MIS Quarterly

SPECIAL ISSUE: BUSINESS INTELLIGENCE RESEARCH

# BUSINESS INTELLIGENCE IN BLOGS: UNDERSTANDING CONSUMER INTERACTIONS AND COMMUNITIES[1]

**Michael Chau**
School of Business, The University of Hong Kong, Pokfulam, HONG KONG {mchau@business.hku.hk}

**Jennifer Xu**
Computer Information Systems, Bentley University, Waltham, MA 02452 U.S.A. {jxu@bentley.edu}

*The increasing popularity of Web 2.0 has led to exponential growth of user-generated content in both volume and significance. One important type of user-generated content is the blog. Blogs encompass useful information (e.g., insightful product reviews and information-rich consumer communities) that could potentially be a gold mine for business intelligence, bringing great opportunities for both academic research and business applications. However, performing business intelligence on blogs is quite challenging because of the vast amount of information and the lack of commonly adopted methodology for effectively collecting and analyzing such information. In this paper, we propose a framework for gathering business intelligence from blogs by automatically collecting and analyzing blog contents and bloggers' interaction networks. Through a system developed using the framework, we conducted two case studies with one case focusing on a consumer product and the other on a company. Our case studies demonstrate how to use the framework and appropriate techniques to effectively collect, extract, and analyze blogs related to the topics of interest, reveal novel patterns in the blogger interactions and communities, and answer important business intelligence questions in the domains. The framework is sufficiently generic and can be applied to any topics of interest, organizations, and products. Future academic research and business applications related to the topics examined in the two cases can also be built using the findings of this study.*

**Keywords**: Business intelligence, Web mining, blog mining, social networks, design science

## Introduction

There is an explosion of user-generated content on the Web, attributable to the growth in popularity of Web 2.0 applications in recent years. The availability of a wide range of user-friendly Web 2.0 applications allows users to post content on the Web more easily than ever before. Blogs are one of the earliest and most popular Web 2.0 applications. Bloggers can write about almost anything: personal stories, ideas, reviews, opinions, feelings, emotions, etc. They can also form social links with other bloggers by joining groups, usually known as blogrings, based on their shared interests or opinions and by interacting with one another in different manners, such as by subscribing to another blogger, commenting on a blog entry (post), or citing the content of a blog entry. These activities build the interaction relations between bloggers and their readers (Lin and Kao 2010) and form a complex social network, which is often called the blogosphere. Information, ideas, propaganda, and opinions flow and spread in the blogosphere through the interaction and communication between bloggers (Adar and Adamic 2005; Ali-Hasan and Adamic 2007; Gruhl et al. 2004; Kumar et al. 2005; Nahon et al. 2011).

As a result, blogs have become an important type of online media, potentially useful for various types of business intelli-

[1] Hsinchun Chen, Roger Chiang, and Veda Storey were the accepting senior editors for this paper. Ee-Peng Lim served as the associate editor.

gence analysis. For instance, by analyzing the blog contents of its stakeholders (e.g., customers or pressure groups), a company can obtain first-hand knowledge of customers' feedback about its products and services (Liang et al. 2009), about its own or its competitors' brand images (Chau et al. 2009; Pikas 2005), or about what is happening in the external environment (Chung et al. 2005). In addition, by analyzing characteristics and dynamics of blogger communities, it is possible to study the formation, growth, and evolution of online consumer networks and identify new ideas in the blogosphere (Chau and Xu 2007). These insights enable companies and organizations to make better decisions on critical business matters such as investments (O'Leary 2011), marketing (Kozinets et al. 2010), and planning (Lewis 2008).

Studying the linkage and social structure in the blogosphere is an important topic for both researchers and business practitioners. Academically, it is important to study the nature and topology of the social networks of bloggers and compare them with other online social networks. Such studies will reveal their characteristics and help improve our understanding of information flow and dissemination in these networks. In practice, companies can identify the clusters of customers for their products and services and conduct target marketing to these groups. For example, companies can find the most influential people in their consumer networks and devise more effective and efficient marketing strategies accordingly (Yang and Counts 2010; Zhu and Tan 2007).

