq j}(a_{ij} - \bar{a})(a_{ji} - \bar{a})}{\sum_{i \neq j}(a_{ij} - \bar{a})^2}, \tag{2.4}$$

where the average value $\bar{a} = \sum_{i \neq j}(a_{ij}/N(N-1)$ and N is the number of users in the graph.

The reciprocity coefficient tells whether the number of mutual links in the network is more or less than that of a random network. If the value of ρ is higher than 0, the network is reciprocal; otherwise, anti-reciprocal.

2.4.1.8 PageRank

PageRank is an interative algorithm that assigns a numerical weight to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. The PageRank algorithm was initially proposed by Brin and Page [37] to rank Web pages as a result of a search on the prototype of Google.

The intuition behind PageRank is that a Web page is important if it has many incoming links or if the page has links coming from highly ranked pages. Thus, the importance of a certain Web page influences and it is influenced by the importance of

other pages. The PageRank is defined as:

$$PR(u) = (1 - d) + d \sum_{v \in B(u)} \frac{PR(v)}{N_v} \tag{2.5}$$

where u represents a web page. $B(u)$ is the set of pages that points to u. $PR(u)$ and $PR(v)$ are rank scores for pages u and v, respectively. N_v denotes the number of outgoing links of page v, and the parameter d is damping factor which can be set between 0 and 1. The PageRank algorithm has been applied to other contexts, as example, to find expertise users in forums [144].

2.4.2 Small-World networks

The concept of small-world became well known with the famous Milgram's experiment [101]. His experiment consisted of a group of volunteers trying to forward a letter to a target person through people they knew. Basically, Milgran sent letters to several people. These letters explained that he was trying to reach a final specific individual in a US city and that the recipient should forward the letter to whomever they thought could bring the letter closer to the final individual (or deliver to the individual himself, if the recipient knew the person). Before forwarding the letter, however, the recipient was asked to add his or her name to the end of the letter, so that Milgram could trace the path to the destination. From the letters that successful reached the final destination, the average number of hops required to the target was six, a result known as the *six degrees of separation* principle.

In terms of the social network properties discussed previously, a graph can be considered to be small-world if it has two basic properties: high clustering coefficient and small network diameter [134]. These properties were verified in a number of different networks such as the Web [14; 38] (Web sites are nodes and Web links between them are edges), scientific collaboration on research papers [111; 114] (researches are nodes and edges connect co-authors of papers), film actors [17] (actors are nodes and edges connect actors that participate on the same movie), and online social networks [10; 105; 54; 97]. Particularly, Mislove *et al.* [105] verified the small-world properties in four online social networks: LiveJournal, Flickr, Orkut, and YouTube.

2.4.3 Power law networks

Power-law networks consist of graphs whose degree distribution follows a power law. In other words, in such networks 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. A number of real world networks showed to have power law degree distributions, including the Internet topology [60], the Web [20], and neural networks [35].

Scale-free networks are a class of power law networks where the high-degree nodes tend to be connected to other high-degree nodes. Barabási *et al.* [20] proposed a model to generate scale-free networks, introducing the concept of *preferential attachment*. In short, preferential attachment says that the likelihood of a node being attached to a new link is in proportional to the node's degree. Barabási *et al.* [20] showed that, under certain circumstances, preferential attachment results in power law networks. More recently, Li *et al.* [98] propose a metric to measure the scale-freeness of graphs and also provides strong definitions about scale-free networks as well as provides an extensive discussion about the theme.

2.5 Discussion

This chapter provides background and definitions used along this thesis. We can note that the wide range of new features and applications that OSN sites provide allow users to interact in multiple forms. In the next chapter we survey some efforts that explore some of these types of interactions.

Chapter 3

Literature Review

In this chapter, we survey related work. First, Section 3.1 discusses recent studies related to structural properties of online social networks. Section 3.2 discusses related work that aims at studying usage and user behavior in OSNs. Then, Section 3.3 surveys efforts towards characterizing and understanding interactions in OSNs. We present a survey of measurement work on video sharing systems in Section 3.4. Finally, Section 3.5 presents efforts and approaches to reduce opportunistic interactions in online systems.

3.1 Structural properties of OSNs

Adamic *et al.* [10] studied an online social network, called Club Nexus, formed by students at Stanford University. The network analyzed exhibits power-law and small-world behavior. Mislove *et al.* [105] studied graph theoretic properties of OSNs, based on the friends network of Orkut, Flickr, YouTube, and LiveJournal. They confirmed the power-law, small-world and scale-free properties of OSN services. Complementarily, Ahn *et al.* [12] studies the network properties of Cyworld, a popular OSN in South Korea. They showed that Cyworld has power-law scaling in degree distribution, large clustering coefficients, and verified the small-world phenomenon. The authors in [54] investigate social network properties considering the network formed by related videos, finding that the links to related videos have small-world characteristics.

Kumar *et al.* [92] approached two OSNs: Flickr and Yahoo! 360°. They examined path properties (such as diameter), graph density, change over time, and presence of giant component. As the main results, they find that both networks have a large SCC. Additionally, by keeping track of the moments when each node and edge arrives in the graph, they noted in both networks the same evolutionary pattern: rapid growth on

the beginning, and then slow but steady growth. Finally, Leskovec *et al.* analyzed the network formed by instant messages in Microsoft Messenger [97]. They found that the network is highly connected (99.9% of the nodes belonging to the largest WCC). Additionally, they verified small-world properties and also verified a strong correlation between instant message communication and user geographic location.

In general, the above studies confirm that online social networks obey power-law scaling characteristics, verify the presence of a dominant large weakly connected component in OSN and firmly establishes them as small-world networks [17].

3.2 OSN usage

The use of OSN sites has been studied in different contexts. Through interviews with Facebook users, Joinson *et al.* [86] identified various reasons for users to use Facebook including, social connection, shared interests, share and retrieve social content, social network surfing, and status updating. Burke *et al.* [39] studied user motivations for contributing in OSN sites based on server log data from Facebook. They found that newcomers who see their friends contributing also share more content themselves. Chapman and Lahav [45] conducted survey interviews and analysis of Web browsing patterns of 36 users of four different nationalities to examine ethnographical differences in the usage of OSNs. Their analyses suggest the existence of three dimensions of cultural difference for typical social networking behaviors: the users' goals, typical pattern of self expression, and common interaction behaviors.

Caverlee *et al.* [41] studied the characteristics of large OSNs analyzing over 1.9 million MySpace profiles in an effort to understand who is using these networks and how they are being used. Among their findings, they show that young users (in their teens and 20s) are most prevalent on MySpace and that half of the user profiles have been abandoned. They also found patterns of language use for users based on their age, location, and gender. Finally, Schneider *et al.* [123] analyzed OSN clickstream data extracted from network traffic and identifying typical user navigation patterns in OSNs.

3.3 Interactions in online social networks

There has been a number of efforts that study properties of user interactions in OSNs. A network formed by textual interactions in Cyworld was recently approached in [51]. They compare the explicit friend relationship network with the implicit network created

by messages exchanged on Cyworld's guestbook, finding several similarities in terms of network structure. Particularly, they found that the in-degree and out-degree distributions are close to each other and social interactions through the guestbook are highly reciprocated. Wilson *et al.* [138] use interaction graphs to report meaning to online social links by quantifying user interactions. They analyzed interaction graphs derived from Facebook user traces and showed that they exhibit significantly lower levels of the "small-world" properties shown in their social graph counterparts. Recently, Gilbert *et al.* [66] used Facebook data to demonstrate that the "strength of ties" varies widely, ranging from pairs of users who are best friends to pairs of users who even wished they weren't friends. Complementarily, Viswanath *et al.* studied the evolution of activity between users at Facebook, investigating how pairs of users in a social network interact and examining how the varying patterns of interaction affect the overall structure of the activity network.

Textual commentaries posted in Forums were also approached in the light of social network analysis. Gómez *et al.* [72] analyzed the social network emerging from user comment activities on Slashdot, a popular technology-news website that publishes frequently short news posts and allows readers to comment on them. They confirm several social network properties such as a giant component, small average path length and high clustering coefficient, but found moderate reciprocity and neutral assortativity by degree. Finally, they studied the dissemination tree in respect to identifying how controversial a post was. Zhou *et al.* [145] use dynamic probability model to predict the tendency of a topic discussion. They test their model in different experiment scenarios, obtaining accurate results. Lastly, a Java Forum, a large online help-seeking community, is analyzed in [144] using social network analysis methods. As the main contribution, the authors propose the use of the PageRank algorithm [37] to identify users with high expertise. Ali-Hasan *et al.* [15] conduct a study about correlation of social relationships and user interactions in blogs. Through an online survey of three communities they show that few of the blogging interactions reflect close offline relationships and many online relationships were formed through blogging.

A few references have studied the use of third-party applications and the properties of the interactions they allow. Golder *et al.* [70] analyzed temporal access and social patterns in Facebook. They analyzed the message header exchanged by Facebook users, revealing periodic patterns in terms of messages exchanged on that network. Particularly, they reveal how different forms of interaction are utilized for performing different functions on the social capital by users of a social network site - they report that while the vast majority of messages are sent to friends (90.6%), a large proportion (41.6%) is sent to friends outside of one's local network - indicating that messaging was

primarily being used for building new connections. In comparison, 'pokes' (a form of application interaction) were primarily exchanged within a network (98.3% of all pokes were within a network) to maintain or strengthen relationships. Gjoka *et al.* [69] have studied application usage workloads in Facebook and the popularity of applications. They showed that the popularity of applications has skewed distribution and that, although the number of users using the applications increases with time, the average user activity decreases. Additionally, they show that users with more applications installed are more likely to install new ones. Nazir *et al.* [109] analyzed application characteristics in Facebook, by developing and launching their own applications. Particularly, they specifically looked at interactions seen on applications as a measure of formation of online communities. By analyzing applications' interaction graphs they found high clustering of nodes, even though the community structures extracted were formed of many different geographical locations.

Valafar *et al.* [131] studied photo interactions in Flickr. As a result, they show that a very small fraction of users in the main component of the friendship graph is responsible for the vast majority of photo interactions and also show that these interactions involve only a small fraction of photos in Flickr. They further characterize temporal properties of fan arrival and show that there is no strong correlation between age (time on the system) and popularity of a photo, but most photos gain a majority of their fans during the first week.

Flickr has also been the theme of other interesting studies. Cha *et al.* [43] investigate how information disseminates through social links in Flickr. They use a theoretical epidemiological model to determine photos with high probability to have popularity in the future. More recently, Cha *et al.* [44] focused on the spatial and temporal popularity patterns of Flickr photos, finding that popular photos gained popularity over an extended period of time. Finally, Mislove *et al.* [104] studied the growth of Flickr and quantified evolutionary aspects of the friendship social network created in Flickr.

The social network created by other communication mechanisms such as cell phones and emails was also approached in the literature. Kossinets *et al.* [89] studied a university email network to identify the information "backbone," where information has the potential to flow the quickest. They found that this backbone is a sparse graph with a concentration of both highly embedded edges and long-range bridges. In [117] the network formed by mobile communication is studied. The authors examine communication patterns of people on mobile phones and argue that the stability of the communication network largely depends on the weak ties (i.e. acquaintances) in the network. Additionally, the social network formed by phone calls was also studied [124]. They found that distributions such as phone calls per costumer and distinct number of

calling partners per customer follow power law distributions. Additionally, they studied evolutive process of their graph, modeling such generative process with a double Pareto LogNormal distribtuion.

Lastly, there is also a rich set of related work focused on the interplay between the social network structure and how information is disseminated as a result of social interactions. Particularly, Gruhl *et al.* [73] studied the diffusion of information in the blogosphere based on the use of keywords in blog posts. Adar and Adamic relied on the explicit use of URL links between blogs to track the flow of information [11]. Recently, Leskovec *et al.* [96] used a framework for tracking short, distinctive phrases that travel relatively intact through online media. They observed a typical lag of around 2.5 hours between the peaks of attention to a phrase in the news media and in blogs respectively, with a "heartbeat"-like pattern in the hand-off between news and blogs. Choundhury *et al.* [47] proposes a model for predicting communication flow in OSNs. Their approach consists of using a set of contextual features to predict the intent of a user to communicate. By using support vector regression over a dataset collected from MySpace they obtain accurate predictions. In [48] they develop a framework to characterize social network dynamics in the blogosphere at individual, group, and community levels. They correlate communication dynamics in the blogosphere with stock market movement, observing considerable correlation in terms of activity. In [50], they propose a mathematical framework to measure the *interestingness* of textual conversations about a certain video on YouTube. They collect commentaries on YouTube videos and show that their method to measure the *interestingness* of a conversation is better than baseline methods (number of comments, number of new participants) comparing measures related to participation in related themes, participant cohesiveness, and theme diffusion. Finally, Choundhury *et al.* [49] propose a framework to predict synchrony of action in online social media. Synchrony is a temporal social network phenomenon in which a large number of users are observed to mimic a certain action over a period of time with sustained participation from early users. They collected a dataset of users from Digg, containing their social graph information as well as information about the stories they vote (i.e. "digg") to show that their model obtain low error (about 15%) in predicting user actions.

3.4 Video sharing systems

Differently from traditional OSN sites, video sharing systems such as YouTube provide features that allow users to interact in unique forms. Next, we survey studies that aim

at understanding properties of these systems.

An in-depth study of two video sharing systems is presented in [42]. The authors analyzed popularity distribution, popularity evolution and content characteristics of YouTube and a popular Korean video sharing service. They also analyzed system issues that could be used to improve video distribution mechanisms, such as caching and peer-to-peer distribution schemes. Additionally, they present the first evidences about the existence of duplicates in user generated content through a dataset of duplicates. They analyze the creation of duplicates and discuss potential problems that duplicates can cause to the system.

Gill *et al.* [67] discusses a characterization of the YouTube traffic collected at the University of Calgary campus network and compares its properties with those previously reported for Web and media streaming workloads. They found that HTTP GET requests, used for fetching content from the server, corresponds to 99.87% of the total requests sent to the server. They show that requests sent from the campus to the server follow typical daily and weekly patterns and quantified typical file sizes, video duration, video bit rate, video age, video rating, and video category. More recently, the same authors used the same workload to characterize user sessions on YouTube [68]. They define the duration of a typical user session as 40 minutes and then they analyze several user session characteristics, such as session duration, inter-transaction times, and the types of content transferred per user sessions. They found that YouTube users transfer more data and have longer think times than traditional Web workloads. Zink *et al.* [146] also characterizes traffic collected from a university campus and provide trace-driven simulations that show that client-based local caching, P2P-based distribution, and proxy caching can reduce network traffic significantly and allow faster access to videos.

Concerning improvements for video sharing systems, Baluja *et al.* [19] proposed a method to provide personalized video suggestions for YouTube users. The method is based on a simple graph where each video is a node linked to other videos often viewed together. Additionally, Halvey *et al.* compares user searching patterns for media and text, finding that many techniques based on text analysis could apply to the video context [76].

The quality of YouTube's videos is approached in [53]. The authors monitor a set of YouTube videos, showing that the activity on these videos follows a Poisson process. Additionally, they propose to use time series as a dynamic signature to distinguish three different types of videos, according to their quality: (1) Viral videos are those with precursory word-of-mouth growth resulting from epidemic-like propagation through a social network. (2) Quality videos are similar to viral videos but experience a sudden

burst of activity (e.g., featured videos on the first page of YouTube site). (3) Junk videos are those that experience a burst of activity for some reason (spam, chance, etc.) but do not spread through the social network.

In [126], the authors present a method for predicting the popularity of YouTube and Digg content from early measurements of users access. For Digg, measuring access to given stories during the first two hours allows them to forecast their popularity 30 days ahead, while downloads of YouTube videos need to be followed for 10 days to attain the same accuracy. Such difference is shown to be due to differences in how content is consumed: Digg stories quickly become outdated, whereas YouTube videos are still found long after they are initially submitted to the portal.

3.5 Opportunistic interactions

Next we discuss different forms of opportunistic interactions in OSN sites as well as some of the approaches used to minimize the problem. Later in this thesis, we propose a mechanism to detect users on YouTube that create this form of interaction using the video response feature.

Opportunistic interactions have been observed in various online systems, including e-mail [71], Web search engines [62], blogs [128], and OSNs [147]. The problem appears in different forms in OSNs, including unwanted invitations to join communities or create friendship connections and the disruption of legitimate communication with unsolicited marketing or propaganda posted as comments (or videos) to forums, blogs, and videos. As example of new forms of spam, Jindal *et al.* [84] studied spam in the context of Amazon's product reviews. They call opinion spam the reviews for products that are automatically generated as an attempt to promote an advertise or fool consumers with fake reviews about products. Based on a characterization of spam in this environment, they propose initial solutions to detect opinion spam.

With spam increasing in different online systems, different strategies to detect them have been proposed. A common strategy is based on content filtering. This technique requires that the recipient part inspects the content of the message and remove the undesirable ones. Such strategy is widely used in email filtering systems like SpamAssassin [spa]. Similarly, content filtering was also applied to detect blog spam [128; 102]. Another strategy is white listing, a technique in which each user specifies a list of users who they are willing to receive content from. Particularly, "RE" [65] is a white-listing system for email based on social links that allows emails between friends and friends-of-friends to bypass standard spam filters. Socially-connected users

provide permissions for each others' email messages while keeping users' contacts private. In the context of OSNs, LinkedIn uses implicit white listing of a user's friends and offers a manual introduction service based on the social network.

There has been some efforts that rely on machine learning to detect spam. Castillo *et al.* [40] proposed a framework to detect Web spamming which uses a classification approach and explore social network metrics extracted from the Web graph. Classification has also showed to be efficient to detect image-based email that contains spam [18; 139].

Finally, a way of making the life of originators of interactions harder is to exploit the difficulty in establishing and maintaining relationships in OSNs. Particularly, Mislove *et al.* [106] propose Ostra, a mechanism that imposes a cost for senders for each communication. Although Ostra showed to be an effective solution to filter unwanted communication in OSNs, it is susceptible to sybil attacks, since it assumes that users have unique identities. A Sybil attack occurs when a single attacker creates a large number of online identities, which when colluding together, allows the attacker to gain significant advantage in a distributed system. Sybil identities can work together to distort reputation values, out-vote legitimate nodes in consensus systems, or corrupt data in distributed storage systems. SybilGuard [143; 142] proposes using social network structure to detect Sybil identities in an online community to protect distributed applications. It relies on the fact that it is difficult to make multiple social connections between Sybil identities and legitimate users. The result is that Sybil identities form a well-connected sub-graph that has only a limited number of connection edges (called attack edges) to the legitimate network. The success of SybilGuard relies on the premise that Sybil identities cannot easily establish trusted social relationships with legitimate users, and hence have few "attack edges" in the social network. However, SybilGuard requires connected users to exchange encryption keys.

3.6 Discussion

Compared to the studies discussed in this chapter, this thesis focuses on characterizing the user behavior across activities, beyond the use of a single application. Additionally, although previous studies provide a valuable study on the usage of some OSNs features, they do not approach the interactions that emerge from the use of OSNs features. Our work uniquely addresses a number of issues related to interactions. First, we do not reconstruct user interactions from "visible" actions, such as text messages posted to blogs, videos, or profiles. Instead, we study a number of user actions in an OSN system

accounting with interactions that come from actions such as browsing a profile page or viewing a photo. Second, we investigate the characteristics of the interactions that emerge from video responses on YouTube, a new form of interaction that occurs based on video objects.

Lastly, the current strategies used to reduce or detect opportunistic interactions are not directly applied to video interactions since content based detection techniques are not easily applied to non-textual video objects. Additionally, white listing would not be deployable given that users do not watch video only from their friends in systems such as YouTube. In Chapter 7 we propose the use of a machine learning technique to detect the opportunistic users that post different forms of spam using the video response feature of YouTube.

Chapter 4

Datasets and Measurement Methodology

In a recent past, social network analysis was a domain of sociologists and anthropologists, when typical tools to collect social network data were surveys and interviews [133]. As a consequence, most of these efforts relies on small (and sometimes biased) datasets. With the emergence of OSN sites, the computer science community entered into the game and started gathering large scale datasets from OSNs.

In this chapter we first discuss common methods to collect data from OSNs sites (Section 4.1). Then, we describe the characteristics of the datasets used to study interactions in OSN as well as the methodology used to collect them (Sections 4.2.1 and 4.2.2).

4.1 Gathering OSN data

OSN data have been obtained through a large number of strategies. In fact, OSN data can vary from simple interviews obtained directly from users to traces collected from proxies, servers or third-party applications. Figure 4.1 illustrates a number of points where OSN data can be gathered. Next, we discuss how other researchers have collected OSN data on these points.

4.1.1 Data from users

A common method used to analyze OSN usage consists of conducting survey interviews with users. Particularly, this strategy has been widely used in the human computer

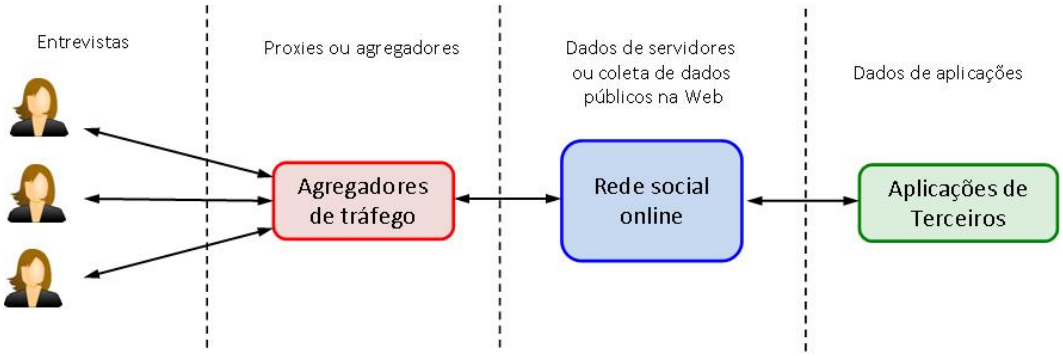


Figure 4.1: Possible data collection points

interaction community [127; 86; 45; 31; 118], where structured interviews are conducted with users of OSN sites.

4.1.2 Data from intermediary points

Two common techniques used for gathering data from an aggregation point consist of filtering OSN data from an ISP (Internet Service Provider) or obtaining the data directly from a social network aggregator. Next, we discuss efforts that made use of these approaches.

4.1.2.1 Proxy servers

Collecting data from a proxy server has been used as strategy for a number of Internet traffic studies [55; 74; 100; 137]. A proxy server can be understood as an aggregator of traffic. Such servers are usually used to delimit a portion of network, in which nodes have the same geographic location. There are few efforts that used proxy servers to collect OSN data. Gill *et al.* [67] characterized YouTube traffic collected from a proxy at the university of Calgary. They also studied session's characteristics of YouTube users [68]. The impact of caches for YouTube videos was approached in [146], based on data collected with similar methodology. More recently, Schneider *et al.* [123] extracted OSN data from an Internet service provider (ISP) and reconstructed user actions in OSN sites from the collected requests in order to study typical usage of OSN features.

4.1.2.2 Social network aggregator

Social network aggregators pull content from multiple social networking sites to a single location, thereby helping users who belong to multiple networks to manage diverse profiles more easily [King; Schroeder]. Upon logging into a social network aggregator,

users can access their social network accounts through a common interface, without having to login to each OSN site separately. This is done by a two-level real-time HTTP connection: the first level is between a user and a social network aggregator site and the second is between the social network aggregator site and the OSN sites. Social network aggregators typically communicate with OSN sites using Open APIs that OSN sites provide [ope]. All content from OSN sites are shown to users through a social network aggregator's interface. Figure 4.2 depicts the scheme interaction among users, a social network aggregator site, and OSN sites. Through the interface of the social network aggregator, a user can enjoy all features that are provided by OSN sites, for instance, checking updates from friends, sending messages, and sharing photos.

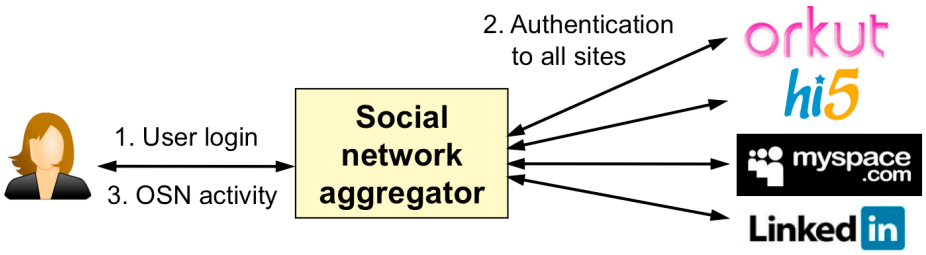


Figure 4.2: Illustration of a user connecting to multiple OSNs through the social network aggregator

4.1.3 Data from OSN servers

Ideally, OSN servers are the most suitable places to collect traces of user interactions and friendship links. However, most of these sites are hesitant to provide even anonymized data. There are few works that use data obtained directly from an OSN server. Particularly, Chun *et al.* [51] obtained guestbook logs of Cyworld directly from the server. Burke *et al.* [39] obtaining logs from approximately 140,000 newcomers in Facebook. Baluja *et al.* [19] used YouTube logs obtained from the server. Finally, Duarte *et al.* [58] studied blogs from the server point of view and we used data from a Brazilian video sharing system to characterize user navigation patterns [23].

Given the difficult to obtain data directly from servers, a common strategy used to obtain OSN data consists of visiting pages of OSN sites and gathering public information about users and objects. This strategy is called *crawling*. Such strategy has been applied to a different range of work, including studies about topology structure [105; 12], access patterns [42] and interactions reconstructed based on visible actions such as textual messages [132].

Crawling an OSN is usually different from crawling traditional Web pages. The Web pages of OSNs are typically well-structured, as they are usually automatically generated, unlike traditional Web pages which can be created by any person. Therefore, since each individual in an OSN is identified by a unique identifier, we can be sure about what pieces of data we can obtain after crawling a particular individual's Web pages.

Typically in OSN links between users can also be crawled automatically, representing an opportunity to measure and study social networks at large scale. By crawling these links we can construct the graph of connections between all the users in the social network. Ideally, we would like to obtain the entire social network graph that would correspond to the desirable measurement scenario in order to avoid measurement bias that can lead to inaccurate results. However, most of times, there is no systematic way of collecting the entire graph available on the OSN service. For these cases, it is necessary to sample part of the social network graph. In order to minimize any possible bias when crawling OSNs there are some strategies that are usually employed:

- **Crawling the largest WCC**: Since most real-world graphs have a dominant large weakly connected component (WCC) [38], a common approach to sample OSN graphs consists of collecting this entire component. Particularly, Mislove *et al.* [105] shows that the largest WCC is structurally the most "interesting" part of a social network, where the most activity users can be found. They show that the users not included in the largest WCC tend to be either part of very small, isolated clusters or are not connected to other users at all.

  There are two well known strategies to collect a WCC in an OSN, called breadth first search (BFS) and depth first search (DFS). Crawling only a subset of a graph by ending a BFS early (called snowball method) is known to produce a biased sample of nodes [93; 12], underestimating power-law coefficient for node degree.