Although blogs provide considerable potential for business intelligence, two unique characteristics of blogs present major challenges for collecting blog data, evaluating blog content, and analyzing the underlying social networks. First, blogs are dynamic and are frequently updated. Contents and linkages can be added or removed any time. Second, bloggers have their own styles of linking to each other. These linkages, which represent the interactions between bloggers, are different from traditional hyperlinks between Web documents.

Consequently, automated techniques are needed to collect and analyze the sheer volume of blog data in order to have a good understanding and make effective use of the underlying information and structure. Most previous studies have focused on traditional forms of online content such as Web pages or forum posts (Abbasi and Chen 2008; Chen et al. 2001; Cooley et al. 1997; Sack 2000; Viegas and Smith 2004). These studies have shown that automated analysis and visualization systems are very useful for obtaining a quick understanding of the contents and social interactions in online communities.

This research intends to tackle the challenges of generating business intelligence based on blogs. In this paper, we present our design of a framework and a system for content and social network analysis of blogs. Following the design science methodology described in Hevner et al. (2004), we incorporate automated data collection, content analysis, and social network analysis of blogs in our design.

The rest of the paper is organized as follows. We first review the characteristics of blogs and the blogosphere and their potential value for business intelligence. We also discuss the importance of conducting content and network analysis on blogs and related techniques. The following section introduces the proposed framework, and we discuss how the framework was used to guide our design of the blog mining system, which is then presented. To provide a proof-of-concept evaluation for our framework, we present two case studies in which we applied the framework for business intelligence and report our findings. We conclude our research and suggest some future research directions in the final section.

## Blogs and Blog Content Analysis

Blogs in the early days were primarily Web pages containing links to other useful resources and were usually maintained manually (Blood 2004). When free blog software and blog hosting sites became widely available, the number of blogs grew significantly. People use personal blogs to record their daily lives and express their opinions and emotions (Gill et al. 2009; Nardi et al. 2004). Some corporations and organizations create and maintain corporate blogs to interact with their customers, suppliers, and other stakeholders (Liang et al. 2009; Tsai et al. 2007). For example, Microsoft has created blogs for MSDN to inform developers about the company's latest developments.

Early research on blogs has focused on studying the characteristics of blogs and bloggers, such as the demographics of bloggers (Adar and Adamic 2005; Ali-Hasan and Adamic 2007; Gruhl et al. 2004; Kumar et al. 2005), blogging behavior (Nardi et al. 2004), or the blogging process (Blood 2004). To extract valuable knowledge from blogs, various data and text mining techniques have been proposed to collect and analyze blog contents. Different models have been proposed for identifying blog topics (Agarwal et al. 2010; Kumar et al. 2010; Tsai 2011) and opinions and sentiments expressed in blogs written in English (Abbasi et al. 2008) and non-English languages (Bautin et al. 2008; Feng et al. 2009). The Text Retrieval Conference (TREC) has organized a blog track and attracted researchers' interest in blog content analysis (Macdonald et al. 2010). TREC has created two large blog corpuses, and various classification-based and lexicon-based techniques have been proposed for finding opinions and sentiments relevant to a given topic using these corpuses (e.g., He et al. 2008; Lee et al. 2008; Zhang et al. 2009).

## Blogger Communities and Social Network Analysis

Bloggers are connected in various ways such as subscriptions, comments, and citations, forming networks of bloggers. Many blogger communities, which can be categorized into explicit communities and implicit communities, exist in the blogosphere. Explicit communities are often called blogrings or groups. Most blog hosting sites allow bloggers to create a new group or join any existing groups. In contrast, implicit communities are not explicitly defined as groups or blogrings, but are formed organically by the interactions among bloggers. For instance, a blogger may subscribe to another blog, hoping to get notifications when the subscribed blog is updated. A blogger can also post a link to or leave comments on another blog. These connections signify the social interactions among bloggers. Because such interactions are rather different from simple hyperlinks between Web pages, these blogger communities, which involve social interactions between online users and are characterized by memberships, sense of belonging, relationships, shared values and practices, and self-regulation (Erickson 1997; Roberts 1998), are more similar to *virtual communities* of users than to the traditional *cyber communities* of Web documents (Kumar et al. 1999).

An online survey (Ali-Hasan and Adamic 2007) revealed that different types of relationships between blogs have different characteristics and play different roles in facilitating interactions between bloggers. By analyzing these relationships, the hidden *social structure* that may represent the real social relationships between bloggers can be extracted (Tang et al. 2012; Tang et al. 2009).

Social network analysis (SNA) is a sociological methodology (Wasserman and Faust 1994) that can be used to reveal patterns of relationships and interactions and discover the underlying social structure in the blogger communities. In the following, we will review the three major types of analyses in SNA, namely topological analysis, centrality analysis, and community analysis.

*Topological analysis* is used to find the structural properties of a network, which is often represented by a set of nodes connected by links. Some widely used statistics, such as the average shortest path length, efficiency, clustering coefficient, and degree distribution, can be used to characterize the network (Albert and Barabási 2002; Crucitti et al. 2003). Three models have been proposed to characterize the overall topology of a network, namely, random graph model (Bollobás 1985), small-world model (Watts and Strogatz 1998; Xu and Chau 2006), and scale-free model (Barabási and Albert 1999). Different network topologies have different implications for the functions of a network (Albert and Barabási 2002).

*Centrality analysis* aims to find the key nodes in a network. Central nodes often play an important role by providing leadership or bridging different communities. Traditional centrality measures such as degree, betweenness, and closeness can be used (Freeman 1979). In the context of blogs, degree centrality, which is defined as the number of direct interactions a blogger has made, measures how active a particular blogger is. “Popular” bloggers with high degree scores are the leaders, experts, or hubs in a blogger network. Betweenness centrality measures the extent to which a blogger lies between other bloggers in a network. The betweenness of a blogger is defined as the number of geodesics (shortest paths between two nodes) passing through it. Bloggers with high betweenness scores often serve as bridges and brokers between different communities. They are important communication channels through which information is spread. Closeness centrality is the sum of the length of geodesics between a particular blogger and all other bloggers in a network. A blogger with low closeness may find it very difficult to communicate with other bloggers in the network. Such nodes are thus more peripheral and can become outliers in the network (Xu and Chen 2005).

*Community analysis* is intended to identify implicit communities in social networks. A subset of nodes in a network is considered a community or a social group if nodes in this group have denser links with nodes within the group than with nodes outside the group (Wasserman and Faust 1994). Community analysis finds implicit communities in a network by maximizing within-group link density while minimizing between-group link density. In the context of blogger networks, these implicit communities represent the real interactions (e.g., subscription and comment) between the bloggers and may reveal more important business intelligence information than the explicit groups. Researchers have proposed techniques to detect implicit communities of bloggers. Lin et al. (2006) defined blog communities based on mutual awareness and extracted them using a PageRank-based algorithm. Bulters and de Rijke (2007) utilized both link and content information of blogs to identify communities. As community detection can be seen as a graph problem, graph-based algorithms have also been developed to find the experts and communities in blogs (e.g., Lakshmanan and Oberhofer 2010; Liu et al. 2011).

In addition, as relationships among blogs can enable and facilitate various network functions and processes (e.g., information dissemination, innovation diffusion, and knowledge sharing) in the blogosphere (Adar and Adamic 2005; Gruhl et al. 2004), it is important to study how a blogger network's structural properties affect the outcomes of these processes. Papagelis et al. (2009) proposed a model to study the information dissemination in the blogosphere and the effects of different factors on the diffusion process. Some studies have attempted to identify the bloggers who are the most important in the information dissemination process using graph algorithms (Agarwal et al. 2008; Mathioudakis and Koudas 2009).