- **Crawling both directions on directed graphs**: Some OSN systems offer direct friendship links between users. As example, in Twitter a user can follow another user but not be followed back. Crawling directed graphs, as opposed to undirected graphs, presents additional challenges. Crawling a graph using only one direction of the links (e.g., forward links) does not necessarily cover an entire WCC. Instead, it collects the reachable nodes from a set of seed nodes. This limitation is typical of studies that crawl online networks, such as the Web [38]. Typically, the Web is often crawled by following links in the forward direction since one cannot easily determine the set of nodes which point into a given page.

4.1.4 Data from OSN applications

Another approach used to study part of the activities that users do in OSNs consists of studying user access to third-party applications. Third-party applications are characterized by the presence of the OSN server as an intermediary for all communication between the client and the application server. Typically, a client forwards a request to the OSN server, which forwards it to the application server. Then, the application server sends the response back to the OSN server, which sends it to the client [110]. OSN applications can be used to study not only how users use certain applications, but they can also be useful to gather user information. Third-party applications can ask users permission to access information such as list of friends and activities performed during a session.

4.2 Our datasets

Next, we present the two datasets used in this work. The first dataset consists of data from an OSN aggregator. The second dataset consists of data from YouTube, used to study video interactions.

4.2.1 Dataset 1: OSN aggregator dataset

Online social network aggregators are one-stop shopping sites for OSNs and provide users with a common interface for accessing multiple social networks [King]. We collaborated with a popular OSN aggregator in Brazil, obtaining clickstream data from OSNs, which capture each "click" of the users that accessed the OSN aggregator site during the period the data were collected. Next we describe the characteristics of our dataset.

4.2.1.1 Clickstream dataset

The clickstream data that we analyzed were collected over a 12-day period (March 26 through April 6, 2009). The data consist of summaries of HTTP header information for traffic exchanged between the social network aggregator server and users. The dataset summarizes 4,894,924 HTTP requests, including information about time stamp, HTTP status, IP address of the user, login ID in the social network aggregator site, URL of the social network site, login ID within the social network site, session cookies, and the traffic bytes sent and received. After discarding events with missing fields or HTTP status associated with error codes (e.g., 301, 302), there were 4,649,595

valid HTTP requests. HTTP requests in the trace are grouped into sessions, where a session represents the sequence of a user's requests during a single visit to the social network aggregator. The trace included 77,407 sessions, covering 16,175 distinct user IP addresses and 37,137 distinct login IDs in the social network aggregator site.

Not all log entries in the trace were related to accessing OSNs. Some log entries reflect users accessing non-OSN features of the aggregator site, such as listening to an Internet radio or watching videos. Other log entries result from the automatic display of advertisements and the aggregator site's website logo. After discarding non-OSN related log entries, 802,574 or 17% of the HTTP requests were related to accessing the following four OSNs: Orkut, Hi5, MySpace, and LinkedIn. In Chapter 5 our analyses are focused on these HTTP requests related to accessing OSNs.

Table 4.1 displays the number of users, sessions, and HTTP requests for these OSNs. Among them, Orkut had the largest number of users and accounted for nearly 98% of all HTTP requests. The remaining OSN sites take up only 2% of the trace.

OSNs	# users	# sessions	# requests
Orkut	36,309	57,927	787,276
Hi5	515	723	14,532
MySpace	115	119	542
LinkedIn	85	91	224
Total	37,024	58,860	802,574

Table 4.1: Summary of the clickstream data

4.2.1.2 Social network topology of Orkut

In order to gain insight into user behaviors over the social graph, we crawled the largest OSN site in the trace, Orkut. Because of the sheer size of the Orkut network, we decided to crawl friendship information for only those users that appear in the clickstream dataset. We used the Orkut user IDs that appear in the trace, prior to anonymization. We implemented a crawler which downloaded the profile page of each of these Orkut users. A profile page contained a variety of information about users. Certain profile information is made publicly available to all Orkut users, for instance, the list of friends, the list of community memberships, name, gender, and country. On the other hand, other information like email, phone number, and age is set private and is shown only to friends by default. When crawling Orkut, we stored all profile information that is made publicly available.

We gathered the profile information of the 36,309 Orkut users the week after the clickstream data were gathered, during April 10–17, 2009. The average number of friends was 211.4 and the median number of friends was 152. Some users had no listed friends at all, while the user with the highest number had 998 friends. Orkut allows a user to have at most 1,000 friends. The IDs of users in the crawled social graph were anonymized in the same way as the clickstream data.

4.2.1.3 Dataset limitations

Although the clickstream data give us a unique opportunity to study user activities across multiple OSNs, the dataset has limitations.

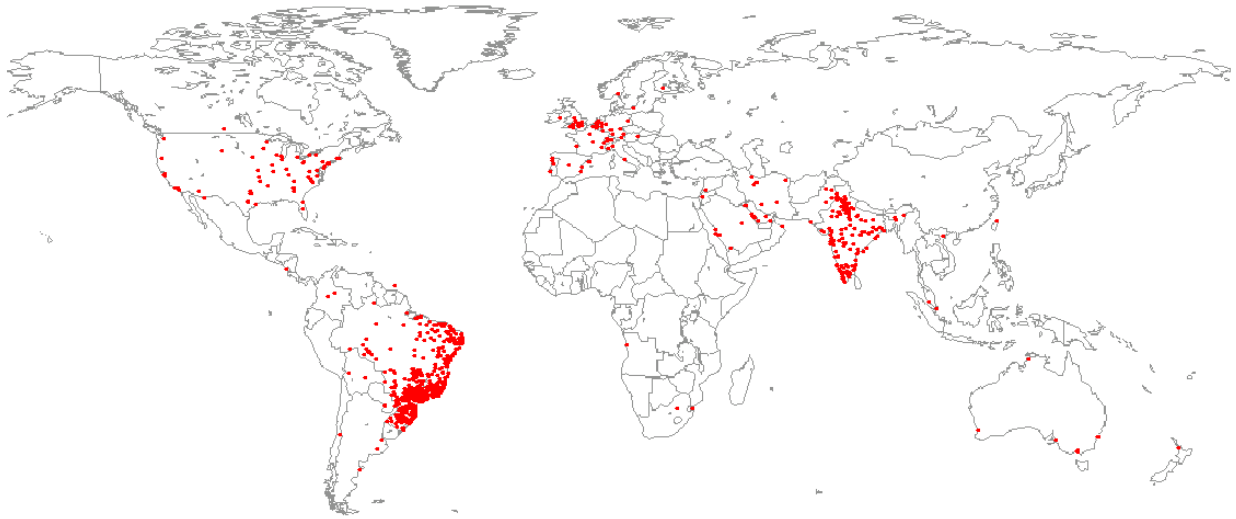


Figure 4.3: Location of the social network aggregator users

First, the dataset is biased towards the set of users in the social network aggregator portal. One evident bias is the demographics of users in Orkut. To examine the geographical distribution of users, we used the GeoIP database [MaxMind] to identify the location of 16,175 IP addresses that appeared in the trace. Figure 4.3 shows the location of the social network aggregator users. These users were located across all continents in the world, spanning several countries. However, certain geographical locations contained more users than others. Particularly, Brazil had the highest presence both based on the number of IP addresses (71%) and the number of the HTTP requests (70%). The second largest user base came from India and accounted for 12% of the IP addresses and 14% of the requests. The third most common location was the United

States. The bias in user samples may raise a concern about how representative our results are for other social networks in the data, i.e., Hi5, MySpace, and LinkedIn.

Second, we are not able to study user navigation and interactions over a long term period (e.g., several months) since the data were collected only over a 12-day period.

Despite the partial control of personal data, Orkut allows anyone with an account to freely browse lists of friends. Thus, we do not have limitations related to restrictions to crawl the friendship topology of the users in our clickstream data. Additionally, since we crawled this data on the week after the clickstream data were gathered, we believe that the friendship does not present significant changes able to affect our results.

4.2.2 Dataset 2: Video response dataset

In order to analyze social network aspects of video interactions, we need to obtain information about users and their interactions (via video responses) from YouTube. To do it, we can sequentially visit pages on the YouTube site (that is, crawl) and gather information about YouTube video responses and their contributors. Every YouTube video post has a single contributor, who is a registered YouTube user. We say a YouTube video is a *responded video* or *video topic* if it has at least one video response. A video topic has a sequence of video responses listed chronologically in terms of when they are uploaded to the system[1]. We say a YouTube user is a *responded user* if at least one of her contributed videos is a responded video. Finally, we say that a YouTube user is a *responsive user* if she has posted at least one video response.

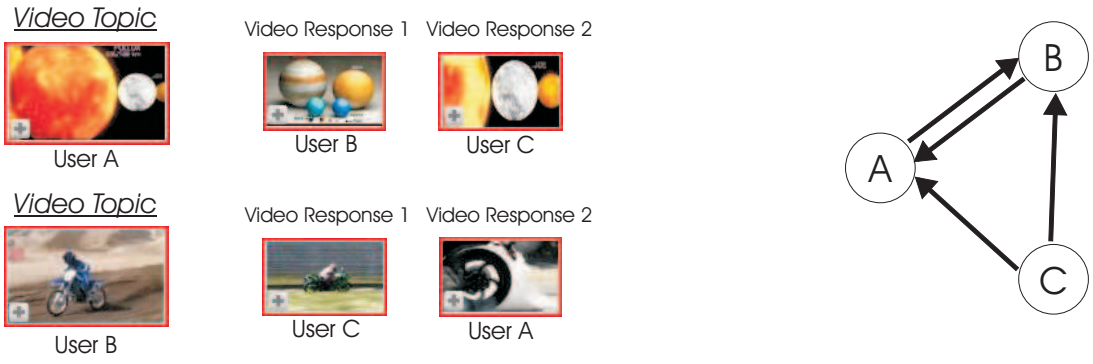


Figure 4.4: Video responses posted to two video topics (left) and the graph created by these video interactions (right)

[1]We note that YouTube does not allow a video to be posted as response to more than one video topic.

4.2.2.1 Video response user graph

The first step to collect an OSN site consists of defining the sequence of nodes and links to be gathered. A natural user graph emerges from video response interactions. At a given instant of time t, let X be the union of all responded users and responsive users. The set X is, of course, a subset of all YouTube users. We denote the *video response user graph* as the directed graph (X, Y), where (x_1, x_2) is a directed arc in Y if user $x_1 \in X$ has responded to at least one video contributed by user $x_2 \in X$. Figure 4.4 illustrates two video response sequences and the graph established by these interactions. We note that the video response user graph may have multiple weakly connected components.

Ideally, we would like to obtain the complete video response user graph (X, Y), which is equivalent to finding *all* the video responses and responded videos. Unfortunately, it is difficult to locate all video responses in YouTube. Although the YouTube site provides lists of the 100 most responded videos of all time, it does not currently provide a means to systematically visit all the responded videos. YouTube has numerous hyperlinking mechanisms from users to videos (e.g., users' favorite videos) and from videos to videos (e.g., related videos). But even by following all the hyperlinks among the videos and users, we may not find all the users in X, since some of the users in X are not hyperlinked from any of the crawled pages. In particular, it is difficult to find the users in the small components of the graph (X, Y).

Instead, we design a sampling procedure that allows us to obtain a large representative subset A of X, described next.

4.2.2.2 Sampling strategy

For a given sampled subset A of X, let (A, B) be the directed graph, where (a_1, a_2) is a directed arc in B if user $a_1 \in A$ has responded to a video contributed by user $a_2 \in A$. Note that sampled graph (A, B) is a sub-graph of (X, Y). In order to sample YouTube data, we designed a distributed crawling framework similar to the one presented in [46]. Our distributed crawler is composed of a master node and a number of slave nodes. The master node maintains a centralized list of users to be crawled, which is initialized with a set of seeds. The master is also responsible for coordinating the operation of the slaves, sending non-overlapping subsets of this list to them, thus preventing redundant crawling. The slaves, after obtaining the user identifiers from the master, crawl YouTube following Algorithm 1, return all new users collected to the master, and wait for a signal from it. The master, in turn, eliminates duplicate or previously crawled users from the received lists, and, in case there are still uncrawled users, starts

a new round of crawling by sending new user identifiers to the slaves. Otherwise, the master sends a signal for termination to the slaves and stops execution. We note that the crawling process is terminated only after an entire weakely connected component of graph (X, Y) has been collected. We ran our crawler using 10 Linux boxes (1 master and 9 slaves) located at the Federal University of Minas Gerais in Brazil.

Input: A list L of users received from master node
1: **for each** User U in L **do**
2: Collect U's information and video list;
3: **for each** Video V in the video list **do**
4: Collect information of V;
5: **if** V is a responded video **then**
6: Collect information of V's video responses;
7: Insert the responsive users in list of new users NL;
8: **end if**
9: **if** V is a video response **then**
10: Insert the responded user in list of new users NL;
11: **end if**
12: **end for**
13: **end for**
14: Return NL to the master node;

Algorithm 1 Video response crawler (run by slave nodes)

We used the set of all users who contributed with the videos listed by YouTube as the all-time top-100 most responded videos[2] as seeds to our crawler. These consisted of 92 users. Our crawler produced the graph (A, B), composed of a large weakly connected component of graph (X, Y). The crawler ran for 5 days (September 21^{st}-26^{th}, 2007), gathering a total of **1**60,765 users, **2**23,851 responded videos and **4**17,759 video responses. For each video that was crawled, we collected a number of pieces of information available, including video identifier, video contributor identifier, title, category, description, tags, upload time, video duration, number of ratings, average rating, number of views, number of users who set the video as favorite, number of comments received and number of video responses received. We also collected the author of the video responses of each video, and the sequence order in which the video responses were posted.

4.2.2.3 Sample validation

It is desirable that the sampled set of users A have the following three properties:

[2]YouTube usually provides lists of top videos or users with size 100.

- **Property 1**: Each connected component in (A, B) is a connected component in (X, Y), that is, the sampled subgraph (A, B) consists of (entire) connected components from (X, Y). As discussed in Section 4.1.3, this property is important in order to analyze the social networking interactions.

- **Property 2**: The subset A covers a large fraction of X (more than 50% indicates that A is the largest weakly connected component). Our sample would then include the majority of the responded and responsive users. Note that our interest is in crawling a sample of only those YouTube users who make use of the video response feature, who is a (possibly much smaller) subset of all YouTube users.

- **Property 3**: The most responded users are included in A. This last property ensures that we are including the most important users, and only neglecting users who have few responded videos.

Input: A list of words from a dictionary

1: Randomly select a word from the dictionary;
2: Search for a tag on YouTube, using the selected word as tag;
3: **for each** Contributor C of the videos found in step 2 **do**
4: **if** C is a responded **OR** a responsive user **then**
5: Add user to the result list of users;
6: **end if**
7: **end for**

8: Randomly select 100 users from the final list of users;

Algorithm 2 Selecting random seeds

Since our crawler strategy produced the large weakly connected component of graph (X, Y), our strategy satisfies property 1. In order to verify properties 2 and 3 of our graph (A, B), we ran Algorithm 1 using a random set of users as seeds. Algorithm 2 describes the mechanism used to select users randomly on YouTube, which basically consists of search for random words on YouTube and gather the user identifier that appear in the search results. Out of the 4,633 users collected with Algorithm 2, 3,231 of them belong to A. Moreover, 67 out of the 100 randomly selected seed users are also in A. Both results are evidence that our sampling scheme may satisfy property 2. In fact, our second crawl gathered a total of 182,725 users, out of which 146,799 are found in our first dataset. To verify property 3, we ranked the 10, 100 and 1000 most responded users from our second dataset. We found that graph (A, B) contains all 10

of the 10 most responded users, 98 of the 100 most responded users, and 951 of 1000 most responded users. Thus, Property 3 is verified as well.

4.3 Discussion

In this chapter we provide an overview of methods used to collect data from OSNs sites and we also describe the characteristics of the two datasets used in this thesis. The OSN clickstream data obtained from an OSN aggregator is able to capture all the interactions performed by users that accessed the OSN aggregator site during the data collection period. Since clickstream data include not only visible interactions, but also "silent" user actions like browsing a profile page or viewing a photo, they can provide a more accurate and comprehensive view of the OSN interactions. Complementarily, our second dataset allow us to explore an unique form of interaction that occur on the YouTube video response, the video interactions.

Chapter 5

User Navigation and Activities

This chapter provides a first look into the usage of OSN services from the viewpoint of a social network aggregator. We conduct three sets of analysis of OSN activities based on the clickstream dataset collected. First, we characterized the traffic and session patterns of OSN workloads (Section 5.1). We examined how frequently people connect to OSN sites and for how long. Based on the data, we provide best fit models of session inter-arrival times and session length distributions. Second, we developed an analysis strategy, which we call the *clickstream model*, to characterize user activity in OSNs (Section 5.2). The clickstream model captures dominant user activities and the transition rates between activities. We profiled user activities for four OSN services: Orkut, MySpace, Hi5, and LinkedIn. Third, to gain insight into how users interact within a given social network, we additionally crawled the Orkut website and analyzed user activity along the social graph (Section 5.3). Our analysis reveals how often users visit other people's online profiles, photos, and videos.

5.1 Connection pattern analysis

In this section, we characterize OSN workloads at the session level. We describe how sessions are identified in the social network aggregator, then examine the duration and frequency of connections to OSN services.

5.1.1 Defining a session

The social network aggregator considers the following events to determine end of a session (a) when a user closes the web browser or logs out or (b) when a user does not engage in any action for more than an arbitrarily set period of time. The social

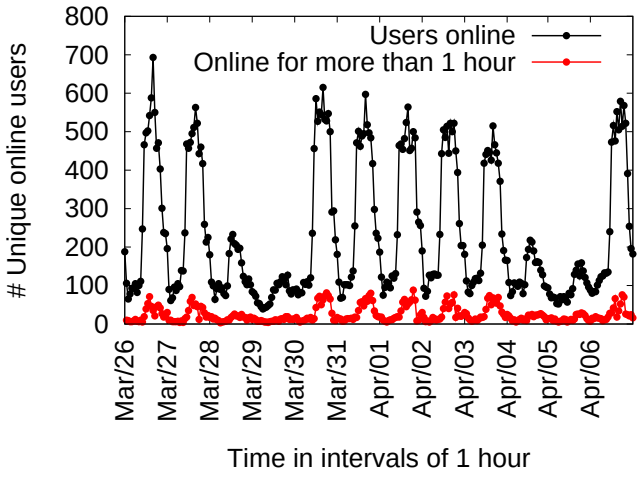
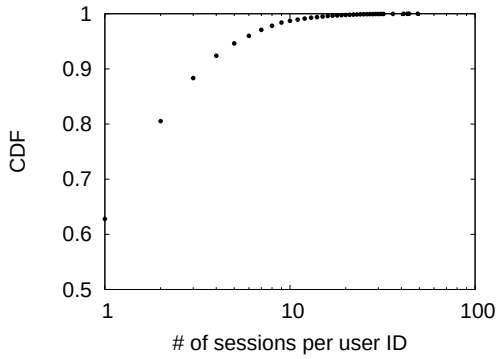


Figure 5.1: Number of online users over time

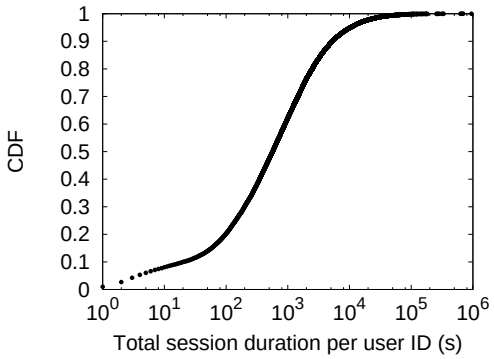
network aggregator uses a 20 minute threshold. To check the sensitivity of this session threshold, we examined whether any two consecutive sessions of the same user had a shorter interval than 20 minutes. For 22% of all sessions (generated by 13% of all users), an earlier session by the same user ended less than 20 minutes prior (i.e., 22% of sessions were solely identified by events of closing of web browsers or logging out). For analysis, we used the session information that is identified by the social network aggregator.

Utilizing the session information, we first examined the number of concurrent users (i.e., concurrent sessions) that accessed any of the four OSN sites (Figure 5.1). The beginning of each day is marked in the horizontal axis. We see a diurnal pattern with strong peaks around 3 PM (in Brazil). At all times, there are at least 50 people who are using the social network aggregator service. At peak times, the number of concurrent users surpasses 700, more than a 10-fold increase over the minimum. Drops in usage on certain days indicate clear weekly patterns, where weekends showed a much lower usage than weekdays. The strong diurnal pattern in OSN workloads has also been observed in accessing messages and applications on Facebook [70] and in the content generation of blog posts, bookmarks, and answers in user generated content (UGC) websites [75; 58].

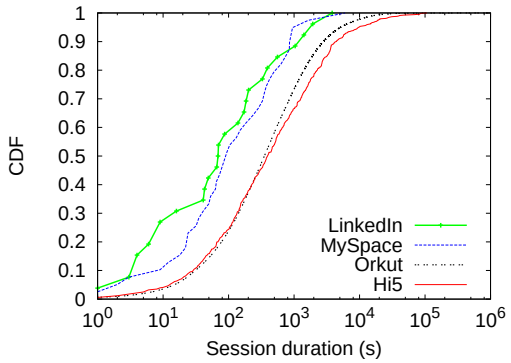
To see the usage pattern of heavy OSN users, we also show in Figure 5.1 the number of users who stayed online for more than 1 hour at any given point in time. The daily peaks for heavy users coincide with the peaks from all users. The total number of online users and the number of heavy users showed a strong correlation; the Pearson's correlation coefficient was 0.84. This indicates that the ratio between the



(a) Frequency of accesses to OSN sites



(b) Total time spent on OSNs per user



(c) Session duration per OSN site

Figure 5.2: Session level characteristics of OSN workload

heavy users and all users is oblivious to the time of day. The gap between the two data points in the figure also indicates that there are users who login and connect for less than an hour throughout the day.

5.1.2 OSN session characteristics

So, how often and for how long do people connect to OSN sites? To estimate these quantities, we measure the frequency and duration of sessions for each user. We calculate session duration as the time interval between the first and the last HTTP requests within a session. This approach allows us to infer the duration of any session with two or more HTTP requests. 87% of all sessions in the dataset contained at least two HTTP requests.

Individuals varied widely in the frequency with which they accessed social networks. Figure 5.2(a) shows the cumulative distribution function (CDF) of the total number of sessions per user. The majority of users (63%) accessed the social network aggregator's site only once during the 12-day period. The most frequently logging in

user accessed the social network aggregator's site on average 4.1 times a day. The total time spent accessing social networks also varied largely per individual, as shown in Figure 5.2(b). On one hand, 51% of the users spent no more than 10 minutes at the social network aggregator's site over the 12 days. On the other hand, 14% of the active users spent in total more than an hour and the most active 2% of the users spent more than 12 hours (i.e., an average of an hour a day).

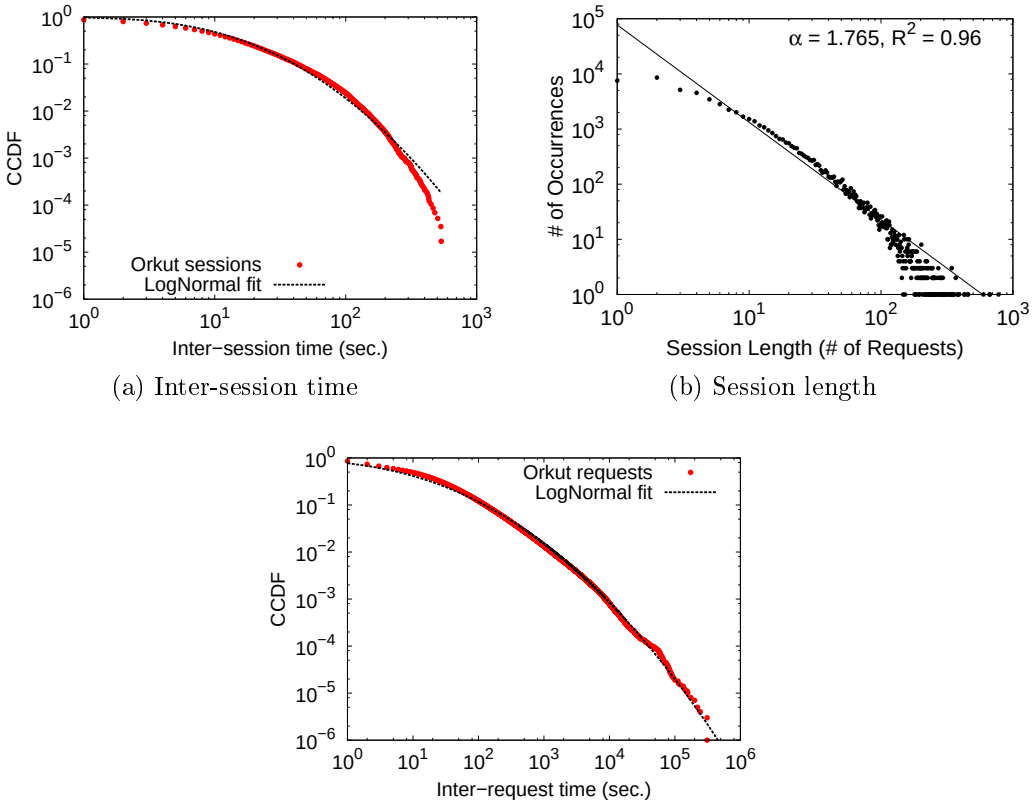
Across all users, we did not see a high correlation between the frequency and duration of OSN accesses (correlation coefficient 0.27). This means that the amount of time a user spends on social networks is not strongly correlated to the specific number of times that the user logins to social networks. We also did not see a strong correlation between a session duration and the number of HTTP requests made during the session (correlation coefficient 0.16). The correlation became relatively stronger when we considered relatively short sessions that lasted less than 20 minutes (correlation coefficient 0.49). This may suggest that long sessions tend to have idle users. For short sessions, the longer the session duration, the more activities the session contains.

In addition to widely varying OSN usage per individual, session durations also varied widely across the four OSN sites. Figure 5.2(c) shows the CDF of the session durations for each OSN site. All four OSN sites exhibit a consistent heavy-tailed pattern in their session durations. However, the median session durations vary across OSNs. The median session durations of Orkut, Hi5, and MySpace are 13.4 minutes, 2.7 minutes and 24 seconds, respectively, indicating that users likely engage in a series of activities when they connect to these sites. In contrast, the median session duration of LinkedIn is very short (3 seconds). In the following section, we take a deeper look into which activities are popular across these sites.

5.1.3 Modeling Orkut sessions

To understand the dynamics of user arrival and departure processes from a system's perspective, we measure the session inter-arrival times. Here, we present a case study for Orkut. More formally, we utilize a time series $t(i), i = 1, 2, 3, ...$ to denote the arrival time of the ith session in the trace. The time series $a(i)$ is defined as $t(i+1) - t(i)$ and it denotes the inter-arrival time of the ith and $i+1$th sessions, where sessions may belong to different users. Figure 5.3(a) shows the complementary cumulative distribution function (CCDF) of $a(i)$, which we fitted to a Lognormal distribution. The probability distribution function for the lognormal distribution is given by:

$$f(x) = \frac{1}{\sigma x \sqrt{2\pi}} e^{-(log(x)-\mu)^2}/2\sigma^2 \tag{5.1}$$



(a) Inter-session time

(b) Session length

(c) Inter-request time

Figure 5.3: Characteristics of Orkut sessions and the best fit functions

with parameters $\mu = 2.245$ and $\sigma = 1.133$.