## A Design Science Approach

Our objective in this research is to design, implement, and apply a framework for generating business intelligence based on blog data. We adopt the design science methodology. In this section, we present the design of our framework and system (i.e., the artifacts that address the problem of business intelligence analysis based on blogs). Hevner et al. (2004) provided seven guidelines for conducting effective and high-quality design science research in the field of information systems. It is suggested that these guidelines be followed closely to ensure that the research process and outcome are scientific. In the following, we discuss how our current research has followed and addressed these guidelines.;

- *Design as an Artifact*: Both our framework and our system are artifacts for addressing the problem, which is to gather and generate business intelligence from consumer blogs. The artifacts can be applied to different domains, and future research and applications can be built upon them.

- *Problem Relevance*: As discussed earlier, business intelligence based on user-generated contents is highly useful for decision making at various managerial levels in organizations in the current fast-evolving business environment. Analyzing business intelligence can reveal valuable information and discover novel knowledge critical to the success of a business (Abbasi et al. 2008; Chau et al. 2007; McGonagle and Vella 1999).

- *Design Evaluation*: A design must be evaluated in order to show its usefulness and quality. In this research, we use an observational evaluation method to evaluate the design (Hevner et al. 2004). In particular, we conduct two case studies, which will be reported in later sections, as a proof-of-concept to demonstrate the feasibility of our approach and the value of the design (Albert et al. 2004).

- *Research Contributions*: The main contributions of this research are twofold. First, we demonstrate the feasibility and usefulness of applying our framework to blog mining for business intelligence. We investigate how content analysis and social network analysis can reveal useful information for business. We have created two artifacts, namely the business intelligence analysis framework and the blog mining system. Second, by conducting the two case studies, we improved our understanding of the characteristics and networks of the blogs on a consumer product and a company, and reported some interesting findings.

- *Research Rigor*: This research relies on rigorous elements from multiple academic fields, including business intelligence, marketing, information retrieval, social network analysis, Web mining, and system design. Both the construction and evaluation of the artifact are based on the knowledge base from these fields.

- *Design as a Search Process*: Our framework design and the application of techniques is a scientific process in which we searched for a potential solution to address the problem of generating business intelligence by mining blog contents and blogger community structure. In the early stages of our research, we obtained initial feedback from users and the design was revised a number of times. We iteratively revised the design in order to search for the best artifact that served our purpose.

- *Communication of Research*: We present the research in this paper to both technology-oriented and management-oriented audiences. Both the artifacts and the evaluation study are presented in this paper in the following sections, such that both can be easily replicated by researchers or practitioners.

## The Framework for Analyzing Business Intelligence in Blogs

In this section, we present our proposed framework for conducting business intelligence collection and analysis of blogs on a topic of interest, such as a consumer product or an organization. Our framework, shown in Figure 1, consists of the following steps (components):