To characterize the period of time during which a session is active, we use a time series $l(i)$ which denotes the length of the ith session in the trace, defined as the number of requests in that session. Figure 5.3(b) shows the marginal distribution of $l(i)$ for all sessions identified in the Orkut trace. We observe a heavy-tail distribution; most of the sessions involve very few HTTP requests, while a small number of sessions involve a large number of HTTP requests. This implies significant deviations in the number of actions (or clicks) users make in a single session.

The distribution was fitted to a Zipf distribution of the form $\beta x^{-\alpha}$ with parameters $\alpha = 1.765$ and $\beta = 4.888$. A Zipf-like distribution suggests that session lengths are highly variable when users connect to online social networks. Such high variability is in line with the patterns seen in web surfing. Huberman et al. [79] also found strong variability in the number of clicks a user exhibits in a session, as well as when navigating a given website.

The last variable we characterize at the session layer is the inter-arrival time

between requests within a single session. Figure 5.3(c) displays the CCDF distribution that was fitted to a Lognormal distribution, with parameters $\mu = 1.789$ and $\sigma = 2.366$. Large inter-arrivals would correspond to users leaving Orkut pages to spend time on other social networks or other features of the social network aggregator then returning back to Orkut. On the other hand, small inter-arrivals would correspond to users constantly interacting with the social networking site. We found that the average session lengths and the session starting times are not correlated (the Pearson's correlation coefficient is -0.027). This suggests that the high variability in session length is not due to diurnal pattern in user behaviors (as was the case with the number of active clients), but rather it is a fundamental property of the interaction of OSN users.

The combination of request inter-arrival time and session length provides an important model for understanding the behavior of OSN users, for the two quantities reflect the inherent nature of OSN users and are not related to load (e.g., the number of active sessions) or time of the day. The best fit distribution functions presented in this section can be used to generate synthetic (parameterizable) traces, that mimic actual OSN workloads.

5.2 User navigation patterns in OSNs

In this section we present a comprehensive view of user behavior in OSNs by characterizing the type, frequency, and sequence of activities users engage in. We developed a new analysis strategy, which we call the *clickstream model*, to identify and describe representative user behaviors in OSNs based on clickstream data.

The modeling of the system implies two steps. The first step is to identify dominant user activities in clickstreams. This step involves enumerating all features users engaged in on OSNs at the level of basic unit, which we call *user activity*. We manually annotated each log entry of the clickstream data with the appropriate activity class (e.g., friend invitation, browsing photos), based on the information available in the HTTP header. Because a user can conduct a wide range of activities in a typical OSN site, we further tried to group semantically similar activities into a *category* by utilizing the Web page structure of OSN sites (i.e., which set of activities can be conducted in a single page) and manually grouping related activities into categories.

The second step of modeling is to compute the transition rates between activities. To represent the sequence in which activities are conducted, we built a first-order Markov chain of user activities and compute the probability transition between every pair of activity states. To gain a holistic view, we built a Markov chain that describes

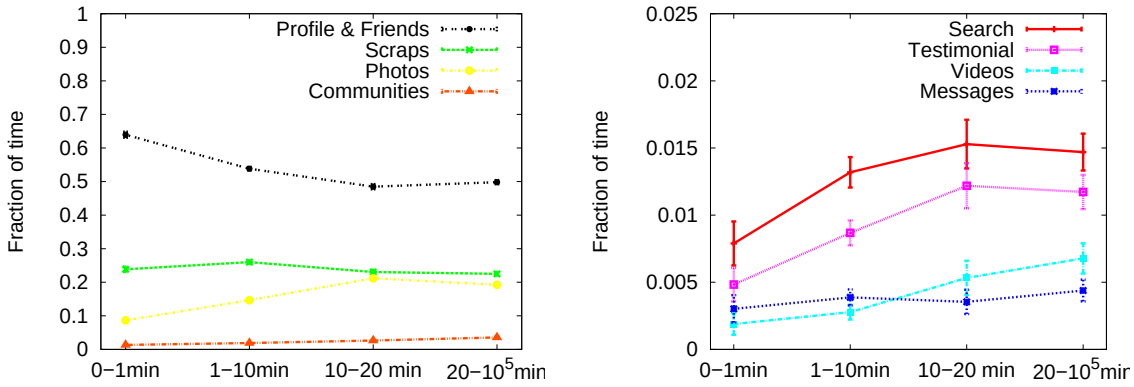
how users transition from actions in one category to another.

Different OSNs provide different features, potentially leading to a substantial variation in the set of popular user activities. Our analysis in this section highlights the similarities and differences in user behaviors across four different social networks in the trace. Below we present the full clickstream model only for Orkut, which is the most accessed OSN in the trace.

5.2.1 Identifying activities in Orkut

In the first step of modeling, we identified 41 activities with at least one HTTP request in the clickstream data. We grouped these activities into the following categories: Search, Scrapbook, Messages, Testimonials, Videos, Photos, Profile & Friends, Communities, and Other. Table 5.1 displays the list of 41 activities with the number and share of users who engaged in the corresponding activity at least once, the number and share of HTTP requests, and the total traffic volume both received and sent by users.

The activity categories listed in Table 5.1 represent the following features in Orkut, which are described in more detail in [6; 7]:



(a) Probability of dominant activities (b) Probability of subdominant activities

Figure 5.4: Probability of the user activity as a function of session duration (error bars indicate 95% confidence interval)

- **Universal search** (activity 1) allows users to search for other people's profiles, communities, and community topics (or forums) in the entire Orkut website. A search box appears at the upper right corner of every Orkut page, allowing users to engage in the search feature from any page.
- **Scrapbook** (activities 2 and 3) displays all text messages sent to a given user. Unlike personal messaging or email, Scrapbook entries are public, meaning that

Category	ID	Description of activity	# Users	(%)	# Requests	(%)	Bytes (MB)
Search	1	Universal search	2,383	(2.1)	15,409	(2.0)	287
Scrapbook	2	Browse scraps	17,753	(15.9)	147,249	(18.7)	2,740
	3	Write scraps	2,307	(2.1)	7,623	(1.0)	113
Messages	4	Browse messages	931	(0.8)	3,905	(0.5)	64
	5	Write messages	70	(0.1)	289	(<0.1)	5
Testimonials	6	Browse testimonials received	1,085	(1.0)	3,402	(0.4)	57
	7	Write testimonials	911	(0.8)	4,128	(0.5)	65
	8	Browse testimonials written	540	(0.5)	1,633	(0.2)	26
Videos	9	Browse the list of favorite videos	494	(0.4)	2,262	(0.3)	44
	10	Browse a favorite video	390	(0.3)	862	(0.1)	13
Photos	11	Browse a list of albums	8,769	(7.8)	43,743	(5.6)	871
	12	Browse photo albums	8,201	(7.3)	70,329	(8.9)	2,313
	13	Browse photos	8,176	(7.3)	122,152	(15.5)	1,147
	14	Browse photos the user was tagged	1,217	(1.1)	3,004	(0.4)	47
	15	Browse photo comments	355	(0.3)	842	(0.1)	16
	16	Edit and organize photos	82	(0.1)	266	(0.0)	3
Profile &	17	Browse profiles	19,984	(17.9)	149,402	(19.0)	3,534
Friends	18	Browse homepage	18,868	(16.9)	92,699	(11.8)	3,866
	19	Browse the list of friends	6,364	(5.7)	50,537	(6.4)	1,032
	20	Manage friend invitations	1,656	(1.5)	8,517	(1.1)	144
	21	Browse friend updates	1,601	(1.4)	6,644	(0.8)	200
	22	Browse member communities	1,455	(1.3)	6,963	(0.9)	133
	23	Profile editing	1,293	(1.2)	7,054	(0.9)	369
	24	Browse fans	361	(0.3)	1,103	(0.1)	17
	25	Browse user lists	126	(0.1)	626	(0.1)	9
	26	Manage user events	44	(<0.1)	129	(<0.1)	2
Communities	27	Browse a community	2,109	(1.9)	8,850	(1.1)	164
	28	Browse a topic in a community	926	(0.8)	9,454	(1.2)	143
	29	Join or leave communities	523	(0.5)	3,043	(0.4)	43
	30	Browse members in communities	415	(0.4)	3,639	(0.5)	56
	31	Browse the list of community topics	412	(0.4)	2,066	(0.3)	38
	32	Post in a community topic	227	(0.2)	1,680	(0.2)	24
	33	Community management	105	(0.1)	682	(0.1)	12
	34	Accessing polls in communities	99	(0.1)	360	(<0.1)	6
	35	Browse the list of communities	47	(<0.1)	337	(<0.1)	8
	36	Manage community invitations	20	(<0.1)	63	(<0.1)	1
	37	Community events	19	(<0.1)	41	(<0.1)	1
Other	38	Accessing applications	1,092	(1.0)	4,043	(0.5)	61
	39	User settings	403	(0.4)	2,020	(0.3)	32
	40	Spam folder, feeds, captcha	48	(<0.1)	150	(<0.1)	2
	41	Account login and deletion	39	(<0.1)	76	(<0.1)	1
		Total	36,309 (distinct)		787,276		17.3 GB

Table 5.1: Enumeration of all activities in Orkut and their occurrences in the click-stream data.

anyone with an Orkut account can read others' *scraps*. By default, anyone can leave a scrap in a user's scrapbook. However, users can set their scrapbook to be private, so that only friends or friends of friends in the network can leave a scrap. Table 5.1 shows that browsing and writing scraps is one of the most popular forms of user interaction in Orkut.

- **Messages** (activities 4 and 5) are a private way to communicate. Messages can be sent by anyone. Table 5.1 shows that the messages feature is not widely used in Orkut.

- **Testimonials** (activities 6 to 8) are comments that users leave about his or her friends. Testimonials can only be written by friends, but can be viewed by anyone by default. A user can set options so that testimonials are kept private, and only the user's friends can view the testimonial page. Compared to the interaction through scrapbook, we see much less interaction through testimonials.

- The **Videos** (activities 9 and 10) and **Photos** (activity 11–16) categories incorporate all activities in which users share multimedia content. The photos category is another popular activity in Orkut. A photo can be tagged and commented on only by friends. However, a photo can be viewed by anyone by default. To share a video, Orkut asks users to first upload their videos to YouTube then to add the video URLs at the Orkut's video page.

- **Profile & Friends** (activities 17–26) represent all activities in which users manage their own profiles or visit other people's profiles. Orkut allows anyone to visit anyone's profile, unless a potential visitor is on the "Ignore List" (a list where a user specifies other users who he or she wants to block from any form of interaction). Users can customize their profile preferences and can restrict the information that appears on their profile page from other users.
  A user's homepage displays a short list of updates about the user's friends. The homepage also displays a short list of friends ordered by login time, where the first person is the one who logged in most recently.

- **Communities** (activities 27–37) can be created by anyone with an Orkut account. Community members can post topics, inform other members about an event, ask questions, or play games. Users can freely join any public community, while a moderated community requires explicit approval. Invitations to join a community are sent through messages.

The statistics of user activity in Table 5.1 suggest interesting trends in the usage of Orkut. First, we can note that the most popular activities, both in terms of the number of users and the request volume, are related to profile and friends. In fact, Orkut interface is designed in a way that users need to browse a user profile in order to do some activities, such as view the scrapbook of that user. Additionally, the user homepage works like the main portal in the social network where users can check updates from their friends, new messages received, upcoming birthdays, etc. Thus, it is natural to expect that users visit their homepage to check updates and spend more time on this page due to the high number of available information. The second and third most popular groups of activities are related to photos and scrapbook.

Interestingly, note that browsing is the most common user behavior across all categories of activities. Thus, we take a closer look at browsing related activities. We categorize them into four types: browsing of media content such as photos and videos, profile content (both one's own and others'), text messages of testimonials, scraps, and messages, and community content which belongs to not a user but a community within Orkut. Table 5.2 displays the popularity of these categories based on the fraction of associated requests and the average time spent on each state. In total, browsing accounted for **92.7%** of all requests. Compared to other non-browsing activities in the same category, browsing typically engaged 2 to 100 times more users. For instance, the number of users who ever browsed messages was 13 times larger than those who sent messages. In fact, other behaviors that require more user engagement were less prominent in the trace; time-intensive behaviors like browse a favorite video (activity 10) and participation-oriented behaviors like posting in a community topic (activity 32) are not popular.

Our findings demonstrate that many Orkut users primarily use the service for passive interactions such as browsing updates from their friends through homepage, profile pages, and scrapbook, while occasionally engaging in more active interaction such as writing scraps, searching, editing photos, and accessing applications.

Category	**Request percentage**	**Time spent**
Browsing media	30.8%	11.2 min
Browsing profile	39.1%	15.4 min
Browsing text messages	19.8%	9.6 min
Browsing community	3%	13.3 min

Table 5.2: Browsing activity in Orkut

Admittedly, these activities are not independent but are interrelated. Certain updates from friends (i.e., browsing) can lead to interaction and search will be followed by browsing activity. Therefore, it is very important to understand the relationships among these activities. In the next Section we study these relationships.

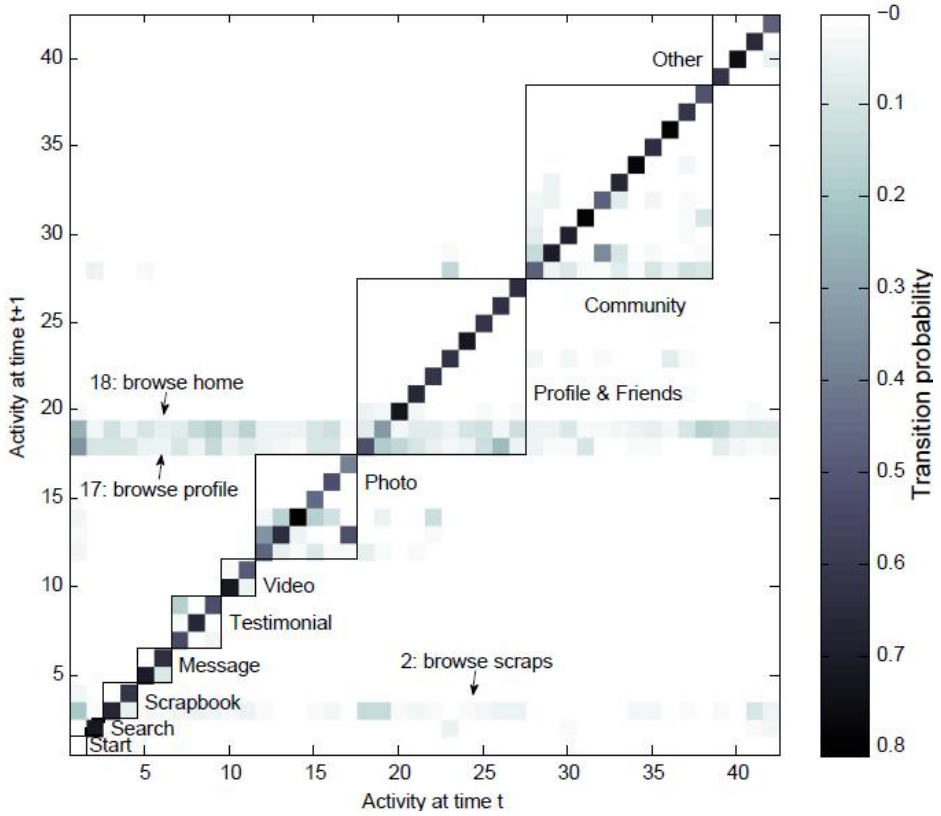


Figure 5.5: Transition probability among activities in the clickstream model for Orkut

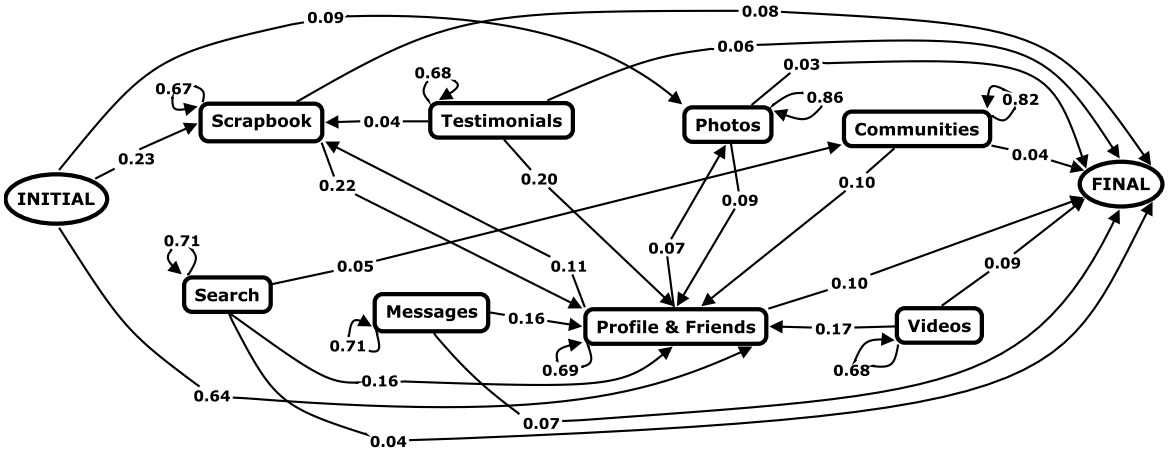


Figure 5.6: Transition probability among categories in the clickstream model for Orkut

5.2.2 Transition from one activity to another

In the second step of modeling, we constructed a first-order Markov chain of user activity based on the sequence of activities seen from all sessions. We added two states, *initial* and *final*, which we appended to the sequence of requests at the beginning and the end of the user sessions, respectively.

Figure 5.5 shows the transition probability between all pairs of activities. A color pixel at (x,y) represents the probability of transition from activity x in the horizontal axis to activity y in the vertical axis. Activity IDs in the figure are identical to the activity IDs in Table 5.1. We also visually show the boundaries for categories. Darker pixels indicate higher transition probability. For visual clarity, probabilities below 0.01 are shown as zero probability in the figure.

When users log in the social network aggregator site, they are immediately exposed to a small selection of updates from all social networks. Users can then click on any of the displayed web objects or the logo of a social network to further browse a given social network. These events are shown as dark pixels on the first column in Figure 5.5. For example, x="Start" and y="browsing homepage" illustrates the case when a user clicked on the logo of a social network and the homepage of the social network was displayed. A typical session started with one of the following activities: browsing scrap, browsing profile, and browsing homepage.

Once a user engaged in a particular activity, the user was likely to repeat the same activity. This is shown by a strong linear trend in $y = x$. For instance, after browsing one photo, a user was likely to immediately browse other photos. In total, 67% of the user activities were repeated.

Next there were more transitions of activities within the same category (77%) than across categories (23%). This means that users typically conduct a sequence of activities that are conceptually related. For instance, a user is likely to browse photos immediately after browsing the list of photo albums, rather than after conducting a less related activity like accessing applications.

We also notice that popular activities like browsing homepage, browsing profiles, and browsing scraps display characteristic horizontal stripes in the graph. This is because every Orkut page embeds hyperlinks to a user's homepage, profile page, and scrapbook page. This suggests that providing a means for users to access a particular feature easily can motivate users to use the given feature frequently.

5.2.3 Transition from one category to another

Finally we examined the sequence of user activities at the level of categories (Figure 5.6). Again we added two synthetic states, Initial and Final, at the beginning and the end of each session. Nodes now represent categories and directed edges represent the transition between two categories. Edges with probability smaller than 4% were removed to reduce the figure complexity. The sum of all outgoing probabilities (including the omitted edges) for each state is 1.0. Compared to Figure 5.5, user behaviors at the category level provide a more holistic view of OSN usage.

We observe that most users initiated their sessions from the Profile & Friends, Scrapbook, or Photos category, as mentioned earlier. We also observe that self loops are present in almost all states. For example, one Community activity was followed by another Community activity with a probability of 0.82. Similarly, Photos activities showed high repetition with a probability of 0.86. Repetition also occurred in Search (probability 0.71). Repetition in Scrapbook was related to users replying to received scraps after browsing them. In Orkut, users can directly reply to an existing (received) scrap from one's own Scrapbook page. We found that 65% of write scrap events (activity 3) immediately followed browsing scrap events (activity 2). Except for self loops, Profile & Friends was the most common preceding state for most activities.

5.2.4 Probability of activity over time

We next investigated whether there is any correlation between the occurrence of a particular activity and session duration. To check for such a correlation, we categorized user sessions into four non-overlapping classes based on their session durations: (a) less than 1 minute, (b) 1 to 10 minutes, (c) 10 to 20 minutes, and (d) 20 minutes or longer. For sessions belonging to each of these intervals, we examined the average proportion of the total session duration that a user spent on each activity.

Figure 5.4 shows the fraction of time spent on each activity as a function of session duration. The results are shown in two separate plots to more easily exhibit the trends for both dominant and subdominant activities. We found two key patterns. First, irrespective of session duration, users spent the most time on Profile & Friends and Scrapbook activities. In very short sessions (i.e., less than 1 minute), users spent 90% of their time on these activities. However, even for a long session (i.e., 20 minutes or longer), the two activities accounted for 75% of the total. Second, the remaining categories of activities became more prevalent for longer sessions. The fraction of time spent consuming media content (i.e., Photos and Videos activities) increased by a factor

of 2 when comparing sessions shorter than 1 minute to those longer than 20 minutes. The probability of seeing Community activity also increased with the session duration.

5.2.5 Comparison of user activity across OSNs

To get perspective on how user behaviors vary across different social networks, we repeated the analysis in Table 5.1 for other social networks that appear in the trace (i.e., MySpace, LinkedIn, and Hi5). All four OSNs exhibited a common pattern in that the most popular activity was browsing profiles. Some activities, however, could only be observed in a subset of these four networks, because the four social networks provided different features to users. For example, MySpace uniquely provided Blogs and News pages and LinkedIn uniquely provided Jobs and Companies pages. Also video and photo features are not supported in LinkedIn.

	Orkut		MySpace		LinkedIn		Hi5	
Rank	Category	Share	Category	Share	Category	Share	Category	Share
1	Profile & Friends	41%	Profile & Friends	88%	Profile & Friends	51%	Profile & Friends	67%
2	Photos	31%	Messages	5%	Other (login)	42%	Photos	18%
3	Scrapbook	20%	Photos	3%	Messages	4%	Comments	6%
4	Communities	4%	Other (login)	3%	Search	2%	Other (login)	4%
5	Search	2%	Communities	1%	Communities	<1%	Messages	3%

Table 5.3: Comparison of popular user activities across four OSN sites

Table 5.3 displays for all four social networks the top five categories based on the number of HTTP requests and the share of corresponding HTTP requests. The statistics are normalized for each social network, so that the sum of share of all activity categories is 100% for each social network.

We make several observations. First, the Profile & Friends category is the most popular across all social networks. Users commonly browsed profiles, homepage, and the list of friends across all four networks.

Second, LinkedIn shows a much lower degree of interaction among users using messages than Orkut. Only 4% of the requests in LinkedIn are related to messaging between users. Because Linkedin is a network used mainly for professional networking (e.g., finding jobs or employees), it is natural to expect that users primarily browse profiles and create links with each other, rather than exchanging messages.

Third, MySpace showed a different profile from Orkut, despite the similarity of its service to that provided by Orkut. MySpace showed a much lower interaction through Photos. A detailed look into the data reveals that 90% of the MySpace users also accessed one of the other three social networks (75% accessed Orkut). Thus, it seems

that users who accessed MySpace using the social network aggregator use Orkut as their primary social network and access MySpace to keep in touch with friends that use only MySpace.

Fourth, the popular user activities in Hi5 were similar to those of Orkut: the most frequent user activity involved browsing friends' updates through Profile & Friends and Photos. The next most popular user activity in both OSNs was a form of message interaction among users: Scrapbook in Orkut and Comments and Messages in Hi5.

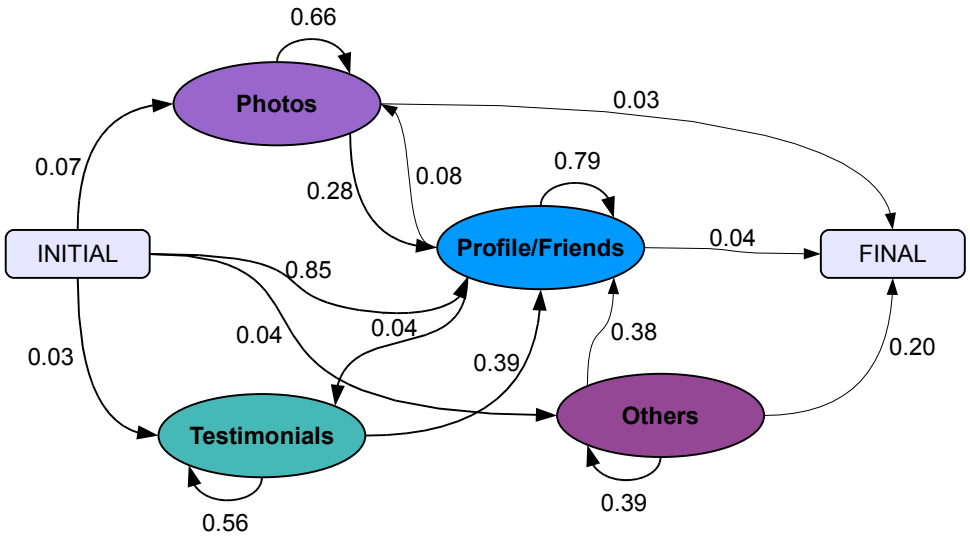


Figure 5.7: Transition probability among categories in the clickstream model for Hi5

We further looked at the first-order Markov chain of user activity in the category level for Hi5, represented in Figure 5.7. Similarly to Orkut, we also observe that self loops are present in almost all states and we also noted that Profile & Friends plays a central role in connecting the other categories of activities. We expect to see similar usage trends for other social networks that possess similar service characteristics to Orkut and Hi5.

5.3 Social interactions in Orkut

One crucial aspect of OSNs is the wide range of features that support communication between users. In this section, we investigate how users interact with each other through the various features OSNs provide, considering the social network distance as well as the physical distance.

5.3.1 Overview

Understanding social interactions has been of great interest in various research fields like sociology, economy, political science, and marketing. Until recently, obtaining large-scale data was one of the key challenges in studying social interactions. Nowadays, we get around this challenge by the wealth of OSN data available on the Internet. A few studies have used publicly crawled OSN data (e.g., comments, testimonials) to characterize social interactions [138; 51; 132; 80; 28]. Although these initial studies have identified several important properties of social interaction, there are behaviors of users that cannot be measured with datasets that contain only visible activity.

One such activity is browsing, which, as demonstrated in the previous section, is one of the most frequent activities in OSNs.[1] As opposed to "visible" interactions that are inferred from crawled data like writing a scrap, browsing a friend's web content can be considered "silent" social interaction. Although visible and silent interactions serve different purposes, both are interesting for understanding the social behaviors of users.

In this section we provide the first look at user interactions in social networks that considers both visible interactions and silent interactions. Our goal is two-fold: (*a*) We would like to know what fraction of user interactions is silent, compared to visible. If we consider browsing a friend's profile or photos as social interaction among users, how much increase would we observe in the number of friends a user typically interacts with? We highlight the potential bias in studies of user interactions using only visible data. (*b*) We are interested in knowing the interaction patterns among users along the social graph distance. In particular, how often do users visit their friends' profiles or even traverse multiple hops to visit the profile of friend of a friend?