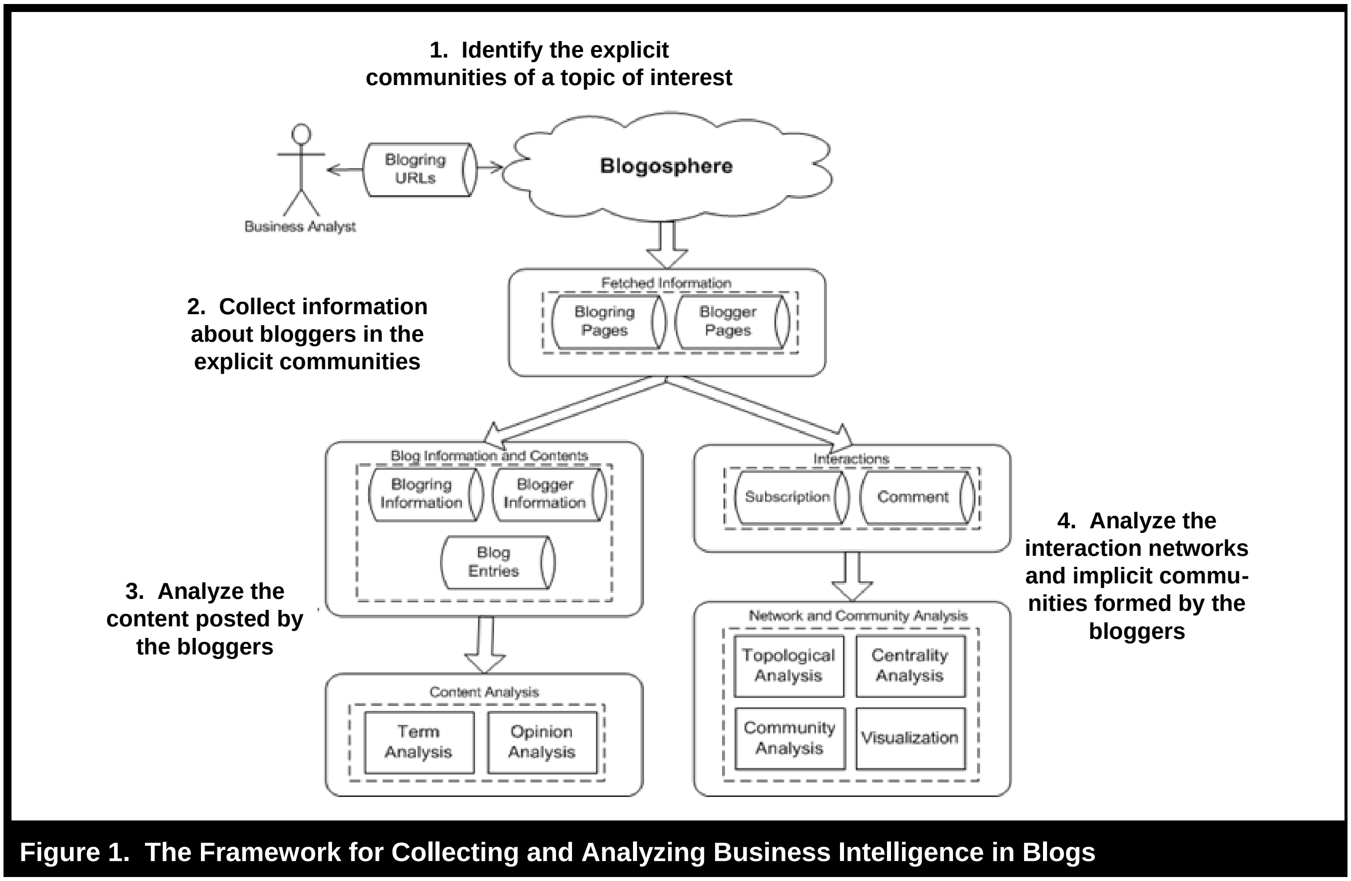
1. Identify the explicit communities of a topic of interest
Business Analyst
Blogring URLs
Blogosphere
2. Collect information about bloggers in the explicit communities
Fetched Information
Blogring Pages
Blogger Pages
Blog Information and Contents
Blogring Information
Blogger Information
Blog Entries
Interactions
Subscription
Comment
3. Analyze the content posted by the bloggers
4. Analyze the interaction networks and implicit communities formed by the bloggers
Content Analysis
Term Analysis
Opinion Analysis
Network and Community Analysis
Topological Analysis
Centrality Analysis
Community Analysis
Visualization

**Figure 1. The Framework for Collecting and Analyzing Business Intelligence in Blogs**

1. Identify the explicit communities of a topic of interest.
2. Collect information about bloggers in the explicit communities.
3. Analyze the content posted by the bloggers.
4. Analyze the interaction networks and implicit communities formed by the bloggers.

In the remainder of this section, we describe each step in this framework.

### Identify the Explicit Communities of a Topic of Interest

After deciding on the topic of interest, one can begin finding the explicit communities on this topic in the blogosphere. These explicit communities are represented by blogrings or interest groups, which are searchable on some blog hosting sites. These communities' information and membership can be retrieved either manually or using a software program such as a Web spider or crawler, depending on the number of communities. Each community can be manually examined for its relevance, authenticity, validity, and suitability to ensure data quality. This involves reading the description and sampled contents of these explicit communities and selecting the ones to be retrieved and analyzed. The communities can also be manually classified according to their characteristics, such as their attitudes toward the topic of interest.