[1] Most social networks do not log browsing events of users. However, one exception is Orkut. In Orkut, the list of "recent visitors" to every profile page is shown. Users can also turn this option off and hide their browsing patterns.

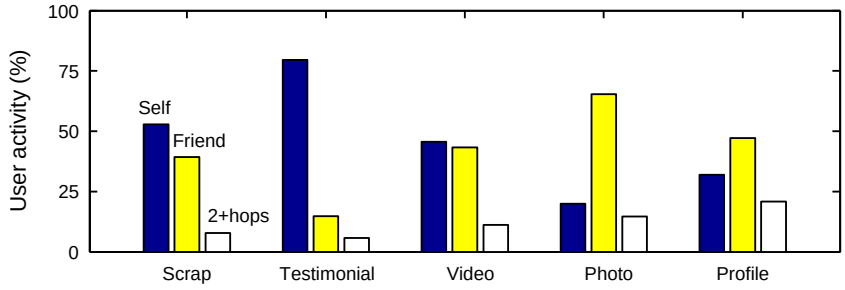


Figure 5.8: Webpage accesses along the social network distance

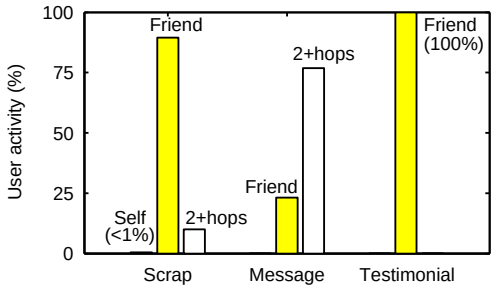


Figure 5.9: Interaction in writing

5.3.2 Interaction over social network distance

We only considered explicitly visiting another user's page to be silent user interaction. It is possible that a user can silently "interact" with a friend by viewing the short list of updates about that friend that are automatically shown on the user's own homepage. However, we do not count these views as interaction, because we cannot be certain whether a user noticed these updates.[2] For example, a user may find a thumbnail of photo update from a friend at her homepage. Only when the user clicks on the photo (thereby visiting the friend's photo page), do we then consider the event as a valid social interaction with a 1-hop friend.

To gain a comprehensive understanding on the social behavior of a user, we needed an essential piece of information: the list of friends of a given user. The clickstream dataset does not include information about the list of friends. Therefore, as described in Section 4.2.1, we gathered information about the list of friends for all users in the workload trace by crawling the public data on Orkut website.

[2]User studies using eye tracking devices will be able to distinguish whether users noticed the exposed content or not.

5.3.2.1 Webpage access patterns

To investigate the patterns of interactions among users, we first examined how often users visit their friends' pages, compared to visiting their own. Not all accesses in the trace were related to interactions among users. Therefore, we focused on the following activities as a form of user interaction: scrapbook, messages, testimonials, videos, photos, and profile & friends. This list comprises activities from 2 to 26 in Table 5.1. We excluded all activities related to search, communities, and others in Table 5.1.

Figure 5.8 shows, for each category of user activity, the fraction of times a user was accessing one's own page (denoted *self* in the figure), a page of an immediate friend (denoted *friend*), or a page of a non-immediate friend (denoted *2+hops*). The result for Messages is omitted, because users can only access their own Messages page. Unless a user has explicitly restricted access, Orkut users can browse any other user's pages containing scrapbook, testimonials, video, photo, and profile. However, the bar chart shows that users mostly accessed pages of their own or their immediate friends; 80% of all accesses remain within a 1-hop neighborhood in the social network topology.

We examined each of the activity categories in detail. Users most frequently accessed their own pages when it comes to scrapbook and testimonials. Yet, users did visit scrapbook and testimonial pages of their 1-hop friends and read what messages are written about their friends. With a small probability, users also visited beyond the 1-hop neighborhood. In total, Orkut users accessed their friends' pages more frequently (59%)[3] than their own pages. When visiting friends' pages, Orkut users not only interacted with immediate friends, but also had significant exposure to non-immediate friends (22%=13/59).

Focusing on each category of interaction, Figure 5.8 shows that users accessed their own video pages as often as they accessed their friends' video pages. On the other hand, in accessing photos, which is a popular activity in Orkut, users were more likely to access their friends' photo pages than their own. Accessing profile pages was well-divided among one's own, immediate friends, and non-immediate friends; 20% of the browsed profiles were 2 or more hops away.

Next we focused on visible interactions and examined which friends users interacted with. We considered the following three visible activities: write scraps (activity 3), write messages (activity 5), and write testimonials (activity 7), because for these activities we could determine the interaction partner from the URL of the trace.

Figure 5.9 shows the division of the times when a user wrote to oneself, a 1-hop

[3]This probability is computed as the count of activities at 1-hop divided by the total occurrences of all activities.

friend, or a 2 or more hop away friend. When using the scrapbook feature, users mostly interacted with immediate friends. Self posts were rare (0.5%), but could serve as a broadcast message to everyone who visits the scrapbook. Interestingly, 10% of scraps were sent to users that are 2 or more hops away. On the other hand, users did not interact much with immediate friends through Messages. Instead, we observed frequent interactions with non-immediate friends through Messages (76%). Testimonials were only sent to immediate friends, as written in the Orkut policy. We discuss the implications of these findings in the following section.

5.3.2.2 What leads users to visit other people's pages?

Having studied the frequency at which users access their friends' pages, we now take a closer look at how a user navigates from one friend's page to another. Particularly, we are interested in understanding what activities lead users to visit a page of a friend or a non-friend. We performed the following analysis. Each time a user visited a page of a friend, we examined which preceding page the user was at: one's own page, an immediate friend's page, or a non-immediate friend's page? Table 5.4 shows the fraction of preceding locations for every first access to a friend's page in each session. In addition to the navigation statistics, Table 5.4 also shows the list of top activities that preceded the navigation event.

Current location (First access to)	Preceding location: 0-hop		1-hop		2 or more hops	
Immediate friend's page (1-hop)	Total 68%		Total 25%		Total 7%	
	◦ Browse homepage	(36%)	◦ Browse profiles	(8%)	◦ Browse profiles	(4%)
	◦ Browse scraps	(12%)	◦ Browse scraps	(6%)	◦ Browse scraps	(1%)
	◦ Browse friends list	(9%)	◦ Browse photos	(3%)	◦ Browse photos	(<1%)
Non-immediate friend's page (2 or more hops)	Total 37%		Total 30%		Total 33%	
	◦ Browse homepage	(22%)	◦ Browse profiles	(9%)	◦ Browse profiles	(15%)
	◦ Browse scraps	(6%)	◦ Browse scraps	(9%)	◦ Browse friends list	(5%)
	◦ Browse profiles	(3%)	◦ Browse friends list	(5%)	◦ Browse scraps	(5%)

Table 5.4: How users arrive at other people's pages: preceding locations and activities for every first visit to an immediate and non-immediate friend's page

The majority of accesses (68%) to an immediate friend's Web page originated from browsing one's own Web page (the first row of Table 5.1). The remaining accesses occurred when the user was navigating the social network; accesses to an immediate friend's Web page were followed by browsing of another immediate friend's Web page (25%) or browsing of a non-immediate friend's Web page (7%). When it comes to visiting a non-immediate friend's Web page (the second row of Table 5.1), the preceding location of the user was well distributed across 0-hop, 1-hop, and 2 or more hops.

Interestingly, the most popular activity that leads a user to an immediate or non-immediate friend's Web page is browsing one's own homepage. A user's homepage contains a short list of updates from friends as well as a list of the subset of friends who recently logged in. Such updates can contain links to non-immediate friends when they interacted with mutual friends through photo comments, testimonials, or applications. Therefore, updates from friends can also drive users to visit the Web page of a friend of a friend.

Another interesting observation we make is the high fraction of accesses that originated from an immediate friend's Web page, which accounted for 25% of the accesses to another immediate friend's Web page and 30% of the accesses to a non-immediate friend's Web page (the third column of Table 5.1). This reinforces the previous findings that users in social networks find new content and contacts through their 1-hop friends [43; 122; 44]. Browsing an immediate friend's profile was the most common gateway that led users from one friend to another.

Lastly, we note that browsing scraps (activity 2) appears in the top three activities in all the rows of Table 5.4. This may mean that Orkut users are keen on reading other users' scrapbook content and also are curious about checking out new contacts that they encounter through such activity.

5.3.3 Interaction over physical distance

So, to what extend interactions are related to physical proximity between users in OSNs? In order to answer this question we need to know the location of users involved in the interaction. In our dataset, we can use the IP address to identify the location of users that sent requests to the social network aggregator. However, we cannot identify the users that are in the receiving part of the interactions (e.g. users that received a scrap, users that had a photo browsed, etc.). Thus, in order to obtain the location of these users we further crawled the Orkut profile information of these users.

The location information available in user profiles is in free text form and often contained invalid location like "Mars." We used the Yahoo Maps Web service Geocoding API to filter out invalid locations and infer user locations. In this way, we identified the location of 276,558 users, of which we know the location at the city level for 128,836 users and at the country level for 276,558 of the users. Users were located in 4,297 different cities across 226 countries.

Figure 5.10 shows how the probability of interaction varies as a function of the physical distance between two users. Physical distance between users is computed based on their longitude and latitude. We grouped distances in units of 10km, so that

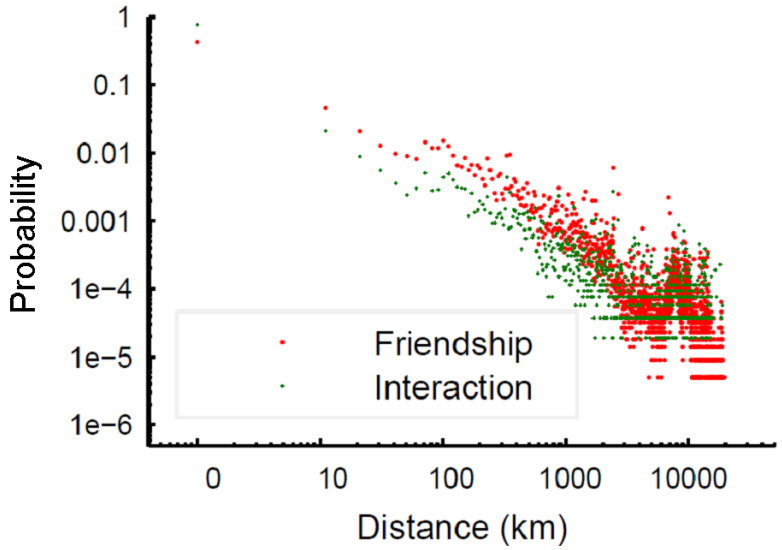


Figure 5.10: Interaction across physical distance

0km means a distance between (0,10km]. The graph shows the probability for each distance d, which is the physical distance among all pairs u, v of users who interactered in our dataset. For comparison, we also show the probability of friendship over physical distance between users.

We can see that there is significant correlation between interaction and physical proximity in the Orkut social network. Although Orkut works as a means to bridge the distance between friends, two users within a short distance (e.g. 10 km) have a high probability of interacting. Current OSNs infrastructure could exploit the physical proximity between content producers and consumers. We also observe a strong correlation between the probability of interaction and the probability of forming friendship links. This is expected as users tend to interact more with their friends in the social graph. Although this could be a result of the OSN's interface design (e.g. updates from a user are pushed only to the user's friends), even when there are no such limitations, the very nature of social content (e.g. personal pictures) might make it appealing to a small set of people. For example, in Orkut users can access any other user's profiles by default. However, our data show that users mostly accessed profiles of their immediate 1-hop friends in the social network. This suggests that users in the social network tend to be geographically closer to each other when the interaction occurs mainly due to the presence of social links.

The observations that social content is primarily of interest to friends of the uploader and friends are usually located in a close physical distance could be exploited

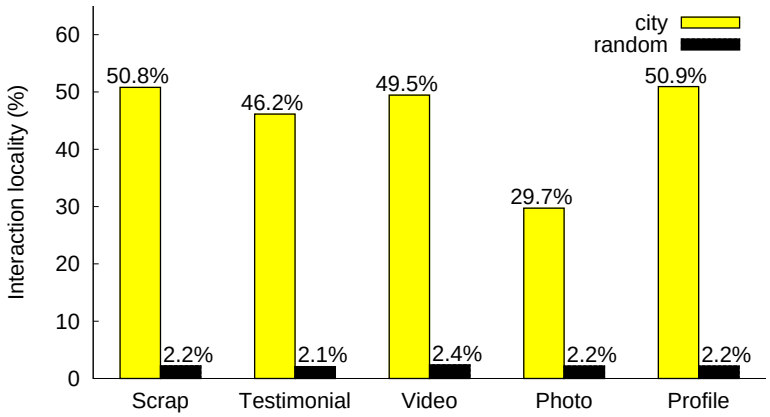


Figure 5.11: Locality between content producers and consumers

for caching design and content delivery networks (CDNs). Thus, we next explore the characteristics of locality of content producers and content consumers. Figure 5.11 shows the probability that content producers and consumers are located in the same city. In order to test if these results are affected by a biased sample of users, we randomly picked 500,000 pairs of users and computed the cities they are located. The results are shown as "random" in the figure. We can see that the fraction of interaction between users located in the same city is much smaller for the randomized set of pairs of users, for all the types of interaction analyzed. Consequently, in contrast to geographically-diverse content uploads, our findings indicate that social content is consumed *locally* in Orkut.

5.3.4 Number of friends interacted with

Finally we investigated how silent interactions affect the level of user interactions along the social network topology. We compare the number of friends (including multi-hop friends) a user interacts with through all activities with the number interacted with through only visible activities, as a function of the number of friends in the social graph.

Figure 5.12 shows these quantities. Overall, the degree of interaction is very low; the average user interacted (whether visibly or silently) with 3.2 friends in total over the 12-day period and interacted visibly with only 0.2 friends. This low level of interaction has also been observed in other work. According to Wilson et al. [138], in the Facebook social network nearly 60% of users exhibit no interaction at all over an entire year. Therefore, our workload trace of 12 days is expected to show a much lower level of interaction.

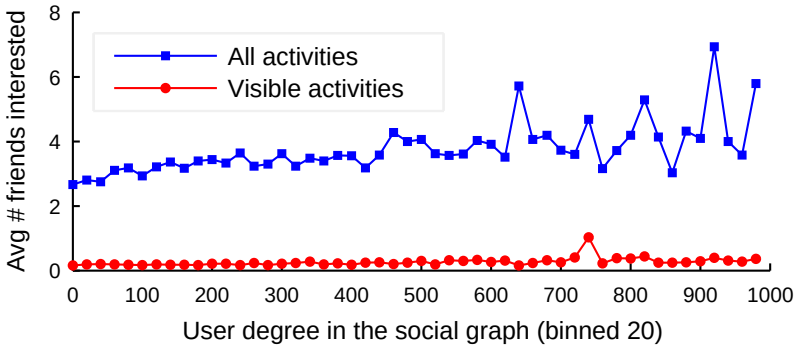


Figure 5.12: Comparison of the Orkut social graph degree and interaction degree

Interestingly, even for a short trace period, the degree of all interaction is 16 times or an order of magnitude greater than the degree of visible interaction. The stark difference in the two quantities may be because in OSN usage the majority of time is spent browsing, which cannot be captured by visible interactions. Another trend that we observe is that interaction degree does not grow rapidly with the user degree in the social graph; users with low degree interacted with a similar number of friends as users with high degree. This indicates that it is easier to form friend links than to actually interact with those friends.

In total, 55% of users in the workload trace interacted with at least one other user during the 12-day period; 8% showed at least one visible interaction and 47% showed only silent interactions. Thus, if one were to measure the strength of social ties based only on visible traces, such analysis would be biased because 85% (=47/55) of the users would be completely disregarded.

In summary our analysis of social interaction in this section brought out many interesting findings. When we consider silent interactions like browsing friends' pages, the measured interaction among users significantly increased, compared to only considering visible interactions. Furthermore, we showed that if one were to measure the strength of social ties based on visible traces, 85% of the users would be disregarded.

5.4 Discussion

In this chapter we presented a thorough characterization of social network workloads, based on detailed clickstream data summarizing HTTP sessions over a 12-day period of 37,024 users. The data were collected from a social network aggregator website, which after a single authentication enables users to connect to multiple social networks: Orkut, MySpace, Hi5, and LinkedIn. We analyzed the statistical and distributional

properties of most of the important variables of OSN sessions. We presented the clickstream model to characterize user behavior in online social networks and also explore how users interact using the various features that OSNs provide, as a function of the social network distance as well as the physical distance.

Our study uncovered a number of interesting findings, some of which are related to the specific nature of social networking environments. Next, we summarize our findings and discuss their implications.

Modeling of OSN sessions: In Section 5.1, we characterized the properties of individual session properties in OSN workloads. Among various findings, here we highlight that session quantities like inter-session times, session lengths, and inter-request times follow a heavy-tailed distribution. For example, the majority of the sessions remain short (on the order of tens of minutes), but some sessions last several hours to days. As a result of the asymmetry of the distribution, user behaviors cannot be represented as a normal distribution with comparable mean and variance. Also the typical behavior of users will not be the same as their average behavior [79].

To incorporate the large variation in user behaviors, we provided statistics for the average behavior as well as the best fit distribution functions that capture this asymmetry. Such distribution functions can be used to generate synthetic (parameterizable) traces, that mimic actual OSN workloads. We hope that the statistics summarized in the paper and the session modeling will be valuable in evaluating and testing potential OSN services.

Understanding user activity in OSNs: In Section 5.2, we characterized the type, frequency, and sequence of user activities in OSNs. Using clickstream data, we presented a complete profile of user activity in Orkut in Table 5.1 and Figures 5.5 and 5.6. Our analysis demonstrated that browsing, which cannot be identified from visible data, is the most dominant behavior (92%).

We believe that understanding user navigation is important for OSN service providers and portals [138; 39] as well as for advertising agencies [B. Williamson]. This is because frequently repeated activities (e.g., browsing home, browsing scraps) naturally serve as good targets for advertisements and the sequence of activities can be analyzed to improve the website design. Furthermore, the fact that users frequently check updates from their social contacts without involving any other action naturally makes OSNs a great place for advertisement.

Another application of our analysis is that an OSN service provider may consider providing a personalized web interface for users based on the users' activity profiles.

For example, a user login page can be reorganized so that frequently repeated activities are more easily accessible. OSN service providers may also use aggregate patterns in clickstreams to identify users with similar behaviors (e.g., belonging to the same communities, possessing similar profile description) and recommend popular content within the site.

Interaction over the social graph: In Section 5.3 we used both the clickstream data and the social graph topology to study how users interact with friends in OSNs. Among various findings, we observed that Orkut users not only interact with 1-hop friends, but also have substantial exposure to friends that are 2 or more hops away (22%). This exposure to friends' pages has significant implication for information propagation in OSNs: OSNs exhibit "small-world" properties [105; 12; 138], which means that the network structure has a potential to spread information quickly and widely. Our observation highlighted that users actively visiting immediate and non-immediate friends' pages serves as an empirical precondition for word-of-mouth-based information propagation.

Interaction over the physical distance: In Section 5.3 we also studied the physical distance between users as well as the distance between content producers and consumers. Our results suggest that users in a social network tend to be geographically closer to each other when the interaction occurs mainly due to the presence of social links. Additionally, we showed that social content is mostly consumed *locally* in Orkut.

These results call for a reexamination of today's content distribution infrastructures, which does not exploit the physical proximity between content producers and consumers in OSNs. Considering that content in social workloads is typically produced by geographically-diverse users but consumed locally, one could allow users to upload content to a local server in the corresponding geographical area, like a city. Such mechanism could significantly reduce the amount of wide-area bandwidth needed compared to an upload to a centralized, remote server. The local server can then handle requests coming from users in the same geographical area and content needs to leave the local area only when a remote user requests it. However, this is expected to happen infrequently. In particular, high locality in content access at the city level in our dataset indicates that placing a server in every city can greatly reduce the amount of cross-city traffic. The exact benefit depends on the size of a city, with more bandwidth savings attained in larger urban areas.

Chapter 6

Video Interactions

A recent discussion on the needs and challenges of multimedia research in the context of Web 2.0 pointed out that understanding how users typically interact in OSNs is of great relevance as users play an important role in the social network system [32]. In this chapter we focus on providing a large and representative characterization of users interacting with each other essentially via video objects. Particularly, we characterize the use of the YouTube video response feature.

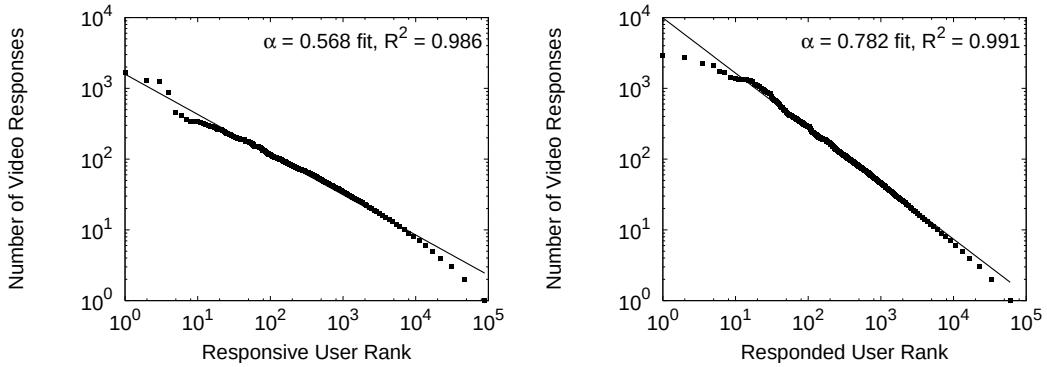
Video responses have been largely used as a mean to exchange knowledge and express ideas and opinions, allowing users to discuss themes and to provide review for products or places using a much richer media rather than simple text. Video responses were also widely used during USA presidential detabes. Promoted by CNN and YouTube, several individuals send their questions to candidates, that answered live on CNN. The answers of the candidates were displayed on YouTube so that YouTube users could post their video responses with their own opinions about the candidates answer [you]. In addition to provide valuable statistical models for various characteristics, our study uncovers typical user behavioral patterns as well as show evidences of opportunistic behavior, corresponding to some kind of interaction and interface pollution.

Our analysis of video interactions relies on a large dataset of videos and users collected from YouTube, described in Section 4.2.2. Next, Section 6.1 presents statistics of video responses. In the following two sections we present a characterization of video interactions at two levels of granularity, the interactions engendered by a single responded video and interactions created by communication in multiple topics. Section 6.4 discusses opportunistic behavior and unwanted information in video interactions. Finally, Section 6.5 concludes the chapter and discuss implications of our findings.

6.1 Video response statistics

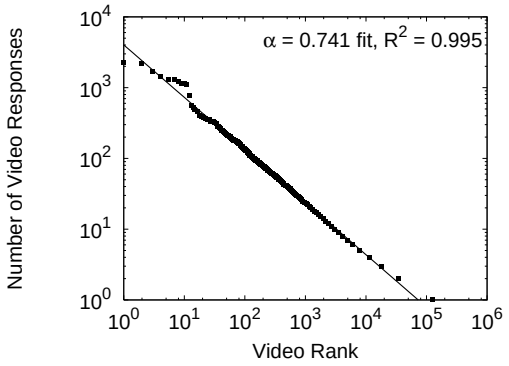
This section presents a characterization of several aspects of the video responses and responded videos collected.

6.1.1 Posting characteristics



(a) Number of video responses per responsive user

(b) Number of video responses per responded user



(c) Number of video responses per responded video

Figure 6.1: Ranking of videos and users in terms of number of video responses received or posted

Experimental data have been used in the literature to show that the behavior of many real-world systems can be modeled by a power law distribution [Barford and Crovella; 16; 36]. In terms of video interactions, we are interested in verifying the existence of the power law phenomenon to provide answers for three questions: (1) *Does a small fraction of responsive users account for posting the majority of the video responses?* (2) *Does a small fraction of responded users receive the majority of the video responses?* (3) *Does a small fraction of responded videos receive the majority of the video responses?*

Figure 6.1(a) shows the distribution of the number of video responses posted by different users. We note that the top 20% most responsive users contributed with 65% of all video responses, whereas 84% of all responsive users posted, each, less than 5 video responses. In fact, the distribution is well fitted by a power-law distribution (Prob(user with rank i posts a video response) $\propto 1/i^{\alpha}$) with $\alpha = 0.568$. Similarly, Figures 6.1(b) and 6.1(c) show the distributions of the numbers of video responses per responded user and responded video, respectively. Both distributions are also well fitted by power-laws with $\alpha = 0.782$ e $\alpha = 0.741$, respectively. To assess the accuracy of our proposed models, we measure the R^2 factor of the linear regression [130] for each single distribution found. In our models, the values of R^2 are above 0.91 in all cases, which show a good agreement ($R^2 = 1$ means perfect agreement).

In comparison with textual interactions, prior research showed that ranking blogs in terms of the number of comments received is also fitted by a power-law distribution with $\alpha = 0.7$ [57].

6.1.2 Video response categorization

It is natural to expect that video responses posted to a video topic are related and approach the same subject. Next, we study the video categories of our collected data. The definition of a category is done by the users, contributors of the posted video, and the categories are chosen among 12 pre-defined subjects provided by YouTube.

Table 6.1 shows the distribution of video responses across categories. Each line refers to *responded* videos falling into one category, and reports the average fractions of their video responses falling in each category. The concentration of responses into different topics varies depending on the category of the responded video. For instance, most of the responses posted to “Gadgets & Games" videos are of the same category, whereas “Travel & Places" videos receive responses from different categories, with emphasis on “People & Blogs" and “Entertainment". Nevertheless, in most cases, the majority of the responses fall into the same category of the responded video. In other words, the responded video and most of its responses create a large group around the same subject. However, this subject is not strong enough to categorize all interactions triggered by a video, as a non-negligible fraction of the responses (24% - 61%) falls outside the responded video category, creating multiple variable-size groups around each other category.

Since users can freely assign pre-defined categories to their uploaded videos, we can not assume that categories correctly describe the subject of the video content. However, some of the video responses might really be unrelated to the responded video.

Category	Com	N&P	Ent	Mus	G&G	A&V	P&A	T&P	P&B	F&A	H&D	Spo
Comedy	**54.6**	2.8	10.1	6.6	1.8	0.5	1.4	0.8	13.8	5.5	1.1	0.9
News & Politics	5.3	**66.0**	4.4	4.0	0.4	0.4	0.6	1.2	13.5	2.1	1.7	0.5
Entertainment	8.8	2.1	**49.2**	10.1	3.9	0.5	0.8	1.1	10.4	10.3	1.1	1.5
Music	4.0	1.0	7.8	**72.6**	1.2	0.3	0.4	0.6	6.3	4.9	0.6	0.4
Gadgets & Games	2.8	0.3	6.0	3.3	**76.4**	0.4	0.1	0.2	4.3	5.2	0.7	0.4
Autos & Vehicles	4.1	2.9	6.8	1.9	4.3	**58.8**	0.4	4.3	6.1	5.7	2.2	2.5
Pets & Animals	6.4	1.9	4.3	2.9	0.4	0.2	**73.0**	1.4	6.1	1.9	1.0	0.6
Travel & Places	6.3	4.1	14.4	6.8	3.2	2.3	1.6	**38.9**	12.5	5.8	1.8	2.4
People & Blogs	5.6	4.6	6.8	5.3	1.1	0.4	0.9	1.3	**67.6**	3.3	2.4	0.8
Film & Animation	6.6	1.1	12.0	7.6	3.6	0.5	0.5	0.7	5.0	**61.2**	0.8	0.4
Howto & DIY	6.3	4.8	8.7	4.9	2.9	1.3	1.3	1.8	14.2	6.0	**46.6**	1.3
Sports	4.6	1.1	8.8	2.7	1.1	1.1	0.8	1.2	6.4	1.9	0.8	**69.4**

Table 6.1: Video responses across categories

This, in turn, may be a result of some sort of opportunistic behavior, such as spamming (see Section 6.4). Thus, information services such as advertising and recommendation should not take for grant that a video topic is directly related to its video responses.

6.1.3 Video response duration

Except for directors and other kinds of special accounts, YouTube users can post videos with a maximal duration of 10 minutes. Figure 6.2 (left) shows the distributions of video durations, considering, separately, only responded videos and only video responses. Both distributions are very skewed, with 80% of all samples being under 5 minutes. In fact, both follow Weibull distributions[1], with parameters $\alpha = 0.0023$ and $\beta = 1.15$, for video responses, and $\alpha = 0.00054$ and $\beta = 1.35$ for responded videos. However, video responses have durations slightly more skewed towards shorter values.

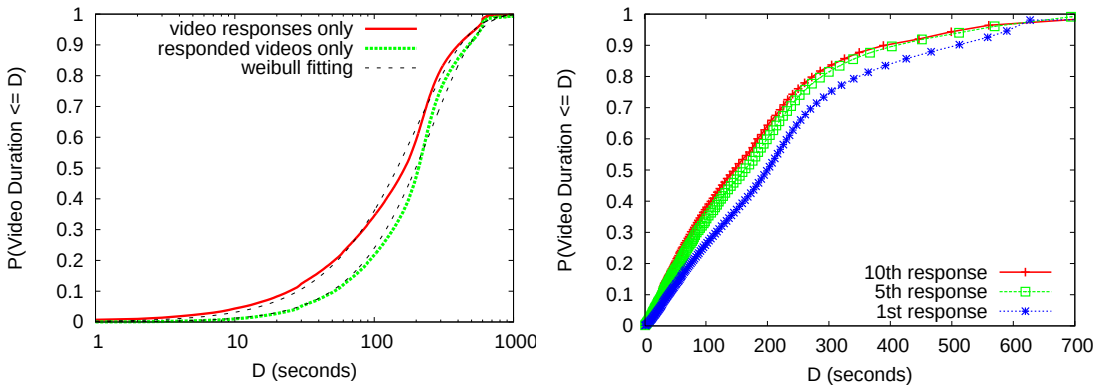


Figure 6.2: Duration of responded videos and video responses (left) and correlation of the video topic duration and its video responses (right)

[1]The Probability Density Function of the Weibull distribution is given by: $f(x) = (\beta/\alpha)^{-\beta}(x/\beta)^{(\beta-1)}e^{(-x/\alpha)^{\beta}}$.

Interestingly, we found that, although the duration of a responded video has practically no impact on the number of responses it receives (correlation coefficient $C = -0.008$), there is a strong correlation between the duration of the responded video and the average duration of its responses ($C = 0.51$). Longer video topics tend to receive longer responses, as well as shorter video topics tend to receive shorter responses. Moreover, Figure 6.2 (right) shows that, considering only the i^{th} responses of videos that had at least i responses, the duration distribution becomes more skewed as i increases. These results mimic the expected pattern in real-life human interactions, whereas longer (but interesting) expositions tend to trigger longer replies, initially. However, replies tend to become shorter as the interaction progresses, and the discussion dies down.

6.1.4 Video response geographical distribution

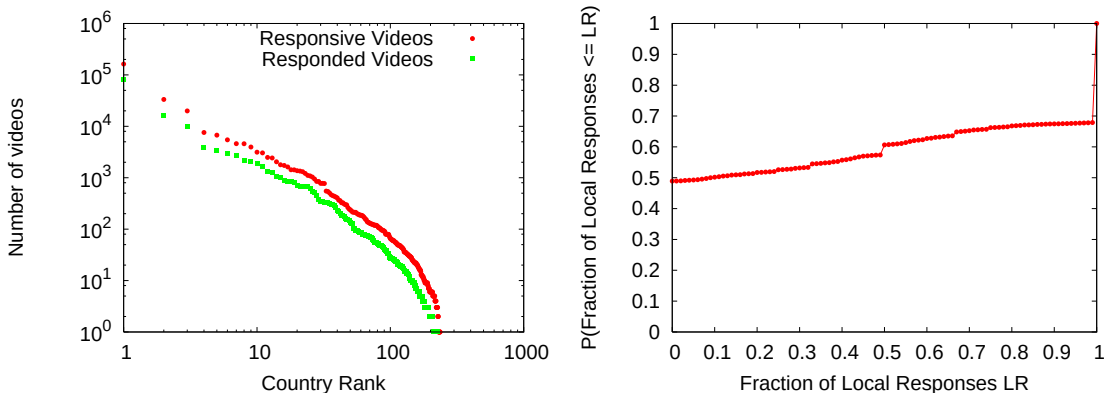


Figure 6.3: Distribution of video responses over countries (left) and distribution of the percentage of local video responses (right)

An important question that is often asked regarding Web characterization studies has to do with the geographical representativeness of the sample. Location and friendship have been used to build real social network models [99]. Figure 6.3 (left) shows the distribution of countries in terms of number of videos shared in YouTube. The sample characterized is fairly diverse in terms of the number of countries, identified by the country described in the user profile, and of the numbers of video responses and of responded videos uploaded by users from different countries. Using the country identification, we are able to map the contributor population to over 236 countries. The top five countries in our sample account for almost 77% of all video responses uploaded to YouTube. The plots suggest a power law-like profile, with parameter $\alpha = 2.12$ ($R^2 = 0.92$) and $\alpha = 2.22$ ($R^2 = 0.93$) for video responses and responded videos, respectively. We also defined the percentage of local (i.e. from the same country) video

responses as the ratio of video responses from the same country of the owner of the responded video to the total number of video responses. Figure 6.3 (right) shows the cumulative distribution of the percentage of local video responses over all responded videos of our sample. We notice that slightly more than half of the "video conversations" involves participants from the same country of the original contributor. For example, 40% of responded video has a percentage of local responses larger than 60%. This observation may be useful for designing geographical distributed video placement mechanisms.

6.1.5 Self-responses

Interestingly, 25% of all video responses are self-responses, that is, responses posted by the user who posted the original video. Figure 6.4 shows the cumulative distribution of the fraction of self-responses posted to each responded video. Roughly 35% of the responded videos received at least one self-response, and around 12% of them received *only* self-responses.

Such phenomenon was also observed in textual communication in the Cyworld's guestbook, where 39% of all posts are self-posts [51]. In a guestbook, a self-post may have two purposes, namely, a reply to a previous message and a note, i.e., a message to other users who access the guestbook. In the YouTube context, we conjecture that there are, at least, two purposes for self-responses in video interactions. Whereas some of these responses might actually be replies to other responses, others might be an attempt at self-promotion, that is, an attempt to inflate the number of video responses posted to a video to place it in a top position of the most-responded lists provided by YouTube, thus, gaining visibility in the system.

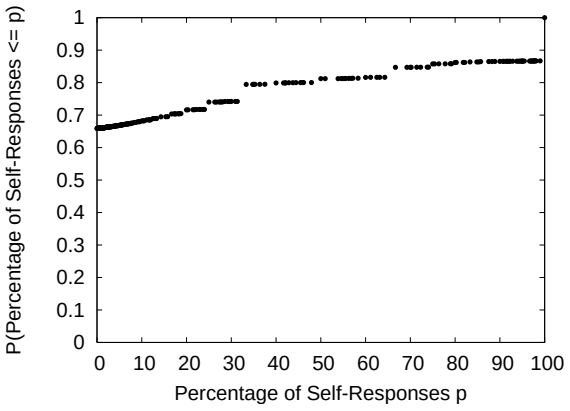


Figure 6.4: Fraction of self-responses per responded video

Therefore, a question that arises is whether a user can exploit the video response feature to raise the popularity of a video topic, i.e., *is a video that receives many responses also viewed many times?* The correlation between the number of responses and the number of views of a responded video shows this is not often the case (correlation coefficient $C = 0.16$). If we disregard all responded videos with at least one self-response, the correlation increases somewhat ($C = 0.24$), but remains low. These low correlation values indicate that one is not necessarily successful in artificially increasing the popularity of a video by simply posting video responses to it. In other words, posting responses aiming at promotion does not necessarily pay off in YouTube.

6.1.6 Video response interval

A user may post a video response in one of three ways: 1) directly from the user's webcam; 2) choosing a video from one of the user's own, pre-existing YouTube contributions; 3) uploading a video from the user's disk drive. Unfortunately, YouTube does not provide a means to automatically determine in which manner a video was created. We thus propose to categorize video responses based on the time it was uploaded to YouTube, relative to the upload time of the responded video. We define the Video-Response-Interval (VRI) as the upload time of the response minus the upload time of the responded video.

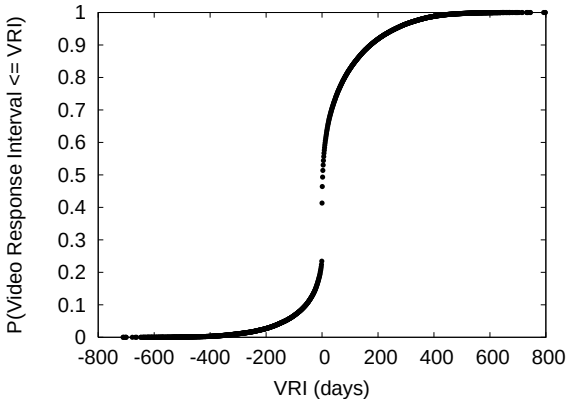


Figure 6.5: Video response intervals (VRIs)

The cumulative distribution of the VRIs is shown in Figure 6.5. We can note that about 27% of the video responses correspond to videos uploaded *before* the responded video, and thus, were certainly not *created* as responses to it. This might be explained by the video content itself. For instance, we observed that one of the responded videos having many previously uploaded responses explicitly actually re-

quested existing YouTube videos for responses. On the other hand, it might also be the result of users uploading existing (and not necessarily related) videos as self-responses. On the other hand, there are videos that clearly motivate users to post their opinions, creating long discussion threads. We found that 42% of the video responses are added within the first month after the video topic was uploaded, indicating a prompt reaction from responsive users. Nevertheless, a non-negligible fraction (17%) of responses are added long after the video appeared in the system (i.e., VRI $\geq$ 100 days), meaning that some videos exhibit long-term popularity, and new interactions are initiated long after it was uploaded.

6.2 Interactions engendered by a single responded video

In this section we study user interactions engendered by a single responded video. Our characterization consists of examining the sequence of video responses following a YouTube video topic. For each responded video V_i, let n_i denote the number of video responses of V_i, and let us denote the sequence of V_i's video responses by $\{VR_{i,1}; VR_{i,2}; ...; VR_{i,j}; ...; VR_{i,n_i}\}$, where $VR_{i,j}$ is the j^{th} video response of video V_i.

A user may add multiple video responses in sequence to the same video. We define a sequence of responses $S_{i,k}$ as a series of *consecutive* responses from the same user to video V_i. Thus, the ordered list of video responses to video i can be also expressed as $\{S_{i,1}; S_{i,2}; ...S_{i,s_i}\}$, $1 \leq s_i \leq n_i$, where $S_{i,k}$ is the sequence $\{VR_{i,j}; VR_{i,j+1}...\}$ of *consecutive* responses from user U_k to video V_i. Note that, the same user may post multiple (non-consecutive) sequences of responses to the same video V_i.

In the following example, video V_i received 7 video responses grouped into 4 sequences, posted by 3 different users (U_1, U_2 and U_3):

$$V_i; \overbrace{\underbrace{VR_{i,1}, VR_{i,2}}_{U_1}}^{S_{i,1}}, \overbrace{\underbrace{VR_{i,3}}_{U_2}}^{S_{i,2}}, \overbrace{\underbrace{VR_{i,4}, VR_{i,5}, VR_{i,6}}_{U_1}}^{S_{i,3}}, \overbrace{\underbrace{VR_{i,7}}_{U_3}}^{S_{i,4}}$$

Our characterization focuses on a simple metric, defined as the ratio U-S of the *Number of Unique Responsive Users* to the *Number of Sequences of Responses*. In the above example, this ratio is 3/4. A video-based interaction with a ratio U-S close to 0 indicates an asynchronous video dialogue between a relatively small number of highly active users, who keep the discussion alive with multiple (not necessarily consecutive) responses to each other. This type of interaction is akin to the exchanges and debates

in a parlor or public forum, in which the communication underscores a many-to-many dialogue between participants. At the other extreme, when the ratio approaches 1, there may be two types of interaction. One type occurs when the number of unique responsive users equals the number of sequences of responses, resembling a register, petition or guest-book, for which the communication is many-to-one, and the purpose of a video response is to record a comment (or support a petition, etc.). The other type has just one user posting a single sequence. Single-user sequences containing multiple responses are typically an attempt at promotion.

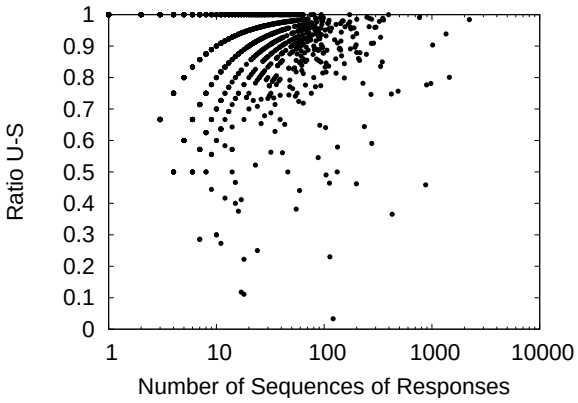


Figure 6.6: Types of video-based interactions

Figure 6.6 shows the ratio U-S versus the number of sequences of responses for each responded video in our data set. Different types of interactions can be seen across all responded videos, characterized by a wide range of U-S values. However, only a relatively small number of videos triggered lively interactions, with each responsive user adding on average at least 3 sequences of responses. In fact, the vast majority (99%) of the responded videos triggered interactions with only one sequence of responses per responsive user (i.e., ratio equal to 1). In fact, Figure 6.6 shows that these interactions occur among groups of responsive users of varying sizes (i.e., number of sequences). We further analyze this issue by looking into the total number of responses included in each sequence. We found that only 6% of all sequences have more than one video response. We also found a few cases in which the responded video received a single sequence with multiple responses: 714 videos received a single sequence with at least 5 responses from the same user, whereas 146 videos had single sequences with 10 or more responses from the same user. Finally, we found that 13% of all video topics with more than two video responses received only self-responses.

6.3 Social aspects of interactions

Social networks are useful for analyzing social phenomena that involve the interactions of a large number of heterogeneous entities, such as videos, contributors, and viewers. In this section, we study the video response user graph (A, B), defined in Section 4.2.2. Table 6.2 presents the main statistics of the network built from the graph (A, B) and its largest strongly connected component (SCC).

Characteristic	Dataset	Largest SCC
Number of Nodes	160,074	7,776
Number of Edges	244,040	33,682
Avg Clustering Coefficient	0.047	0.137
Number of nodes of largest SCC	7,776	7,776
Number of components	149,779	1
r	-0.017	0.017
Avg Distance	8.40	8.40
Avg k_{in} (CV)	1.53 (9.38)	4.33 (3.14)
Avg k_{out} (CV)	1.53 (1.717)	4.33 (1.28)

Table 6.2: Summary of the network metrics

6.3.1 Node degree

The key characteristics of the structure of a directed network are the in-degree (k_{in}) and the out-degree (k_{out}) distributions. As shown in Figure 6.7, the distributions of the degrees for the entire graph follow power laws $P(k_{in/out}) \propto 1/k_{in/out}^{\alpha^{in/out}}$. The scaling exponents of the whole network lie in a range of 2.0 and 3.4, which is a very common range for social and communication networks [59]. Our results agree with previous measurements of many real-world networks that exhibit power law distributions, including the Web and social networks defined by blog-to-blog links [91] as well as the discussions threads in Slashdot [72].

The in-degree exponent is smaller than the exponent of the out-degree distribution, indicating that there are more users with larger in-degree than out-degree. This fact suggests a link assymmetry in the directed interaction network. Unlike other social networks that exhibit a significant degree of symmetry [105], the user interaction network shows a structure similar to the Web graph, where pages with high in-degree tend to be authorities and pages with high out-degree act as hubs directing users to recommended pages [87]. In order to investigate this point further, Figure 6.8 shows the cumulative distribution of the ratios of in-degree to out-degree and out-degree to

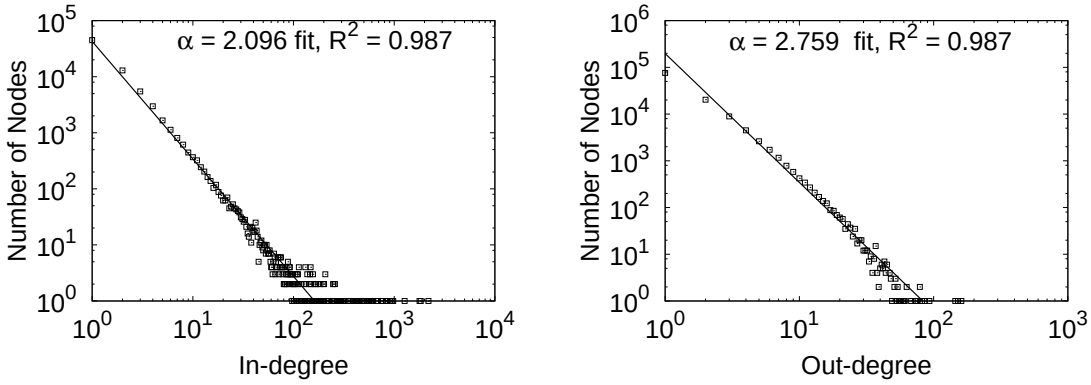


Figure 6.7: In-degree and out-degree distributions

in-degree for the user interaction network. The network has 60% of the users with out-degree higher than in-degree and 5% of the users with significantly higher in-degree than out-degree (ratio > 3). This is evidence that a few users act as "authorities" and "hubs". We have observed in our dataset that authority-like users (that is, highly responded users), with high in-degree, are typically media companies that upload professional content, including sports, entertainment video and TV series. Nodes with very high out-degree may indicate either very active users or spammers, i.e., users that distribute unsolicited content to legitimate users.

According to [116], assortative mixing is a graph theoretical quantity, typical of social networks. We then investigate this structural propertie in the user interaction network. A network is said to exhibit assortative mixing if the nodes with many connections tend to be connected to other nodes with many connections. Social networks usually show assortative mixing. The assortative (or dissortative) mixing of a network is evaluated by the Pearson coefficient r, as described in equation 2.4.1 (see Section 2.4.1). Positive values of r indicate that the network exhibits assortativity, i.e. nodes with high degrees tend to be connected to nodes with high degrees. On the other hand, negative values of r indicate that the network exhibits a dissortative mixing, i.e. nodes with high degrees tend to be connected to nodes with low degrees.

Table 6.2 shows values of r for the directed graph of the interaction network. The video response user graph exhibits a dissortative mixing with $r = -0.017$, where high degree nodes preferentially connect with low degree ones and vice versa. Analyzing only the largest SCC, we can observe an assortative mixing $r = 0.017$. The existence of a significant assortative mixing is associated with the notion of social communities [116]. Thus, whereas the entire user interaction graph does not show evidence of formation of a social community, its largest SCC exhibits some sign (though weak) of such behavior.

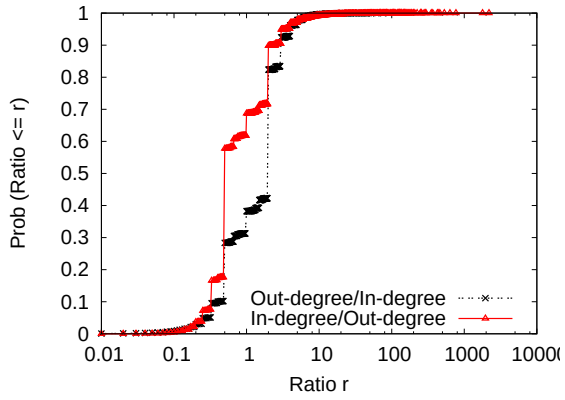


Figure 6.8: In-degree and out-degree ratio

6.3.2 Clustering analysis

It has been suggested in the literature that social networks possess a topological structure where nodes are organized into communities [116], a feature that can account for large values of clustering coefficient and degree correlation. The clustering coefficient of a node i, $cc(i)$ is the ratio of the number of existing edges between i's neighbors over the number of all possible edges between them.

The clustering coefficient of a network, CC, is the mean clustering coefficient of all nodes. The average CC of all nodes of our network is $CC = 0.047$, whereas the mean clustering coefficient for a random graph with identical degree distribution and number of nodes but random links is $CC = 0.007$. Thus, our network presents a significantly larger clustering coefficient than a randomized version of the graph, which points to the existence of some structural hierarchy and the presence of small communities in the graph.

Figure 6.9 (left) shows the cumulative distribution of the node clustering coefficient. The network contains a significant fraction of nodes with clustering coefficient equals to zero. Specifically, 80% of all nodes in the entire user interaction network have $CC = 0$. This feature indicates that there is a clear difference between average clustering in the entire network and in individual nodes. Figure 6.9 (right) shows how the clustering coefficient varies with the node out-degree. Higher values of the clustering coefficient occur among low out-degree nodes, suggesting the lack of large communities around high out-degree nodes.

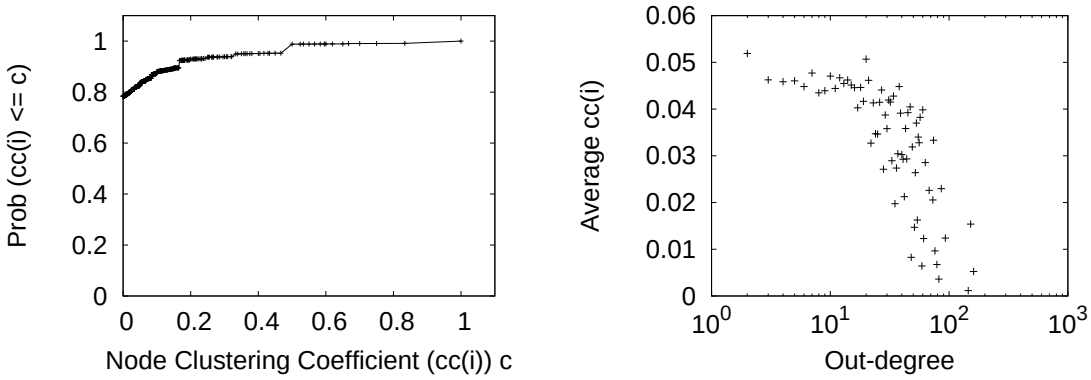


Figure 6.9: Distribution of CC (left) and average CC per out-degree (right)

6.3.3 Component analysis

Since not all the users in the network are both responsive and responded, we adopted a model called *bow tie structure* to study our network structure. The bow tie structure was initially proposed to provide insights of the organization of the network structure of the Web [38]. The key idea is that the Web graph (nodes are web pages and edges are links between the pages) has distinct components depicted in a figure that remembers a bow tie. In our context, the central *core* contains users which frequently communicates via video interactions, by posting and receiving video responses. It is the largest strongly connected component (SCC), meaning that one user can reach every other user following video response links. The *in* component contains users that post video responses to the *core*, but cannot be reached from it. The *out* consists of users that receive video responses posted by users in the *core* but do not link to it, such as directors or companies specialized in providing video content for YouTube. Other users (the Tendrils and Tubes), connect to either the *in* or *out* components, or both, but not to the *core*. They are users who receive video responses from the *in* users or who post video responses to the *out* users. Tendrils and Tubes are not represented in our analysis.

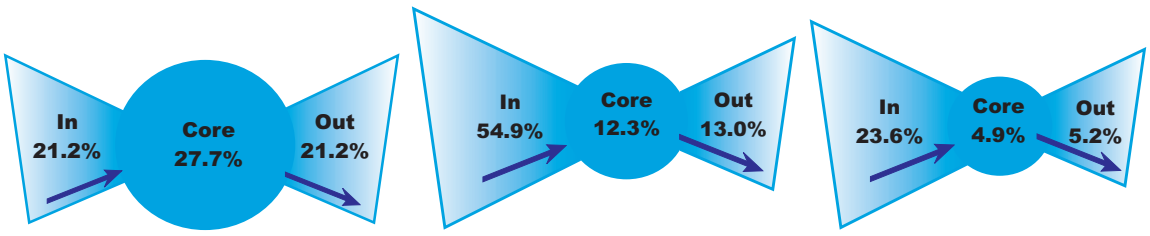


Figure 6.10: Bow Tie Structures of the Web (left), Java Forum (middle), and the Video Response User Graph (right)

Figure 6.10 compares the bow tie structures of the Web [38], of a textual Java Forum [144] (users are nodes and a directed edge from A to B means that A had a question answered by B), and of our video response user graph. Analyzing the fractions of nodes in the three major components, shown in the figures, we make two interesting observations. First, the bow tie structure of the video response interactions is very different from that of the Web, but shares some characteristics of the Java Forum structure. In particular, the fraction of nodes in each component is much more balanced in the Web structure, whereas there is a clear unbalance towards the *in* component in the other two. Moreover, if we divide the fraction of users falling into each component of the Java Forum structure by the corresponding fraction in the video response structure, we get ratios equal to 2.3, 2.5 and 2.5, for the *in*, *out* and *core* component, respectively, indicating a high structural similarity of the video response interactions with those in the textual forum.

Second, in both the video response user graph and in the Java Forum graph, there is a large number of users who post video responses and interact with the *core* users (i.e., users from the *in* and *core* component). However, only a small fraction of these users receive responses back. In fact, the video response user graph has the smallest *core* and *out* components, among the three structures.

There are two particular characteristics of the YouTube video response feature which might impact the video response graph bow tie structure. The first is the fact that a video can receive multiple video responses, but it can be posted as video response to only a single video. Such restriction of the YouTube video response feature allows the *in* component to grow freely, but limits the growth of the *core* component by imposing restrictions to the creation of response back links. The second is that the community structure developed by video interactions might be at an early stage. In fact, in order to further investigate this, we did another crawl following the same methodology presented in Section 4.2.2. The crawl was executed from 01/11/2008 to 01/18/2008. The *core* component of the graph created from this data set has size of 5.02%, which is sligtly larger than the *core* of our previous data set. On the other hand, the *in* component in the new data set is starting to diminish (22.08%). It seems that some new paths from the *core* to nodes in the *in* component have been established, thus increasing the size of the *core* component.