### Collect Information about Bloggers in the Explicit Communities

After the set of explicit communities has been examined, one can gather their member lists and collect a massive amount of data about these members such as profiles, blog entries, and interaction patterns. Unless the number of members is very small, it is nearly impossible to complete this task manually. A blog spider program can be employed to automate this task. The blog spider starts by collecting the description page and retrieving the list of members of the explicit communities. The bloggers' URLs are then extracted and stored into a queue for fetching. The blog spider can be designed to follow only links that are of interest, such as blogger profile pages, blog entry pages, and comment pages, and to exclude pages such as online advertisements. If the blog hosting site supports RSS (really simple syndication), it is possible for the blog spider to easily retrieve blog information and contents in the form of Web feeds. Similar to standard Web spiders, multithreading or asynchronous I/O can also be used such that multiple blog pages can be downloaded in parallel (Chau et al. 2005). This can avoid bottlenecking the process if a particular Web server is sending a malicious response or not responding at all. After a page is downloaded, it can be stored into a relational database or as a simple file. The spider can terminate when a specific number of blog entries have been retrieved or when the data of all bloggers of interest have been collected.

### Analyze the Content Posted by the Bloggers

A downloaded blog page has to be processed to extract useful information. Blogs may be downloaded in HTML or XML format, depending on the blog hosting site. As a blog page may consist of more than one blog entry, the page is first parsed into separate blog entries. This can be done by simple string matching techniques. For example, some HTML formatting tags can be used to identify the beginning of a particular blog entry or comment, depending on the format of the blogs being analyzed. Useful information is then extracted, including the blogger's age, gender, country of residence, and the blog creation date. As blogs, even those hosted on the same site, may have different layouts, it is not trivial to extract such information from blogs in HTML format. For blogs in XML format, usually only partial information is available. Fortunately, some standard data like blogger name and blog entries are often put into specific formats (e.g., as a sidebar or in a table) in the HTML files in large blog hosting sites, and simple rules are often sufficient. Text analysis and Web content mining algorithms, such as linguistic analysis, text classification, or text clustering, can then be applied on the blog entries. Simple analysis includes term frequency analysis and extracting sentences that contain particular keywords of interest. Other text mining techniques such as topic and opinion analysis can also be applied on the contents collected (Abbasi et al. 2008).

### Analyze the Interaction Networks and Implicit Communities Formed by the Bloggers

The profile page and blog entry pages downloaded contain traces of interactions between bloggers. These interactions can be found in different sections of a blog page. For example, subscription links are often located in the blogroll on the left sidebar of a page; comment links are found in the comment section of a blog entry. These links can be extracted from the HTML blog page based on simple pattern matching. Based on all of the interaction links, one can automatically construct the networks formed by these links using software programs. Business intelligence information can be revealed by conducting social network analysis on these networks. Topological, centrality, and community analysis, as discussed earlier, can be applied to these networks to find useful, novel patterns. SNA statistics can be automatically calculated and visualization programs can be used to display a graphical notation of the networks. All of the analysis results can then be manually interpreted by business analysts. Depending on the purpose of the analysis, this may involve studying the profiles of selected bloggers of interest, investigating their link structures, and reading their blog entries.

## Case Studies

Two case studies are presented as a proof-of-concept in applying our framework. In the case studies, we show how our framework can be applied to collect and analyze the characteristics and structural properties of consumer communities in blogs and help generate business intelligence. Apple's iPod music player and Starbucks are chosen as the topics of our two case studies, which are discussed in detail in the following sections..

In the case studies we demonstrate how we applied the proposed framework to study the following important questions for business intelligence:

1. What are the characteristics of the contents of the consumer blogs? How are they related to the product or company of interest?
2. What are the characteristics of the interaction networks of bloggers?
3. Who are the central bloggers in these networks? Are these bloggers effective in disseminating information?
4. Do the implicit communities formed by different types of interactions demonstrate different properties? Which types of interactions are more important in shaping the communities?

## Case Study 1: iPod

The topic we selected for the first case study was Apple's iPod music player. We chose this product as the topic of interest due to its popularity with young people, who are major bloggers. Analyses of such consumer communities in blogs may provide relevant companies and organizations important insights into the characteristics of their current and potential consumers (Baker and Green 2005; Chevalier and Mayzlin 2006; Kozinets et al. 2010) and help them better market their iPod-related products and services.

### Data Set

We chose Xanga (www.xanga.com) as our source of blog data. According to Alexa,[2] Xanga is the second most popular blog hosting site after the Google-owned Blogger (www.blogger.com). It is also ranked 17th in traffic (visit popularity) among all Web sites in English. Xanga was chosen over Blogger because Xanga has more prominent features to support subscriptions and groups, and these features are useful for identifying consumer groups in the blogs and the interactions between bloggers.

We used the onsite search engine to manually identify the online iPod consumer communities in blogs. We first searched for all the blogrings on Xanga that contained the word "iPod" in their titles or descriptions and retrieved 315 blogrings (groups). We then manually examined the details of these blogrings and those that were irrelevant to iPods or invalid were discarded. Groups with only a single member, generally formed by one blogger with no one else joining, were also removed from our list. Our final data set consisted of 204 valid groups. For each group, we read its group description and classified it as having a positive, negative, or neutral attitude toward iPods. The top 20 largest groups are shown in Table 1.

There were 3,493 bloggers in total in this data set. Each blogger maintained one blog, which may contain multiple blog entries. In total there were 75,445 blog entries. Our system automatically fetched and extracted the bloggers' basic information, their blog entries, and their relationships. Because all blog pages were from the same blog host, we used a program based on some simple pattern matching rules (e.g., based on occurrences of some particular HTML tags or headings) to extract the required information from these pages. The basic information about a blogger included the user ID, name, date of birth, city, state, country, and date of registration. Among the 3,493 bloggers, 2,603 indicated their gender. Although these self-reported data may not be very reliable, they provide a rough picture of the sample of bloggers in the blogging consumer groups. The attitude of each blogger toward iPods was also determined based on the attitude of the groups to which the blogger belonged. In our data we found 2,377 bloggers with a positive attitude toward iPods, 225 with a negative attitude, and 891 were neutral.

### Content Analysis

We examined the contents of the blogs to ascertain whether and to what degree they were relevant to iPod, our topic of interest. As a preliminary analysis, we measured the relevance by looking at the number of times (word frequency) that the word iPod was mentioned in each collected blog. We found that the blog with the highest frequency mentioned the word iPod 345 times. However, after careful examination of this blog, we determined that this blog was a splog, which is a type of spam blog used to trick search engines and artificially boost the traffic to other Web sites.

After filtering out similar splogs, we found that the highest word frequency in the legitimate blogs was 115 while the lowest was 0. We plot the percentage of bloggers (i.e., blogs) in logarithm scales against the word frequency in Figure 2. In the chart, we can see that a large percentage of bloggers mentioned the word iPod sparingly, if at all, in their blogs, while only a small percentage of bloggers used the product's name frequently.

We found that 1,573 bloggers mentioned the word iPod at least once in their blogs. The word iPod appears 10,572 times in total in the postings of these bloggers. On the other hand, 1,920 bloggers never mentioned the word iPod in their blogs, although they joined at least one of the iPod-related blogrings identified in our study. This represents more than half (55.0%) of all the bloggers in our data set. This finding is intriguing because it implies that it is not possible to reach these bloggers through standard keyword-based searches (e.g., searching the word iPod in a blog search engine like Technorati or Google Blog Search). These bloggers can only be identified by their group memberships or other approaches.

We further examined the content of the blogs by extracting all the sentences that contain the word iPod. We sampled a random set of 300 of these sentences. Out of this set, 296 sentences talk about the blogger's interaction with an iPod and are neutral toward iPod (e.g., "I want an ipod," "I spent

[2] "Top English Language Sites" (http://www.alexa.com/site/ds/ top_sites?ts_mode=lang&lang=en; accessed June 14, 2008).