Next, we investigate the network properties of the largest SCC, which corresponds to the *core* component, of our video response user graph. In spite of its its small size, the *core* component concentrates 10% of the views and 22% of the video responses, and, thus, deserves further analysis. Our main findings are summarized in Table 6.2. In particular, the average clustering coefficient of the largest SCC is equal 0.137, three

times larger than the clustering coefficient of the entire network. Thus, user interactions captured by the largest SCC form a much more tightly connected community.

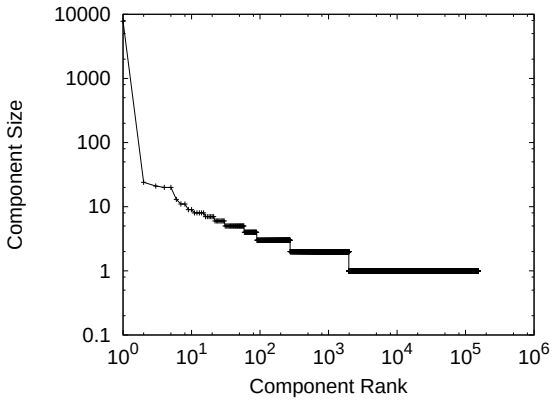


Figure 6.11: Component size rank

Figure 6.11 shows the distribution of the sizes of all strongly connected components of our network, sorted from the largest component to the smallest one. The *core* component is significantly larger than the others. The distribution suggests a general structure that includes the largest SCC, a large number (i.e., 147,805) of components with a single node, and a number of variable (small) sized components (1,974 components, in total). As we are working with a directed graph, the components with size one (singletons) are nodes with links in only one direction. The other smaller components are groups of users which represent small size communities, composed of a few people like a family or a group of friends, which express their interests and establish communication via video responses. In fact, there is a significantly negative correlation between the size and the average clustering coefficient of the components (correlation $C = -0.36$), indicating that small components tend to be more tightly connected.

6.3.4 Link Reciprocity

Another interesting metric to observe is the link reciprocity coefficient ρ, a metric which captures the reciprocity of the interactions in the entire network [64]. Given the adjacency matrix of a directed graph, that is, $a_{ij} = 1$ if there is a link from i to j, and $a_{ij} = 0$ otherwise, the reciprocity coefficient ρ is defined as the correlation coefficient between the matrix entries. In other words:

$$\rho = \frac{\sum_{i \neq j}(a_{ij} - \bar{a})(a_{ji} - \bar{a})}{\sum_{i \neq j}(a_{ij} - \bar{a})^2}, \tag{6.1}$$

where the average value $\bar{a} = \sum_{i \neq j} a_{ij}/N(N-1)$, and N is the number of users in the graph.

The reciprocity coefficient tells whether the number of mutual links in the network is more or less than that of a random network. If the value of ρ is higher than 0, the network is reciprocal; otherwise, anti-reciprocal. The ρ value for the video response user graph is 0.051, which means that the network is more reciprocal than a randomized graph. In contrast, the reciprocity coefficients of the World Wide Web and of Wikipedia[2] have been found to be 0.5161 [64] an In terms of networks created essentially by textual communication, the reciprocity coefficient is 0.231 for e-mail [64], 0.28 for Slashdot[3] [72], 0.58 for Twitter [83], and 0.765 for guestbook communication in Cyworld [51].

Thus, among all these networks, the quantitative link reciprocity of the video response communication is the smallest. In conclusion, our analysis shows that the video response feature triggers weakly reciprocal communication, specially in comparison with networks built from textual interactions.

6.3.5 Average Distance

Lastly, we look at the average distance of our video response network, which is the average number of hops along the shortest paths for all possible pairs of network nodes [116]. The average distance of our network is 8.40, a low value if compared with other directed graphs such as the Web (average distance equal to 16.12) [38]. As shown in [105], the average path length between nodes is usually short in social networks.

Such a small average distance combined with an average clustering coefficient almost 7 times larger than that of the same graph with randomized links constitute the main properties of a small world graph [14; 38]. These properties were also verified in textual communication contexts such as MSN instant messages [97] and in friendship relations in LiveJournal, Flickr, Orkut, and YouTube [105].

6.4 Evidences of unsolicited information

Different forms of unsolicited information are taking a toll on users of social networking services [147]. Unsolicited information opens a large gray area, where videos could be considered spam or promotion. One form of spam occurs when users submit a video with a long list of misleading tags to describe its content in order to fool video

[2] www.wikipedia.com
[3] www.slashdot.com

searching mechanisms [77]. Another form of video spam occurs when a video is posted as a response to a video topic, but whose content is completely unrelated to the video topic. On the other hand, promotion consists of users trying to boost the ranking of a video to make it more visible in the social network ranks. Towards that goal, users upload a large number of video responses to a video topic, most of which are not necessarily related to the responded video.

Due to its intrinsic nature, video response appears to be an attractive feature to users willing to spam in order to promote specific content, advertise to generate sales, disseminate pornography (often as an advertisement) or simply compromise the system reputation. Unlike textual responses or comments, one has to actually watch the video to realize it is or contains some form of spam or promotion, consuming system resources, in particular bandwidth, and compromising user patience and satisfaction with the system.

A simple example illustrates how opportunistic behavior, such as spamming, promotion, and unexpected advertising, could be identified by examining user behavior patterns in a social networking service. Figure 6.1 (left) shows that the distribution of the number of video responses posted by responsive users follows a power law. Experimental data have been used in the literature to show that the behavior of many real-world systems can be modeled by a power law distribution. We notice from the figure that three points (leftmost part of the curve) do not fit well the expected power law. They represent users who have posted a much larger number of video responses than predicted by the model. By identifying these users in our dataset, we realized they share common characteristics, namely: (1) they post video responses to either their own videos or to other specific videos to boost their rankings in order to increase their visibility, and (2) they make use of video responses as a marketing opportunity in the social networking environment, spreading video-based information to influence others or to advertise commercial products and services.

In the following, we discuss two simple metrics that may help understanding different types of user behaviors in video-based interactions, and may serve as a first cut in the identification of opportunistic behavior.

6.4.1 Characterizing user behavior

We define the Inter-Reference Distance (IRD) of user who uploads video responses to a video topic as the sum of the numbers of responses (plus one) that appear between two consecutive video responses of the same user. As an example, consider the IRDs computed to the following two video topics and the contributors of their

video responses.

Video Topic 1: $U_1\ U_2\ U_2\ U_1$ $\quad$ $\text{IRD}(U_1) = 3, \text{IRD}(U_2) = 1$
Video Topic 2: $U_1\ U_1\ U_1$ $\quad$ $\text{IRD}(U_1) = 1{+}1 = 2$

We compute a user's average IRD by first calculating her IRD for each video she responded to, and then taking the average over all videos responded by her. In the above example, the average IRD for user U_1 is computed as $(3 + 2)/2 = 2.5$.

The IRD metric allows us to assess temporal patterns of the users' participation in a sequence of video responses. Previous studies on spam characterization refer to the importance of analyzing temporal issues in order to detect malicious and opportunistic behavior [71]. For example, whereas traditional e-mail traffic is concentrated on diurnal periods, the arrival rate of e-mail spams is roughly stable over time [71].

Figure 6.12 plots the average IRD for each responsive user as a function of the average number of video responses per video responded by the user. A user who uploads many video responses per video, one after the other, mostly following a mechanical process, might be a candidate for further investigation. Thus, the combination of a large number of video responses per video and a small average IRD suggests the user exihibits some type of opportunistic behavior, such as spamming.

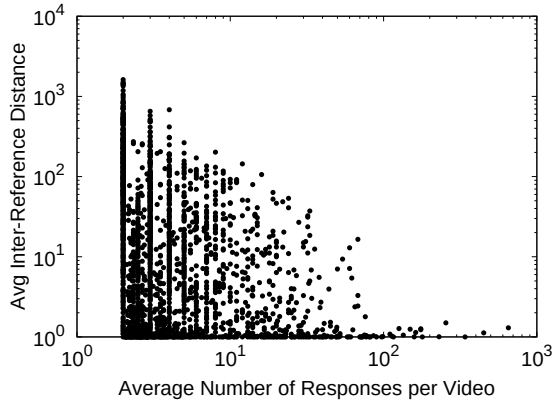


Figure 6.12: Temporal patterns of user participation in sequences of video responses

We conducted an investigation to verify if the combination of these two metrics (i.e., average IRD and average number of responses per responded video) could accurately be used to identify users with opportunistic behavior. Our experiment focused on users that are located on the rightmost part of Figure 6.12. In the dataset, a total of 298 users have average IRD less than 3 and average number of responses per video greater than 10. A group of volunteers in our laboratory randomly selected 95 from

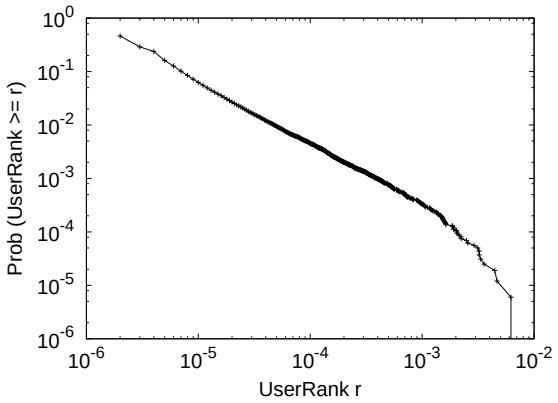


Figure 6.13: UserRank scores

these 298 users to be analyzed[4]. Our volunteers viewed the video responses posted by each user, and, according to their content, classified the responsive user into either social or opportunistic. If at least one video response was considered spam or promotion, the responsive user was labeled as opportunistic. A total of 80% of the analyzed users were classified as opportunistic, suggesting that the proposed metrics could be a starting point to develop heuristics to combat opportunistic behavior in video interactions.

6.4.2 Ranking users

The next step is to use the structure of the social network for detecting opportunistic patterns. Towards that goal, we explore the use of the PageRank algorithm [37], first proposed for the Web context. The intuition behind PageRank is that a Web page is important if it has many incoming links or if the page has links coming from highly ranked pages. Thus, the importance of a certain Web page influences and is influenced by the importance of some other pages.

We explore the application of the PageRank algorithm, discussed in Section 2.4.1, to determine the importance of a user in the video communication network. In this context, we refer to the scores computed by the algorithm as UserRanks, which represent indicators of the importance of users in terms of their participation in video interactions.

Figure 6.13 shows the complementary cumulative distribution of the UserRank scores in our network. Similarly to the Web [119], it clearly follows a power law, with most users with very low scores. However, there are a few users with very high scores.

[4]The sample size was chosen so as to guarantee a 90% confidence level and errors under 7% [82].

We analyzed the profiles and the video responses of a few users in both extremes of the curve. Users with high scores are among the most responded and viewed. Most of them are directors (i.e., a director account has special privileges in YouTube). In comparison, users with low scores own videos which are rarely viewed and that receive, at most, only a few video responses from the video community.

In order to assess if the UserRank scores capture the importance of a user in the video response user graph, we computed the correlation coefficient C between the UserRank and several other characteristics of users and of their videos. As shown in Table 6.3, there exists a strong correlation between UserRank and two metrics, namely, the total number of ratings received by the user's videos and total number of views of the user's videos, suggesting that the algorithm captures the importance of users in terms of the numbers of views and of ratings of their contributions. These strong correlations can also be seen in the scatter plots presented in Figure 6.14. In contrast, the correlations between UserRank scores and all other analyzed metrics such as link reciprocity, number of friends, number of videos uploaded, and clustering coefficient, are much weaker.

Characteristic	**Correlation coefficient**
Total number of ratings received by user's videos	0.44
Total number of views of user's videos	0.27
Total number of user's videos favorited	0.17
Link reciprocity	0.14
Out-degree	0.13
Number of friends	0.11
Number of videos	0.07
Clustering coefficient	0.04

Table 6.3: Correlation coefficient of the UserRank with different characteristics of users and their videos

Next, we assess the potential benefit of using the UserRank score to detect users that exploit the video response feature to boost video ranks, aiming at promoting their content. Intuitively, boosting a video rank can bring some extra visibility it. However, it may also be perceived as a spam or undesired content by other users. Moreover, videos which quickly reach a high rank are strong candidates to be kept in caches or in content distribution networks (CDNs) [42]. Thus, promoted videos can be confused with popular videos, impacting not only the user satisfaction with the system, but also the performance and scalability of online video sharing systems.

Our strategy to detect suspect users consists of searching for users with low UserRank scores who have videos on the top lists. The intuition behind it is the fol-

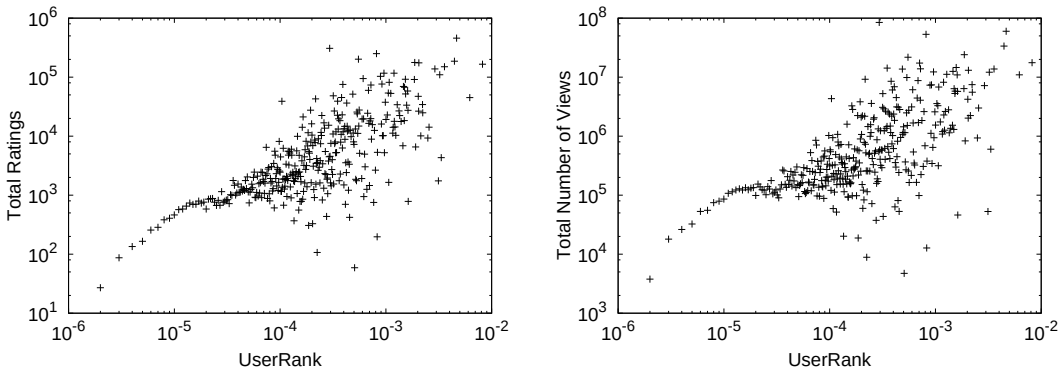


Figure 6.14: Correlation between UserRank and total number of ratings (left), and total number of views (right)

lowing. Suppose a video had its rank boosted by a series of self-responses or video responses posted by (fake) accounts with low rank. Suppose also that the number of (self-)responses posted was such that the video made it to one of the top-100 lists of most responded videos kept by YouTube. Clearly, the UserRank score of the contributor of this promoted video should be lower than the scores of the contributors of the other non-promoted videos in the same top-100 list, as these received video responses from several different users, possibly with different scores.

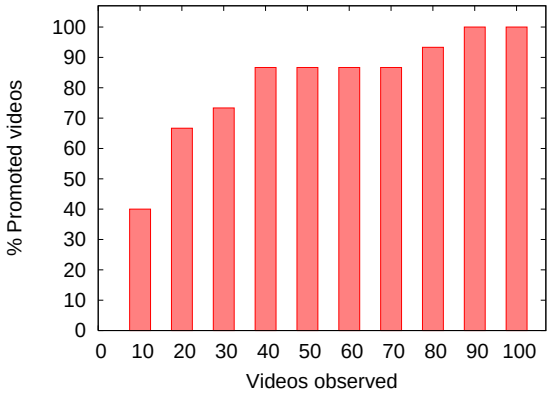


Figure 6.15: Detecting promoted videos in the top-100 most responded list

In order to verify if the UserRank can be useful to pinpoint potential promoters among the contributors of the most responded videos, we conducted the following experiment. We first took the videos in the all-time top-100 most responded video list (used as seeds to our crawler), and sorted them in increasing order according to the video contributor's UserRank. Next, we manually inspected each video of this sorted list, starting from its beginning, and, determined, after each round of 10 videos inspected, the total number of promoted videos found. Figure 6.15 illustrates this

experiment, showing the cumulative percentage of promoted videos found after each round of the experiment. By observing only 40 videos of the users with the lowest UserRank scores, we are able to identify 87% of the promoted videos in the top-100 most responded list, a much higher fraction than what we would have achieved if we had randomly selected the videos to observe.

6.5 Discussion

YouTube provides a video response feature that allows users to post a video as a response to another. Such a feature enriches the online interaction between users, allowing them to exchange knowledge and express ideas through video interactions. In this chapter, we perform an extensive characterization of users interacting with each other through the YouTube video response feature. Our measurement analysis provides many interesting findings that we think will be useful in various ways. We discuss implications of the findings below.

Characterization of usage of the video response feature: we characterize the usage of the video response feature, unveiling interesting user behavior and providing statistical models for various characteristics (e.g., popularity profiles and video duration). Such models and findings provide valuable insights for the future design of realistic synthetic workloads for video sharing systems.

Characterization of video interactions: in order to understand the characteristics of video-based interactions, we provide a characterization of video interactions at two levels of granularity, the interactions engendered by a single responded video and interactions under the perspect of complex network analysis. Characterizing interaction in both levels is valuable in order to build models that describe video-based communication. Not only is it important to identify the intrinsic properties of this type of communication in social networking environments, but it is also important to explain how and why these properties emerge. As example, our work has been used as the basis for research in the field of linguistics that aims at exploring the new paradigm of communication that video sharing systems start to offer [9].

Evidence of opportunistic behavior in video sharing systems: our study unveils evidence of two types of opportunistic behavior by YouTube users, namely, promotion and video spamming. The former consists of artificially boosting a video rank to make

it appear in the top lists provided by YouTube, thus gaining visibility to the content or to its contributor. The latter consists of adding an unrelated video as response to another video in order to promote specific content, advertise to generate sales or disseminate pornography. We preliminarily approached the problem, proposing the use of simple metrics which may serve as a first cut in the identification of opportunistic behavior. However, although IRD and UserRank showed to be promising to detect opportunistic users, we believe that this subject worth a deeper investigation. In the next chapter we focus on using social network aspects and video characteristics to identify opportunistic users in online video sharing systems.

Chapter 7

Spammers and Promoters in Video Interactions

In Chapter 6 we reveal the existence of new forms of opportunistic interactions in video interactions through the use of the YouTube video response feature. In this chapter, we go a step further by addressing the issue of detecting the users that do this kind of interaction. First, we present a few basic definitions and concepts in Section 7.1. Section 7.2 describes our strategies to build a test collection of malicious and legitimate users. Section 7.3 investigates the behavior of these users as well as a set of user attributes and their capability to distinguish users. Section 7.4 describes our approach to detect the malicious users. Finally, Section 7.5 offers conclusions.

7.1 Basic definitions: video spammers and video promoters

We begin introducing a few definitions used in this chapter. We define as *spammer* a user who posts at least *one* video response that is considered unrelated to the responded video (i.e., a spam). Examples of video spams are: (i) an advertisement of a product or website completely unrelated to the subject of the responded video, and (ii) pornographic content posted as response to a cartoon video. A *promoter* is defined as a user who posts a large number of video responses to a *responded video*, aiming at promoting this *video topic*. For instance, we found promoters in our dataset who posted a long sequence (e.g., 100) of (unrelated) video responses, often without content (few seconds) to a single video. A user that is neither a spammer nor a promoter is considered legitimate.

Promoters and spammers are driven by several goals, such as to spread advertise to generate sales, disseminate pornography (often as an advertisement to a Web site), or just to compromise system reputation. Since users cannot easily identify the spam and promotion before watching at least a segment of it, promoters and spammers also provoke an extra consuption of user and system resources.

7.2 User test collection

In order to evaluate our approach to detect video spammers and promoters we need a test collection of users, pre-classified into the target categories, namely, spammers, promoters, and legitimate users. However, no such collection is publicly available, thus requiring us to build one[1].

To build our user test collection we first recollected YouTube and then we selected and manually classified a subset of these users. The strategy used to collect a sample of users who participate in interactions through video responses is exactly the same described in Section 4.2.2. The reason to recollect the data is to obtain updated information about users. Since the creation of a test collection requires manual inspection of users and their videos, it is important that the statistics we collected match the characteristics of the users and videos by the time they were inspected. By using Algorithm 1 (see Section 4.2.2), the crawler ran for one week (01/11-18, 2008), gathering a total of **264,460** users, **381,616** responded videos and **701,950** video responses.

Since the main goal of creating a user test collection is to study the patterns and characteristics of each class of users, the desired properties for our test collection are: (1) having a significant number of users of all three categories; (2) including, but not restricting to, spammers and promoters which are aggressive in their strategies and generate large amounts of spam or promotion in the system; and (3) including a large number of legitimate users with different behavioral profiles. We argue that these properties may *not* be achieved by simply randomly sampling the collection. The reasons for this are twofold. First, randomly selecting a number of users from the crawled data could lead us to a small number of spammers and promoters, compromising the creation of effective training and test data sets for our analysis. Moreover, research has shown that the sample does not need to follow the class distribution of the collection in order to achieve effective classification [136]. Second, selecting legitimate users randomly may lead to a large number of users with similar behavior (i.e. post one video response to a discussed topic), not including examples with different profiles.

[1]This test collection as well as instructions to use it are available at http://homepages.dcc.ufmg.br/~fabricio/testcollectionsigir09.html

Aiming at capturing all these properties, we define three strategies for user selection (described below). Each selected user was then manually classified. However, this classification relies on human judgment that decides whether a video is related to another. In order to minimize the impact of human error, three volunteers analyzed all video responses of each selected user in order to independently classify her into one of the three categories. In case of tie (i.e., each volunteer chooses a different class), a fourth independent volunteer was heard. Each user was classified based on majority voting. Volunteers were instructed to favor legitimate users. For instance, if one was not confident that a video response was unrelated to the responded video, she should consider it to be legitimate. Moreover, video responses containing people chatting or expressing their opinions were classified as legitimate, as we choose not to evaluate the expressed opinions. The volunteers agreed in about 97% of the analyzed videos, which reflects a high level of confidence to this human classification process. The three user selection strategies used are:

(1) In order to select users with different levels of interaction through video responses, we first defined four groups of users based on their in and out-degrees in the video response user graph. Group 1 consists of users with low (≤ 10) in and out-degrees, and thus who respond to and are responded by only a few other users. Group 2 consists of users with high (> 10) in-degree and low out-degree, and thus receive video responses from many others but post responses to only a few users. Group 3 consists of users with low in-degree and high out-degree, whereas very interactive users, with high in and out-degrees, fall into group 4. One hundred users were randomly selected from each group[2], and manually classified, yielding a total of 382 legitimate, 10 spammers, and no promoter. The remaining 8 users were discarded as they had their accounts suspended due to violation of terms of use.

(2) Aiming at populating the test collection with spammers and promoters, we searched for them where they are more likely to be found. We first note that, in YouTube, a video v can be posted as response to at most *one* video at a time (unless one creates a copy of v and uploads it with a different ID). Thus, it is more costly for spammers to spread their video spam in YouTube than it is, for instance, to disseminate spam by e-mail. We conjecture that spammers would post their video responses more often to popular videos so as to make each spam visible to a larger community of users. Moreover, some video promoters might eventually be successful and have their target listed among the most popular videos. Thus, we browsed the video responses posted

[2]Groups 1, 2, 3 and 4 have 162,546, 2,333, 3,189 and 1,154 users. Thus, homogeneous random selection from each one yields a bias towards group 4.

to the top 100 most responded videos of all time, selecting a number of *suspect* users[3]. The classification of these suspect users led to 7 legitimate users, 118 spammers, and 28 promoters in the test collection.
(3) To minimize a possible bias introduced by strategy (2), we *randomly* selected 300 users who posted video responses to the top 100 most responded videos of all time, finding 252 new legitimate users, 29 new spammers and 3 new promoters (16 users with closed accounts were discarded).

In total, our test collection contains 855 users, including 641 classified as legitimate, 157 as spammers and 31 as promoters. Those users posted 20,644 video responses to 9,796 unique responded videos. Our user test collection aims at supporting research on detecting spammers and promoters. Since the user classification labeling process relies on human judgment, which implies in watching a significantly high amount of videos, the number of users in our test collection is somewhat limited.

7.3 Analyzing user attributes

Legitimate users, spammers and promoters have different goals in the system, and, thus, we expect they also differ on how they behave (e.g., who they interact with, which videos they post) to achieve their purposes. Thus, our next step is to analyze a large set of attributes that reflect user behavior in the system aiming at investigating their relative discriminatory power to distinguish one user class from the others. We considered three attribute sets, namely, video attributes, user attributes, and social network (SN) attributes.

Video attributes capture specific properties of the videos uploaded by the user, i.e., each user has a set of videos in the system, each one with attributes that may serve as indicators of its "quality", *as perceived by others*. In particular, we characterize each video by its duration, numbers of views and of commentaries received, ratings, number of times the video was selected as favorite, as well as numbers of honors and of external links. Moreover, we consider three separate groups of videos owned by the user. The first group contains aggregate information of *all videos* uploaded by the user, being useful to capture how others see the (video) contributions of this user. The second group considers only *video responses*. The last group considers only the *responded videos* to which this user posted video responses (referred to as *target* videos). For each video group, we considered the average and the sum of the aforementioned attributes, summing up 42 video attributes for each user, all of which can be easily derived from

[3]For example, the owner of a video with a pornographic picture as thumbnail but posted to a political debate video discussion topic.

data maintained by YouTube. We explicitly choose not to add any attribute that would require processing the multimedia content itself.

The second set of attributes consists of individual characteristics of user behavior. We expect that legitimate users spend more time doing actions such as selecting friends, adding videos as favorites, and subscribing to content updates from others. Thus, we select the following 10 user attributes: number of friends, number of videos uploaded, number of videos watched, number of videos added as favorite, numbers of video responses posted and received, numbers of subscriptions and subscribers, average time between video uploads, and maximum number of videos uploaded in 24 hours.

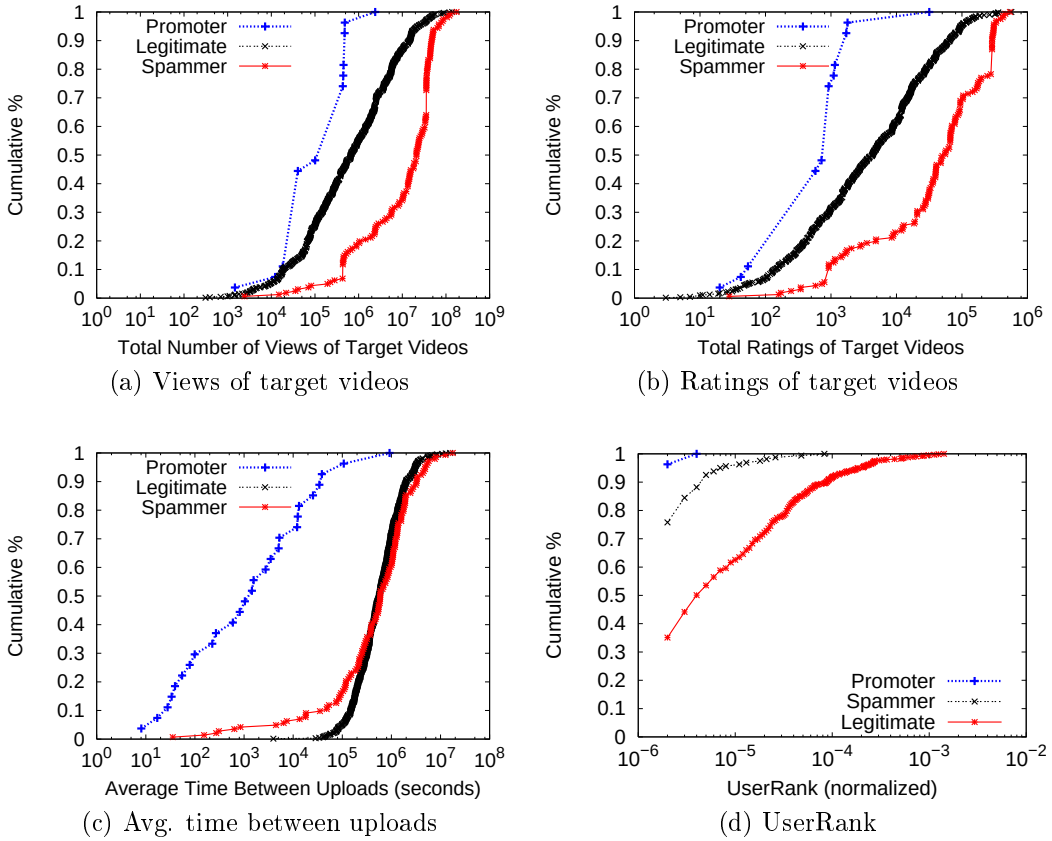
The third set of attributes captures the social relationships established between users via video response interactions, which is one of the several possible social networks in YouTube. The idea is that these attributes might capture specific interaction patterns that could help differentiate legitimate users, promoters, and spammers. We selected the following node attributes extracted from the video response user graph, which capture the level of (social) interaction of the corresponding user: clustering coefficient, betweenness, reciprocity, assortativity, and UserRank. Particularly, node assortativity is defined, as in [40], as the ratio between the node (in/out) degree and the average (in/out) degree of its neighbors. We compute node assortativity for the four types of degree-degree correlations (i.e., in-in, in-out, out-in, out-out). Thus, in total, we selected 8 social network attributes.

Attribute Set	**Top 10**	**Top 20**	**Top 30**	**Top 40**	**Top 50**
Video	9	18	25	30	36
User	1	2	4	7	9
SN	0	0	1	3	5

Table 7.1: Number of attributes at top positions in χ^2 ranking

We assessed the relative power of the 60 selected attributes in discriminating one user class from the others by independently applying two well known feature selection methods, namely, information gain and χ^2 (Chi Squared) [141]. Table 7.1 summarizes the results, showing the number of attributes from each set (video, user, and social network) in the top 10, 20, 30, 40, and 50 most discriminative attributes according to the ranking produced by χ^2. Results for information gain are very similar and, thus, are omitted.

Note that the 9 of the 10 most discriminative attributes are video-related. In fact, the most discriminative attribute (according to *both* methods), is the total number of views (i.e., the popularity) of the *target* videos. Figure 7.1(a) presents the cumulative distributions of this metric for each user class, showing a clear distinction among them.



(a) Views of target videos

(b) Ratings of target videos

(c) Avg. time between uploads

(d) UserRank

Figure 7.1: Cumulative distribution of user behavior attributes

The curve for spammers is much more skewed towards a larger number of views, since these users tend to target popular videos in order to attract more visibility to their content. In contrast, the curve for promoters is more skewed towards the other end as they tend to target videos that are *still* not very popular, aiming at raising their visibility. Legitimate users, being driven mostly by social relationships and interests, exhibit an intermediary behavior, targeting videos with a wide range of popularity. The same distinction can be noticed for the distributions of the total ratings of target videos, shown in Figure 7.1(b), another metric that captures user feedback with respect to these videos, and is among the top 10 most discriminative attributes.

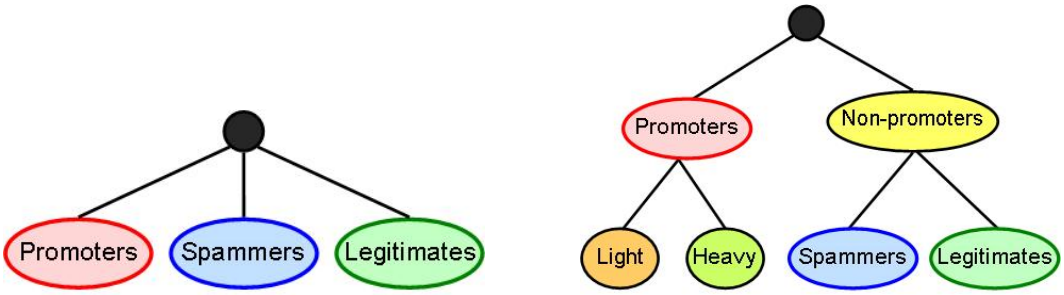
The most discriminative user and social network attributes are the average time between video uploads and the UserRank, respectively. In fact, Figure 7.1(c) and (d) show that, in spite of appearing in lower positions in the ranking, particularly for the UserRank attribute (see Table 7.1), these two attributes have potential to be able to separate user classes apart. In particular, the distribution of the average time between video uploads clearly distinguishes promoters, who tend to upload at a much higher

frequency since their success depends on them posting as many video responses to the target as possible. Figure 7.1(c) also shows that, at least with respect to this user attribute, spammers can not be clearly distinguished from legitimate users. Finally, Figure 7.1(d) shows that legitimate users tend to have much higher UserRank values than spammers, who, in turn, have higher UserRank values than promoters. This indicates that, as expected, legitimate users tend to have a much more participative role (system-wide) in the video response interactions than users from the other two classes, which are much more selective when choosing their targets.

7.4 Detecting spammers and promoters

In this section, we investigate the feasibility of applying a supervised learning algorithm along with the attributes discussed in the previous section for the task of detecting spammers and promoters. In this approach, each user is represented by a vector of values, one for each attribute. The algorithm learns a classification model from a set of previously labeled (i.e., pre-classified) data, and then applies the acquired knowledge to classify new (unseen) users into three classes: legitimate, spammers and promoters. Note that, in this chapter, we do not address the labeling process. Labeled data may be obtained through various initiatives (e.g., volunteers who help marking video spam, professionals hired to periodically manually classify a sample of users, etc). Our goal here is to assess the *potential effectiveness* of the proposed approach as a first effort towards helping system administrators to detect spammers and promoters in online video sharing systems.

We start by presenting, in Section 7.4.1, the metrics used to evaluate our experimental results. Section 7.4.2 describes the classification algorithm, i.e., the classifier, and the experimental setup used. The classifier was applied according to two different strategies, referred to as flat and hierarchical classifications. In the flat classification, illustrated in Figure 7.2(a), the users from the test collection are directly classified into promoters, spammers, and legitimate users. In the hierarchical strategy, the classifier is first used to separate promoters from non-promoters. Next, it classifies promoters into heavy and light promoters, as well as non-promoters into legitimate users and spammers, in a hierarchical fashion shown in Figure 7.2(b). Results from our flat and hierarchical classifications are presented in Sections 7.4.3 and 7.4.4. Section 7.4.5 discusses the impact of reducing the attribute set on the classification effectiveness.



(a) Flat classification (b) Hierarchical classification

Figure 7.2: Classification strategies to detect spammers and promoters

7.4.1 Evaluation metrics

To assess the effectiveness of our classification strategies we use the standard information retrieval metrics of recall, precision, Micro-F1, and Macro-F1 [140]. The recall (r) of a class X is the ratio of the number of users correctly classified to the number of users in class X. Precision (p) of a class X is the ratio of the number of users classified correctly to the total predicted as users of class X. In order to explain these metrics, we will make use of a confusion matrix [88], illustrated in Table 7.2. Each position in this matrix represents the number of elements in each original class, and how they were predicted by the classification. In Table 7.2, the precision (p_{prom}) and the recall (r_{prom}) of the class promoter are computed as $p_{prom} = a/(a+d+g)$ and $r_{prom} = a/(a+b+c)$.

		Predicted		
		Promoter	**Spammer**	**Legitimate**
	Promoter	a	b	c
True	**Spammer**	d	e	f
	Legitimate	g	h	i

Table 7.2: Example of confusion matrix

The F1 metric is the harmonic mean between both precision and recall, and is defined as $F1 = 2pr/(p+r)$. Two variations of F1, namely, micro and macro, are normally reported to evaluate classification effectiveness. Micro-F1 is calculated by first computing global precision and recall values for all classes, and then calculating F1. Micro-F1 considers equally important the classification of *each user*, independently of its class, and basically measures the capability of the classifier to predict the correct class on a per-user basis. In contrast, Macro-F1 values are computed by first calculating F1 values for each class in isolation, as exemplified above for promoters, and then averaging over all classes. Macro-F1 considers equally important the effectiveness in

each class, independently of the relative size of the class. Thus, the two metrics provide complementary assessments of the classification effectiveness. Macro-F1 is especially important when the class distribution is very skewed, as in our case, to verify the capability of the classifier to perform well in the smaller classes.

7.4.2 The classifier and the experimental setup

We use a Support Vector Machine (SVM) classifier [85], which is considered one of the state-of-the-art methods in classification. The goal of a SVM is to find the hyperplane that optimally separates with a maximum margin the training data into two portions of an N-dimensional space. A SVM performs classification by mapping input vectors into an N-dimensional space, and checking in which side of the defined hyperplane the point lies. SVMs are originally designed for binary classification but can be extended to multiple classes using several strategies (e.g., one against all [78]). We use a non-linear SVM with the Radial Basis Function (RBF) kernel to allow SVM models to perform separations with very complex boundaries. The implementation of SVM used in our experiments is provided with libSVM [61], an open source SVM package that allows searching for the best classifier parameters using the *training* data, a mandatory step in the classifier setup. In particular, we use the *easy* tool from libSVM, which provides a series of optimizations, including normalization of all numerical attributes. For experiments involving the SVM J parameter (discussed in Section 7.4.4), we used a different implementation, called SVM light, since libSVM does not provide this parameter. Classification results are equal for both implementations when we use the same classifier parameters.

The classification experiments are performed using a 5-fold cross-validation. In each test, the original sample is partitioned into 5 sub-samples, out of which four are used as training data, and the remaining one is used for testing the classifier. The process is then repeated 5 times, with each of the 5 sub-samples used exactly once as the test data, thus producing 5 results. The entire 5-fold cross validation was repeated 5 times with different seeds used to shuffle the original data set, thus producing 25 different results for each test. The results reported are averages of the 25 runs. With 95% of confidence, results do not differ from the average in more than 5%.

In the following two sections, we discuss the results obtained with the two classification strategies (flat and hierarchical) using all 60 selected attributes, since, as discussed in Section 7.3, even attributes with low ranks according to the employed feature selection methods (e.g., UserRank) may have some discriminatory power, and may be useful to classify users. Moreover, SVMs are known for dealing well with high

dimensional spaces, properly choosing the weights for each attribute, i.e., attributes that are not helpful for classification are given low weights by the optimization method used by the SVM [85]. The impact of using different subsets of the attributes on the classification effectiveness is analyzed in Section 7.4.5.

7.4.3 Flat classification

Table 7.3 shows the confusion matrix obtained as the result of our experiments with the flat classification strategy. The numbers presented are percentages relative to the total number of users in each class. The diagonal in boldface indicates the recall in each class. Approximately 96% of promoters, 57% of spammers, and 95% of legitimate users were correctly classified. Moreover, no promoter was classified as legitimate user, whereas only a small fraction of promoters were erroneously classified as spammers (3.87%). By manually inspecting these promoters, we found that the videos that they targeted (i.e., the promoted videos) actually acquired a certain popularity. In that case, it is harder to distinguish them from spammers, who target more often very popular videos, as well as from some legitimate users who, following their interests or social relationships, post responses to popular videos. Referring to Figure 7.1(a), these (somewhat successful) promoters are those located in the higher end of the curve, where the three user classes can not be easily distinguished.

		Predicted		
		Promoter	**Spammer**	**Legitimate**
	Promoter	**96.13%**	3.87%	0.00%
True	**Spammer**	1.40%	**56.69%**	41.91%
	Legitimate	0.31%	5.02%	**94.66%**

Table 7.3: Flat classification

A significant fraction (almost 42%) of spammers was misclassified as legitimate users. In general, these spammers exhibit a dual behavior, sharing a reasonable number of legitimate videos (non-spam) and posting legitimate video responses, thus presenting themselves as legitimate users most of the time, but occasionally posting video spams. This dual behavior masks some important aspects used by the classifier to differentiate spammers from legitimate users. This is further aggravated by the fact that a significant number of legitimate users post their video responses to popular responded videos, a typical behavior of spammers. Therefore, as opposed to promoters, which can be effectively separated from the other classes, distinguishing spammers from legitimate users is much harder. In Section 7.4.4.1, we discuss an approach that allows one to trade

a higher recall of spammers at a cost of misclassifying a larger number of legitimate users.

As a summary of the classification results, Micro-F1 value is 87.5, whereas per-class F1 values are 63.7, 90.8, and 92.3, for spammers, promoters, and legitimate users, respectively, resulting in an average Macro-F1 equal to 82.2. The Micro-F1 result indicates that we are predicting the correct class in almost 88% of the cases. Complementary, the Macro-F1 result shows that there is a certain degree of imbalance for F1 across classes, with more difficulty for classifying spammers. Comparing with a trivial baseline classifier that chooses to classify every single user as legitimate, we obtain gains of about 13% in terms of Micro-F1, and of 183% in terms of Macro-F1. As a first approach, our proposed classification provides significant benefits, being effective in identifying spammers and promoters in the system.

7.4.4 Hierarchical classification

Our flat classification results show that we can effectively identify promoters, but separating spammers from legitimate users is a harder task. This motivates us to experiment with a hierarchical classification strategy, illustrated in Figure 7.2 (right), which allow us to take advantage of a cost mechanism in the SVM classifier, specific for binary classification. In this mechanism, one can give priority to one class (e.g., spammers) over the other (e.g., legitimate users) by varying its J parameter [108]. The J parameter is the cost factor by which training errors in one class outweigh errors in the other. It is useful, when there is a large imbalance between the two classes, to counterbalance the bias towards the larger one. By varying J, we can study several tradeoffs and scenarios. In particular, we evaluate the tradeoffs between identifying more spammers at the cost of misclassifying more legitimate users (Section 7.4.4.1), and we further categorize promoters into heavy and light, based on their aggressiveness (Section 7.4.4.2). Splitting the set of promoters is also motivated by the potential for disparate behaviors with different impact on the system, thus requiring different treatments. On one hand, heavy promoters may reach top lists very quickly, requiring a fast detection. On the other hand, light promoters may conceal a collusion attack to promote the same responded video, thus requiring further investigation.

The results for the first phase of the hierarchical classification (promoters versus non-promoters) are summarized in Table 7.4. Macro-F1 and Micro-F1 are 93.44 and 99.17, respectively. Similarly to the results with the flat characterization, the vast majority of promoters were correctly classified (both results are statistically indistinguishable). In fact, the absolute number of erroneously classified users in each run of

		Predicted	
		Promoter	**Non-promoter**
	Promoter	**92.26%**	7.74%
True	**Non-promoter**	0.55%	**99.45%**

Table 7.4: Hierarchical classification of promoters vs. non-promoters

a test is very small (mostly 1 or 0).

7.4.4.1 Non-promoters

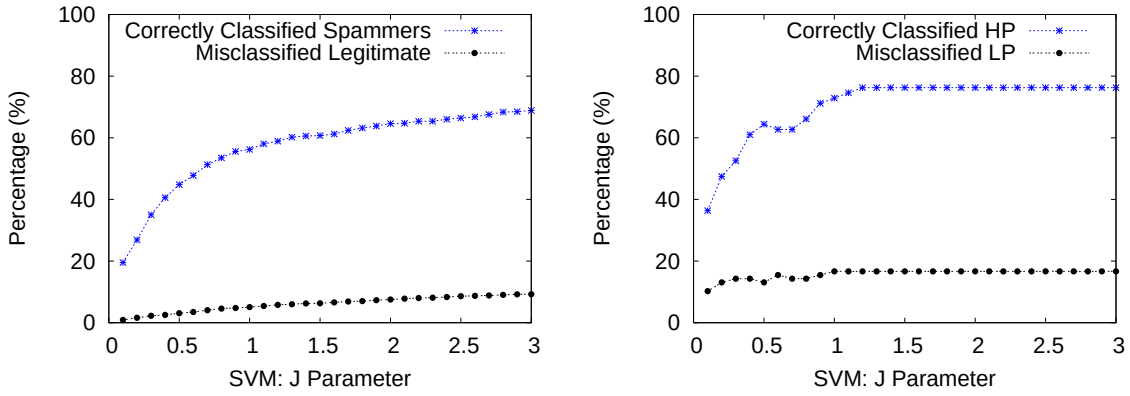
As previously discussed, there are cases of spammers and legitimate users acting similarly, making the task of differentiating them very difficult. In this section, we perform a binary classification of all (test) users identified as non-promoters in the first phase of the hierarchical classification, separating them into spammers and legitimate users. For this experiment, we trained the classifier with the original training data without promoters.

		Predicted	
		Legitimate	**Spammer**
	Legitimate	**95.09%**	4.91%
True	**Spammer**	41.27%	**58.73%**

Table 7.5: Hierarchical classification of non-promoters

Table 7.5 shows results of this binary classification. In comparison with the flat classification (Table 7.3), there was no significant improvement on separating legitimate users and spammers apart. These results were obtained with J=1. Figure 7.3(a) shows that increasing J leads to a higher percentage of correctly classified spammers (with diminishing returns for $J > 1.5$), but at the cost of a larger fraction of misclassified legitimate users. For instance, one can choose to correctly classify around 24% of spammers, misclassifying only 1% legitimate users ($J = 0.1$). On the other hand, one can correctly classify as much as 71% of spammers ($J = 3$), paying the cost of misclassifying 9% of legitimate users. The best solution to this tradeoff depends on the system administrator's objectives. For example, the system administrator might be interested in sending an automatic warning message to all users classified as spammers, in which case they might prefer to act conservatively, avoiding sending the message to legitimate users, at the cost of reducing the number of correctly predicted spammers. In another situation, the system administrator may prefer to detect a higher fraction of spammers for manual inspection. In that case, misclassifying a few more legitimate

users has no great consequence, and may be preferred, since they will be cleared out during inspection. It should be stressed that we are evaluating the potential benefits of varying J. In a practical situation, the optimal value should be discovered in the training data with cross-validation, and selected according to the system administrator goal.



(a) Spammers vs. legitimate (b) Heavy vs. light promoters

Figure 7.3: Impact of varying the J parameter

7.4.4.2 Heavy and light promoters

In order to be able to further classify promoters into heavy and light, we need first a metric to capture the promoter "aggressiveness", and then we must label each promoter as either heavy or light, according to this metric. The metric chosen to capture the aggressiveness of a promoter is the *maximum number of video responses posted in a 24-hour period*. We expect that heavy promoters would post a large number of videos in sequence in a short period of time, whereas light promoters, perhaps acting jointly in a collusion attack, may try to make the promotion process imperceptible to the system by posting videos at a much slower rate. The k-means clustering algorithm [81] was used to separate promoters into two clusters, labeled heavy and light, according to this metric.

Out of the 31 promoters, 18 were labeled as *light*, and 13 as *heavy*. As expected, these two groups of users exhibit different behaviors, with different consequences from the system perspective. Light promoters are characterized by an average "aggressiveness" of at most 15.78 video responses posted in 24 hours, with coefficient of variation (CV) equal to 0.63. Heavy promoters, on the other hand, exhibit an average behavior of posting as much as 107.54 video responses in 24 hours (CV=0.61). In particular,

after manual inspection, we found that all heavy promoters posted a number of video responses sufficient to boost the ranking of their targets to the top 100 most responded videos of the day (during collection period). Some of them even reached the top 100 most responded videos of the week, of the month and of all time. On the other hand, no light promoter posted enough video responses to promote the target to the top lists (during the collection). However, all of them participated in some collusion attack, with different subsets of them targeting different videos.

We performed a binary classification of all (test) users identified as promoters in the first phase of the hierarchical classification, separating them into light and heavy promoters. To that end, we retrained the classifier with the original training data containing only promoters, each one labeled according to the cluster it belongs to. The results are summarized in Table 7.6. Approximately 83% of light promoters and 73% of heavy promoters are correctly classified. Figure 7.3 (right) shows the impact of varying the J parameter, and how a system administrator can trade detecting more heavy promoters (HP) for misclassifying a larger fraction of light promoters (LP). A conservative system administrator may choose to correctly classify 36% of heavy promoters at the cost of misclassifying only 10% of light promoters ($J = 0.1$). A more aggressive one may choose to classify as much as 76% of heavy promoters, if she can afford misclassifying 17% of the light ones ($J \geq 1.2$).

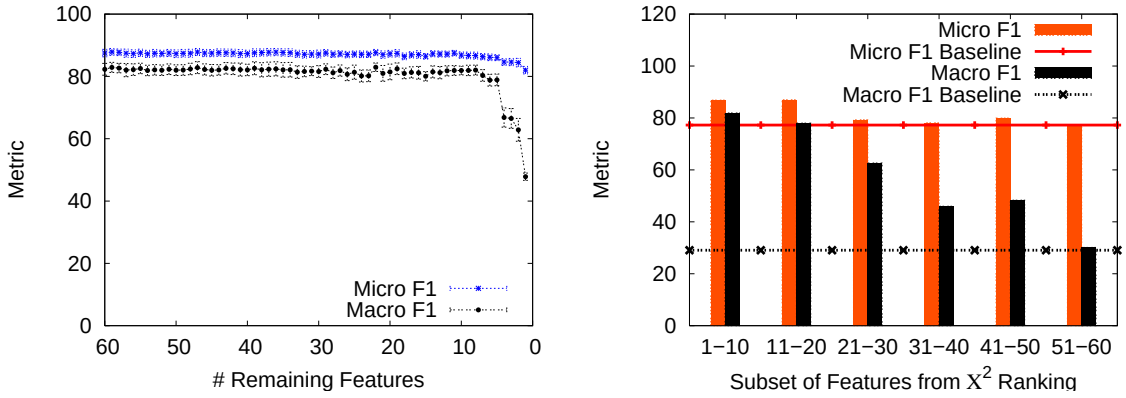
		Predicted	
		Light Promoter	**Heavy Promoter**
	Light Promoter	**83.33%**	16.67%
True	**Heavy Promoter**	27.12%	**72.88%**

Table 7.6: Hierarchical classification of promoters

An interesting finding of our work is with respect to collusion of promoters (especially light promoters). Intuitively, if we identify one element of a collusion, the rest of the collusion can be also detected by analyzing other users who post responses to the promoted video. By inspecting the video responses posted to some of the target videos of the detected promoters, we found hundreds of new promoters among the investigate users, indicating that our approach can also effectively unveil collusion attacks, guiding system administrator towards promoters that are more difficult to detect.

7.4.5 Impact of reducing the attribute set

Once we have understood the main tradeoffs and challenges in classifying users into spammers, promoters and legitimate, we now turn to investigate whether competi-



(a) Scenario 1 (b) Scenario 2

Figure 7.4: Impact of reducing the attribute set

tive effectiveness can be reached with fewer attributes. We report results for the flat classification strategy, considering two scenarios.

Scenario 1 consists of evaluating the impact on the classification effectiveness of gradually removing attributes in a decreasing order of position in the χ^2 ranking. Figure 7.4(a) shows Micro-F1 and Macro-F1 values, with corresponding 95% confidence intervals. There is no noticeable (statistical) impact on the classification effectiveness (both metrics) when we remove as many as the 40 lowest ranked attributes. It is worth noting that some of the most expensive attributes such as UserRank and betweenness, which require processing the entire video response user graph, are among these attributes. In fact, all social network attributes are among them, since UserRank, the best positioned of these attributes, is in the 30^{th} position. Thus, our classification approach is still effective even with a smaller, less expensive set of attributes. The figure also shows that the effectiveness drops sharply when we start removing some of the top 10 attributes from the process.

Scenario 2 consists of evaluating our classification when subsets of 10 attributes occupying contiguous positions in the ranking (i.e., the first top 10 attributes, the next 10 attributes, etc) are used. Figure 7.4(b) shows Micro-F1 and Macro-F1 values for the flat classification and for the baseline classifier that considers all users as legitimate, for each such range. In terms of Micro-F1, our classification provides gains over the baseline for the first two subsets of attributes, whereas significant gains in Macro-F1 are obtained for all attribute ranges, but the last one (the 10 worst attributes). This confirms the results of our attribute analysis that shows that even low-ranked attributes have some discriminatory power. In practical terms, significant improvements over the baseline are possible even if not all attributes considered in our experiments can be

obtained.

7.5 Discussion

Promoters and Spammers can pollute video retrieval features of online video sharing systems, compromising not only user satisfaction with the system, but also system resources and aspects such as caching. We propose an effective solution to the problem of detecting these opportunistic users that can guide system administrators to identify spammers and promoters in online video sharing systems. Relying on a sample of pre-classified users and on a set of user behavior attributes, our flat classification approach was able to detect correctly 96% of the promoters, 57% of spammers, wrongly classifying only 5% of the legitimate users. Thus, our proposed approach poses a promising alternative to simply consider all users as legitimate or to randomly select users for manual inspection. We also investigated a hierarchical version of the proposed approach, which explores different classification tradeoffs and provides more flexibility for the application of different actions to the detected spammers and promoters. As example, the system administrators may send warning messages for the suspects or put the suspects in quarantine for further investigation. In the first case, the system administrators could be more tolerant to misclassifications than in the second case, using the different classification tradeoffs we proposed. Finally, we found that our classification can produce significant benefits even if only a small subset of less expensive attributes is available.

Chapter 8

Conclusions and Future Work

In this thesis we provide an in-depth study of user interactions in OSNs. First, we gathered data from actual OSN sites, including YouTube and Orkut. The clickstream data analyzed offer an accurate view of how users behave and interact when they connect to OSN sites. Our data analysis suggests several interesting insights into how users interact with friends in Orkut and may have implications for efficient system and interface design as well as for advertisement placement in OSNs.

As our second step, we put our efforts towards characterizing the interactions that emerge from YouTube's video responses, a feature that allows users to interact primarily using videos rather than text. We provide an in-depth understanding of a unique type of interaction that emerge from multimedia content. Among a series of interesting findings, our work unveiled novel forms of unsolicited content in OSNs. Then, we developed a method for identifying these users that we named video spammers and video promoters. We created a labeled collection with users "manually" classified as legitimate, spammers, and promoters and we studied a number of characteristics of these users in order to be able to automatically distinguish them. Using attributes based on the user's profile, the user's social behavior in the system, and the videos posted by these users, we investigated the feasibility of applying a supervised learning method to identify spammers and promoters. Our approach is able to correctly identify the majority of the promoters, misclassifying only a small percentage of legitimate users. In contrast, although we are able to detect a significant fraction of spammers, they showed to be much harder to distinguish from legitimate users. These results motivated us to investigate a hierarchical classification approach, which explores different classification tradeoffs.

There is a number of directions towards which this work can envolve. First, an interesting direction is to investigate the impact of friends on the user behavior in

OSNs. The success of an OSN site is directly associated with the quality of content users share. Thus, in order to design social network services, it is key to understand factors that motivate users to join communities, become fans of something, and upload or retrieve media content.

Second, a research topic yet to be explored consists of understanding content distribution patterns across multiple OSNs. Given that users participate in multiple social networks, we expect that a user may share the same content across multiple sites. We would like to know to what extent content is shared across OSN sites as well as explore the impact of age, content, and geographical locality in object popularity. Answering these questions will let us explore opportunities for efficient content distribution, for example, caching and pre-fetching, as well as advertisement and recommendation strategies. For instance, certain types of content may be popular either in a specific geographical region or in a single social network, in which case advertisement algorithms should be based on this characteristic. On the other hand, if content is easily replicated across sites, then it is possible to detect rising content from one social networking site and implant it into another site.

Third, based on our analysis about navigation on Orkut, an interesting step to be taken would be to build a social network workload generator that incorporates many of our findings, including the statistical distributions of sessions and requests and the Markov models for user behavior.

Finally, there are also opportunities for improvements on the detection of spammers and promoters presented in Chapter 7. Since our proposed mechanism to detect opportunistic users consists on a supervised method (i.e., our approach requires labeled data for training), it demands a costly human intervention for the labeling process. In order to reduce the cost of obtain labeled data one could investigate other classes of machine learning techniques that make use of small amounts of labeled training data. Additionally, it is important to explore other refinements to the proposed learning approach such as to use different classification methods (maybe combined). More importantly, we believe that the methodology proposed to detect spammers and promoters could be used to detect other forms of opportunistic users in other OSNs sites and contexts.

Bibliography

[com] comscore: YouTube surpasses 100 million u.s. viewers for the first time. http://www.comscore.com/press/release.asp?press=2741. Accessed in March/2010.

[dev] Developer analytics. http://www.developeranalytics.com. Accessed in March/2010.

[1] Facebook application directory. http://www.facebook.com/apps. Accessed in March/2010.

[2] Facebook platform. http://developers.facebook.com. Accessed in March/2010.

[3] Facebook Press Room, Statistics. http://www.facebook.com/press/info.php?statistics. Accessed in March/2010.

[ope] Google OpenSocial. http://code.google.com/apis/opensocial/. Accessed in March/2010.

[wik] List of social network web sites. http://en.wikipedia.org/wiki/List_of_social_networking_websites. Accessed in March/2010.

[4] Needle in a Haystack: Efficient Storage of Billions of Photos. Facebook Engineering Notes, http://tinyurl.com/cju2og. Accessed in March/2010.

[5] New york times. a web site born in u.s. finds fans in brazil. http://www.nytimes.com/2006/04/10/technology/10orkut.html. Accessed in March/2010.

[nyt] New york times. uploading the avantgarde. http://www.nytimes.com/2009/09/06/magazine/06FOB-medium-t.htm. Accessed in July/2010.

[6] Orkut help. http://www.google.com/support/orkut/. Accessed in March/2010.

[7] Orkut on wikipedia. http://en.wikipedia.org/wiki/Orkut. Accessed in March/2010.

[spa] Spamassassin. `http://spamassassin.apache.org`. Accessed in March/2010.

[you] YouTube CNN debates. `http://www.youtube.com/debates`. Accessed in March/2010.

[8] YouTube fact sheet. `http://www.youtube.com/t/fact_sheet`. Accessed in March/2010.

[9] Adami, E. (2009). 'We/YouTube': exploring sign-making in video-interaction. *Visual Communication*, 4(8):379--399.

[10] Adamic, L., Buyukkokten, O., and Adar, E. (2003). A social network caught in the web. *First Monday*, 8(6).

[11] Adar, E. and Adamic, L. (2005). Tracking information epidemics in blogspace. In *Proceedings of the IEEE/WIC/ACM Conference on Web Intelligence (WI)*, pages 207--214.

[12] Ahn, Y.-Y., Han, S., Kwak, H., Moon, S., and Jeong, H. (2007). Analysis of topological characteristics of huge online social networking services. In *Proceedings of the World Wide Web Conference (WWW)*, pages 835--844.

[13] Albert, R., H., Jeong, and Barabasi, A. (2000). Error and attack tolerance of complex networks. *Nature*, 406(6794):378--382.

[14] Albert, R., Jeong, H., and Barabasi, A. (1999). Diameter of the world wide web. *Nature*, 401:130--131.

[15] Ali-Hasan, N. and Adamic, L. (2007). Expressing social relationships on the blog through links and comments. In *Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM)*.

[16] Almeida, V., Bestavros, A., Crovella, M., and Oliveira, A. (1996). Characterizing reference locality in the WWW. In *Proceedings of the Int'l Conference on Parallel and Distributed Information Systems (PDIS96)*.

[17] Amaral, A., Scala, A., Barthelemy, M., and Stanley, E. (2000). Classes of small-world networks. *Proceedings of the National Academy of Science (PNAS)*, 97(21):11149--11152.

[18] Aradhye, H., Myers, G., and Herson, J. (2005). Image analysis for efficient categorization of image-based spam e-mail. In *Proceedings of the Conference on Document Analysis and Recognition (ICDAR)*.

[B. Williamson] B. Williamson. Social network marketing: ad spending and usage. *EMarketer Report*, 2007. `http://tinyurl.com/2449xx`. Accessed in March/2010.

[19] Baluja, S., Seth, R., Sivakumar, D., Jing, Y., Yagnik, J., Kumar, S., Ravichandran, D., and Aly, M. (2008). Video suggestion and discovery for YouTube: Taking random walks through the view graph. In *Proceedings of the World Wide Web Conference (WWW)*, pages 895--904.

[20] Barabasi, A. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439).

[Barford and Crovella] Barford, P. and Crovella, M. Generating representative web workloads for network and server performance evaluation. *SIGMETRICS Performance Evaluation Review*, 26(1):151--160.

[21] Benevenuto, F., Duarte, F., Rodrigues, T., Almeida, V., Almeida, J., and Ross, K. (2008a). Understanding video interactions in YouTube. In *Proceedings of the ACM Conference on Multimedia (MM)*, pages 761--764.

[22] Benevenuto, F., Magno, G., Rodrigues, T., and Almeida, V. (2010a). Detecting spammers on twitter. In *Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*.

[23] Benevenuto, F., Pereira, A., Rodrigues, T., Almeida, V., Almeida, J., and Gonçalves, M. (2009a). Avaliação do perfil de acesso e navegação de usuários em ambientes web de compartilhamento de vídeos. In *Proceedings of the Brazilian Symposium on Multimedia Systems and Web (WebMedia)*, pages 149--156.

[24] Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., and Gonçalves, M. (2008b). Detectando usuários maliciosos em interações via vídeos no YouTube. In *Proceedings of the Brazilian Symposium on Multimedia Systems and Web (WebMedia)*, pages 138--145.

[25] Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., and Gonçalves, M. (2009b). Detecting spammers and content promoters in online video social networks. In *Proceedings of the Int'l ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 620--627.

[26] Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., and Gonçalves, M. (2009c). Detecting spammers and content promoters in online video social networks. In *Proceedings of the IEEE Int'l conference on Computer Communications Workshops (INFOCOM)*, pages 337--338.

[27] Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., Gonçalves, M., and Ross, K. (2010b). Video pollution on the web. *First Monday*, 15(4):1--14.

[28] Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., and Ross, K. (2009d). Video interactions in online video social networks. *ACM Transactions on Multimedia Computing, Communications and Applications (TOMCCAP)*, 5(4):1--25.

[29] Benevenuto, F., Rodrigues, T., Almeida, V., Almeida, J., Zhang, C., and Ross, K. (2008c). Identifying video spammers in online social networks. In *Proceedings of the Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, pages 45--52.

[30] Benevenuto, F., Rodrigues, T., Cha, M., and Almeida, V. (2009e). Characterizing user behavior in online social networks. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 49--62.

[31] Binder, J., Howes, A., and Sutcliffe, A. (2009). The problem of conflicting social spheres: effects of network structure on experienced tension in social network sites. In *Proceedings of the ACM SIGCHI Conference on Human factors in Computing Systems (CHI)*, pages 965--974.

[32] Boll, S. (2007). Multitube–where web 2.0 and multimedia could meet. *IEEE MultiMedia*, 14(1):9–13.

[33] Boyd, D. (2007). *Why Youth (Heart) Social Network Sites: The Role of Networked Publics in Teenage Social Life*. Cambridge, MA.

[34] Boyd, D. and Ellison, N. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1-2).

[35] Braitenberg, V. and Schüz, A. (1998). *Cortex: Statistics and Geometry of Neuronal Connectivity*. Springer-Verlag.

[36] Breslau, L., Cao, P., Fan, L., Phillips, G., and Shenker, S. (1999). Web caching and zipf-like distributions: Evidence and implications. In *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, pages 126--134.

[37] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107--117.

[38] Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the web. *Computer Networks*, 33:309--320.

[39] Burke, M., Marlow, C., and Lento, T. (2009). Feed me: Motivating newcomer contribution in social network sites. In *Proceedings of the ACM SIGCHI Conference on Human factors in Computing Systems (CHI)*, pages 945--954.

[40] Castillo, C., Donato, D., Gionis, A., Murdock, V., and Silvestri, F. (2007). Know your neighbors: Web spam detection using the web topology. In *Proceedings of the Int'l ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 423--430.

[41] Caverlee, J. and Webb, S. (2008). A large-scale study of myspace: Observations and implications for online social networks. In *Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM)*.

[Cha et al.] Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proceedings of the Int'l AAAI Conference on Weblogs and Social Media (ICWSM)*.

[42] Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.-Y., and Moon, S. (2007). I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement (IMC)*, pages 1--14.

[43] Cha, M., Mislove, A., Adams, B., and Gummadi, K. (2008). Characterizing social cascades in flickr. In *Proceedings of the ACM SIGCOMM Workshop on Social Networks (WOSN)*, pages 13--18.

[44] Cha, M., Mislove, A., and Gummadi, K. (2009). A measurement-driven analysis of information propagation in the Flickr social network. In *Proceedings of the World Wide Web Conference (WWW)*, pages 721--730.

[45] Chapman, C. and Lahav, M. (2008). International ethnographic observation of social networking sites. In *Proceedings of the ACM SIGCHI Conference on Human factors in Computing Systems (CHI)*, pages 3123--3128.

[46] Chau, D., Pandit, Wang, S., and Faloutsos, C. (2007). Parallel crawling for online social networks. In *Proceedings of the World Wide Web Conference (WWW)*, pages 1283--1284.

[47] Choudhury, M., Sundaram, H., John, A., and Seligmann, D. (2007). Contextual prediction of communication flow in social networks. In *Proceedings of the IEEE/WIC/ACM Conference on Web Intelligence (WI)*, pages 57--65.

[48] Choudhury, M., Sundaram, H., John, A., and Seligmann, D. (2008). Multi-scale characterization of social network dynamics in the blogosphere. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*, pages 1515--1516.

[49] Choudhury, M., Sundaram, H., John, A., and Seligmann, D. (2009a). Social synchrony: Predicting mimicry of user actions in online social media. In *Proceedings of the IEEE Conference on Social Computing 2009 (SocialCom)*, pages 151--158.

[50] Choudhury, M., Sundaram, H., John, A., and Seligmann, D. (2009b). What makes conversations interesting? themes, participants and consequences of conversations in online social media. In *Proceedings of the World Wide Web Conference (WWW)*, pages 331--340.

[51] Chun, H., Kwak, H., Eom, Y., Ahn, Y.-Y., Moon, S., and Jeong, H. (2008). Comparison of online social relations in volume vs interaction: a case study of Cyworld. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 57--70.

[52] Cormode, G. and Krishnamurthy, B. (2008). Key differences between web 1.0 and web 2.0. *First Monday*, 13(6).

[53] Crane, R. and Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences (PNAS)*, 105(41):15649--15653.

[54] Dale, X. and Liu, C. (2008). Statistics and social network of YouTube videos. In *Proceedings of the Int'l Workshop on Quality of Service (IWQoS)*.

[55] Duarte, F., Benevenuto, F., Almeida, V., and Almeida, J. (2006). Locality of reference in an hierarchy of web caches. In *Proceedings of the IFIP Networking Conference (Networking)*, pages 344--354.

[56] Duarte, F., Benevenuto, F., Almeida, V., and Almeida, J. (2007a). Geographical characterization of YouTube: a latin american view. In *Proceedings of the Latin American Web Congress (LAWEB)*, pages 13--21.

[57] Duarte, F., Mattos, B., Almeida, J., Almeida, V., Curiel, M., and Bestavros, A. (2008). Hierarchical characterization and generation of blogosphere workloads. Technical report, Boston University, Computer Science Dept.

[58] Duarte, F., Mattos, B., Bestavros, A., Almeida, V., and Almeida, J. (2007b). Traffic characteristics and communication patterns in blogosphere. In *Proceedings of the Conference on Weblogs and Social Media (ICWSM)*.

[59] Ebel, H., Mielsch, L., and Bornholdt, S. (2002). Scale free topology of e-mail networks. *Physical Review E*, 66(3):35103.

[60] Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999). On power-law relationships of the Internet topology. In *Annual Conference of the ACM Special Interest Group on Data Communication (SIGCOMM)*, pages 251--262.

[61] Fan, R., Chen, P., and Lin, C. (2005). Working set selection using the second order information for training SVM. *Journal of Machine Learning Research (JMLR)*, 6:1889--1918.

[62] Fetterly, D., Manasse, M., and Najork, M. (2004). Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. In *Proceedings of the Workshop on the Web and Databases (WebDB)*.

[63] Gabrilovich, E., Dumais, S., and Horvitz, E. (2004). Newsjunkie: Providing personalized newsfeeds via analysis of information novelty. In *Proceedings of the World Wide Web Conference (WWW)*, pages 482--490.

[64] Garlaschelli, D. and Loffredo, M. (2004). Patterns of link reciprocity in directed networks. *Physical Review Letters*, 93(26):268701.

[65] Garriss, S., Kaminsky, M., Freedman, M., Karp, B., Mazières, D., and Yu, H. (2006). Re: Reliable email. In *Proceedings of the USENIX Conference on Networked Systems Design & Implementation (NSDI)*, pages 22--22.

[66] Gilbert, E. and Karahalios, K. (2009). Predicting tie strength with social media. In *Proceedings of the ACM SIGCHI Conference on Human factors in Computing Systems (CHI)*, pages 211--220.

[67] Gill, P., Arlitt, M., Li, Z., and Mahanti, A. (2007). YouTube traffic characterization: A view from the edge. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement (IMC)*, pages 15--28.

[68] Gill, P., Arlitt, M., Li, Z., and Mahanti, A. (2008). Characterizing user sessions on YouTube. In *IEEE Multimedia Computing and Networking (MMCN)*.

[69] Gjoka, M., Sirivianos, M., Markopoulou, A., and Yang, X. (2008). Poking facebook: Characterization of OSN applications. In *Proceedings of the ACM SIGCOMM Workshop on Social Networks (WOSN)*, pages 31--36.

[70] Golder, S., Wilkinson, D., and Huberman, B. (2007). Rhythms of social interaction: Messaging within a massive online network. In *Proceedings of the Conference on Communities and Technologies (ICCT)*.

[71] Gomes, L., Almeida, J., Almeida, V., and Meira, W. (2007). Workload models of spam and legitimate e-mails. *Performance Evaluation*, 64(7-8).

[72] Gómez, V., Kaltenbrunner, A., and López, V. (2008). Statistical analysis of the social network and discussion threads in slashdot. In *Proceedings of the World Wide Web Conference (WWW)*, pages 645--654.

[73] Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A. (2004). Information diffusion through blogspace. In *Proceedings of the World Wide Web Conference (WWW)*, pages 491--501.

[74] Gummadi, K., Dunn, R., Saroiu, S., Gribble, S., Levy, H., and Zahorjan, J. (2003). Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In *Proceedings of the ACM Symposium on Operating Systems Principles (SOSP)*.

[75] Guo, L., Tan, E., Chen, S., Zhang, X., and Zhao, Y. E. (2009). Analyzing patterns of user content generation in online social networks. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 369--378.

[76] Halvey, M. and Keane, M. (2007). Analysis of online video search and sharing. In *Proceedings of the ACM Conference on Hypertext and Hypermedia (HT)*.

[77] Heymann, P., Koutrika, G., and Garcia-Molina, H. (2007). Fighting spam on social web sites: A survey of approaches and future challenges. *IEEE Internet Computing*, 11:36–45.

[78] Hsu, C.-W. and Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. In *IEEE Transactions on Neural Networks*, volume 13, pages 415--425.

[79] Huberman, B., Pirolli, P., Pitkow, J., and Lukose, R. (1998). Strong regularities in world wide web surfing. *Science*, 280(5360).

[80] Huberman, B., Romero, D., and Wu, F. (2009). Social networks that matter: Twitter under the microscope. *First Monday*, 14(1).

[81] Jain, A., Murty, M., and Flynn, P. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264--323.

[82] Jain, R. (1991). *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling*. John Wiley and Sons, INC.

[83] Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the Workshop on Web mining and social network analysis (WebKDD/SNA-KDD)*.

[84] Jindal, N. and Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the Int'l Conference on Web Search and Web Data Mining (WSDM)*, pages 219--230.

[85] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning (ECML)*.

[86] Joinson, A. (2008). Looking at, looking up or keeping up with people?: motives and use of Facebook. In *Proceedings of the ACM SIGCHI Conference on Human factors in Computing Systems (CHI)*, pages 1027--1036.

[King] King, R. When your social sites need networking, *BusinessWeek*, 2007. `http://tinyurl.com/o4myvu`. Accessed in March/2010.

[87] Kleinberg, J. (1999). Hubs, authorities, and communities. *ACM Computing Surveys*, 31:5.

[88] Kohavi, R. and Provost, F. (1998). Glossary of terms. *Machine Learning*, 30(2-3):271–274.

[89] Kossinets, G., Kleinberg, J., and Watts, D. (2008). The structure of information pathways in a social communication network. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 435--443.

[90] Krishnamurthy, B. (2009). A measure of online social networks. In *Conference on Communication Systems and Networks (COMSNETS)*.

[91] Kumar, R., Novak, J., Raghavan, P., and Tomkins, A. (2003). On the bursty evolution of blogspace. In *Proceedings of the World Wide Web Conference (WWW)*, pages 568--576.

[92] Kumar, R., Novak, J., and Tomkins, A. (2006). Structure and evolution of online social networks. In *Proceedings of the ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining (KDD)*.

[93] Lee, S., Kim, P., and Jeong, H. (2006). Statistical properties of sampled networks. *Physical Review E*, 73(30):102--109.

[94] Lerman, K. (2007). Social information processing in news aggregation. *IEEE Internet Computing*, 11(6):16--28.

[95] Leskovec, J., Adamic, L. A., and Huberman, B. A. (2007). The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):228--237.

[96] Leskovec, J., Backstrom, L., and Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 497--506.

[97] Leskovec, J. and Horvitz, E. (2008). Planetary-scale views on a large instant-messaging network. In *Proceedings of the World Wide Web Conference (WWW)*.

[98] Li, L., Alderson, D., Doyle, J., and Willinger, W. (2005). Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathematics*, 2(4).

[99] Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., and Tomkins, A. (2005). Geographic routing in social network. In *Proceedings of the National Academy of Sciences (PNAS)*, volume 102, pages 11623--11628.

[100] Mahanti, A., Eager, D., and Williamson, C. (2000). Temporal locality and its impact on web proxy cache performance. *Performance Evaluation Journal*, 42(2-3):187--203.

[MaxMind] MaxMind. GeoIP Database. `http://www.maxmind.com/app/ip-location`. Accessed in March/2010.

[101] Milgram, S. (1967). The small world problem. *Psychology Today*, 2:60--67.

[102] Mishne, G., Carmel, D., and Lempel, R. (2005). Blocking blog spam with language model disagreement. In *Proceedings of the Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*.

[103] Mislove, A. (2009). *Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems*. PhD thesis, Rice University, Department of Computer Science.

[104] Mislove, A., Koppula, H., Gummadi, K., Druschel, P., and Bhattacharjee, B. (2008a). Growth of the flickr social network. In *Proceedings of the ACM SIGCOMM Workshop on Social Networks (WOSN)*, pages 25--30.

[105] Mislove, A., Marcon, M., Gummadi, K., Druschel, P., and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement (IMC)*, pages 29--42.

[106] Mislove, A., Post, A., Gummadi, K., and Druschel, P. (2008b). Ostra: Leverging trust to thwart unwanted communication. In *Proceedings of the Symposium on Networked Systems Design and Implementation (NSDI)*, pages 15--30.

[107] Moore, C. and Newman, M. (2000). Epidemics and percolation in small-world networks. *Physical Review E*, 61(5):5678.

[108] Morik, K., Brockhausen, P., and Joachims, T. (1999). Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring. In *Proceedings of the Conference on Machine Learning (ICML)*, pages 268–277.

[109] Nazir, A., Raza, S., and Chuah, C. (2008). Unveiling facebook: A measurement study of social network based applications. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement (IMC)*, pages 43--56.

[110] Nazir, A., Raza, S., Gupta, D., Chua, C., and Krishnamurthy, B. (2009). Network level footprints of facebook applications. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement (IMC)*, pages 63--75.

[111] Newman, M. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Science (PNAS)*, 98(2):404--409.

[112] Newman, M. (2002). Assortative mixing in networks. *Physical Review E*, 89(20):208701.

[113] Newman, M. (2003). The structure and function of complex networks. *SIAM Review*, 45:167--256.

[114] Newman, M. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Science (PNAS)*, 101(1):5200--5205.

[115] Newman, M. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2):26113.

[116] Newman, M. and Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E*, 68(3):36122.

[117] Onnela, J., Saramäki, J., Hyvönen, J., Szabó, G., Menezes, A., Kaski, K., Barabási, A., and Kertész, J. (2007). Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, 9(6):179--204.

[118] Otterbacher, J. (2009). 'helpfulness' in online communities: a measure of message quality. In *Proceedings of the ACM SIGCHI Conference on Human factors in Computing Systems (CHI)*, pages 955--964.

[119] Pandurangan, G., Raghavan, P., and Upfal, E. (2002). Using pagerank to characterize web structure. In *Proceedings of the Int'l Conference on Computing and Combinatorics (COCOON)*, pages 330--339.

[Report] Report, N. O. Social networks & blogs now 4th most popular online activity, 2009. `http://tinyurl.com/cfzjlt`. Accessed in March/2010.

[120] Rodrigues, T., Benevenuto, F., Almeida, V., Almeida, J., and Gonçalves, M. (2009). Uma análise contextual de conteúdo duplicado no YouTube. In *Proceedings of the Brazilian Symposium on Multimedia Systems and Web (WebMedia)*, pages 141--148.

[121] Rodrigues, T., Benevenuto, F., Almeida, V., Almeida, J., and Gonçalves, M. (2010). Equal but different: A contextual analysis of duplicated videos on youtube. *Journal of the Brazilian Computer Society*, 16(3):201--214.

[Rodriguez] Rodriguez, P. Web infrastructure for the 21st century. *WWW'09 Keynote, 2009*. `http://tinyurl.com/mmmaa7`. Accessed in March/2010.

[122] Sastry, N., Yoneki, E., and Crowcroft, J. (2009). Buzztraq: predicting geographical access patterns of social cascades using social networks. In *Proceedings of the Workshop on Social Network Systems (SNS)*.

[123] Schneider, F., Feldmann, A., Krishnamurthy, B., and Willinger, W. (2009). Understanding online social network usage from a network perspective. In *Proceedings of the ACM SIGCOMM Internet Measurement Conference (IMC)*, pages 35--48.

[Schroeder] Schroeder, S. 20 ways to aggregate your social networking profiles, *Mashable*, 2007. `http://tinyurl.com/2ceus4`. Accessed in March/2010.

[124] Seshadri, M., Machiraju, S., Sridharan, A., Bolot, J., Faloutsos, C., and Leskove, J. (2008). Mobile call graphs: Beyond power-law and lognormal distributions. In *Proceedings of the ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining (KDD)*, pages 596--604.

[125] Shannon, M. (2007). Shaking hands, kissing babies, and...blogging? *Communications of the ACM*, 50(9):21--24.

[126] Szabó, G. and Huberman, B. (2008). Predicting the popularity of online content. *ArXiv e-prints*.

[127] Thom-Santelli, J., Muller, M., and Millen, D. (2008). Social tagging roles: publishers, evangelists, leaders. In *Proceedings of the ACM SIGCHI Conference on Human factors in Computing Systems (CHI)*, pages 1041--1044.

[128] Thomason, A. (2007). Blog spam: A review. In *Proceedings of the Conference on Email and Anti-Spam (CEAS)*.

[129] Torkjazi, M., Rejaie, R., and Willinger, W. (2009). Hot today, gone tomorrow: On the migration of myspace users. In *Proceedings of the ACM SIGCOMM Workshop on Online social networks (WOSN)*, pages 43--48.

[130] Trivedi, K. S. (2002). *Probability and statistics with reliability, queuing and computer science applications*. John Wiley and Sons Ltd., Chichester, UK.

[131] Valafar, M., Rejaie, R., and Willinger, W. (2009). Beyond friendship graphs: a study of user interactions in Flickr. In *Proceedings of the ACM SIGCOMM Workshop on Online Social Networks (WOSN)*, pages 25--30.

[132] Viswanath, B., Mislove, A., Cha, M., and Gummadi, K. P. (2009). On the evolution of user interaction in Facebook. In *Proceedings of the ACM SIGCOMM Workshop on Online Social Networks (WOSN)*, pages 37--42.

[133] Wasserman, S., Faust, K., and Iacobucci, D. (1994). *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press.

[134] Watts, D. (1999). *Small Worlds: the Dynamics of Networks Between Order and Randomness*. Princeton University Press.

[135] Watts, D. (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences (PNAS)*, 99(9):5766--5771.

[136] Weiss, G. and Provost, F. (2001). The effect of class distribution on classifier learning: An empirical study. Technical report, Department of Computer Science, Rutgers University.

[137] Williamson, C. (2002). On filter effects in web caching hierarchies. *ACM Transactions on Internet Technology (TOIT)*, 2(1):47--77.

[138] Wilson, C., Boe, B., Sala, A., Puttaswamy, K. P. N., and Zhao, B. Y. (2009). User interactions in social networks and their implications. In *Proceedings of the ACM European Professional Society on Computer Systems (EuroSys)*, pages 205--218.

[139] Wu, C., Cheng, K., Zhu, Q., and Wu, Y. (2005). Using visual features for anti-spam filtering. In *Proceedings of the IEEE Int'l Conference on Image Processing (ICIP)*.

[140] Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1(1-2):69--90.

[141] Yang, Y. and Pedersen, J. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the Conference on Machine Learning (ICML)*.

[142] Yu, H., Gibbons, P., Kaminsky, M., and Xiao, F. (2008). Sybillimit: A near-optimal social network defense against sybil attacks. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, pages 3--17.

[143] Yu, H., Kaminsky, M., Gibbons, P., and Flaxman, A. (2006). Sybilguard: Defending against sybil attacks via social networks. pages 267--278.

[144] Zhang, J., Ackerman, M., and Adamic, L. (2007). Expertise networks in online communities: Structure and algorithms. In *Proceedings of the World Wide Web Conference (WWW)*, pages 221--230.

[145] Zhou, Y., Guan, X., Zhang, Z., and Zhang, B. (2008). Predicting the tendency of topic discussion on the online social networks using a dynamic probability model. In *Proceedings of the Workshop on Collaboration and Collective Intelligence (WebScience)*.

[146] Zink, M., Suh, K., Gu, Y., and Kurose, J. (2008). Watch global, cache local: YouTube network traces at a campus network - measurements and implications. In *Proceedings of the IEEE Multimedia Computing and Networking (MMCN)*.

[147] Zinman, A. and Donath, J. (2007). Is Britney Spears spam? In *Proceedings of the Conference on Email and Anti-Spam (CEAS)*.