**Table 1. The Top 20 Largest Groups for iPod**

| Group Title | Number of Members | Group Description | Category |
|---|---|---|---|
| I Can't Live Without My iPod! | 677 | You can't live without your iPod? You bring your iPod to wherever you go? Do you listen to your iPod when you take a sh*t? You feel weird and naked whenever your iPod is not with you? Do you feel like you can't function normally without your iPod? Are you crazy about your iPod? You're not alone. Join! | Positive |
| i <3 my iPod | 422 | as if the novelty will ever wear off… | Positive |
| ! ! iPod Supremacy ! ! | 385 | Own an ipod? JOIN NOW!!! RIGHT NOW!!! all ipods welcome. Old school , New, iPod minis , etc. No we do not execpt the poser iRiver (which sony poses in the iPods place) | Neutral |
| I heart my iPod | 206 | Being a part of this blogring means that not only do I own an iPod, but I also heart my iPod too. | Positive |
| My iPod owns your mp3 player | 170 | Yes We own iPods, there awesome and people envy them saying there Dell DJ or Creative Zen is better, but its not. So Join if you own any ipod. ok? iPod Ipod ipod they are the best. ipod ipod ipod ipod ipod ipod ! | Positive |
| ~~~iPoD~~LoVe~~~ | 75 | ThIs Is 4 AnY1 wHo HaS aN iPoD or WaNtS oNe yeaaaaaa!!!!!!! wE roCk!!!!! I LOVE IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS IPODS | Positive |
| !!**Mini Pink Ipods** | 67 | I want 100 people with pink mini ipods ( orwho want pink mini ipods, or who has an ipod at all) to join this blog!!! Common! THINK PINK! | Neutral |
| i have an ipod | 67 | ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod, ipod,ipod | Neutral |
| I love my ipod!! | 62 | For those who can't get enough listening to their ipod. Whether its a colorful mini or a white 20 gb. we love our ipods! so go get your own and stop looking at mine! get updates and other cool sh*t for ur ipod. | Positive |
| my iPod owns me ♥ | 57 | + iF Y0U CAN'T G0 ANYWHERE WiTH0UT Y0UR iP0D ;; LiSTEN T0 iT N0N-ST0P /// THiS iS DEF. F0R Y0U :] -- J0iN N0W !* | Positive |
| I own an iPod but I'm not rich. | 51 | this blog is for anyone who owns an iPod and knows people who think they are rich since they do have one. but i own an ipod and im not rich! you dont see me walking around in a huge mansion riding in a limo! im not rick..but i do own an iPod. | Neutral |
| i love my ipod | 51 | it is freaking awesome. | Positive |
| .:I Heart my iPod:. | 46 | : : : : Ode to my iPod : : : : : : I love my iPod, yes I do. : : : : I love my iPod, how bout you? : : : : : : : : : : : : : : : : : : : My iPod is my very best friend. : : : : : : : : : : : : : : : : : : Without it I would come to an end. : : : : : : : : : : : : : So, if your iPod makes you sing, : : : : : : : : : : : : : : You should join this blogring! | Positive |

**Table 1. The Top 20 Largest Groups for iPod (Continued)**

| Group Title | Number of Members | Group Description | Category |
|---|---|---|---|
| My MP3 player kicks your iPod's BUTT! | 35 | iPods=crap We are all smart people who bought MP3 players that won't break after 2 months. ahahaha, I laugh at people who own iPods. | Negative |
| iPod Users | 33 | Whether you own a 1st/2nd/3rd Generation Touch-Wheel Apple iPod, 4th Generation Click-Wheel or Apple iPod Mini MP3 player! Join and unite under one blogring as a faithful community to the BEST MP3 player constructed EVER!! SPREAD THIS BLOGRING!! | Positive |
| !!iPod Mini Supremacy!! | 28 | its for people who love there iPod mini. so enjoy. | Positive |
| i luff my iPOd MiNi <3 | 24 | iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi iPOd MiNi just rock. ;] any color, any songs, any kinda ipod, just join cuz you luff iPOd! | Positive |
| ~iPoD OwnERZ~ | 23 | anyone tht havs and liks their ipod join this blog ring~ | Positive |
| iPod's are EVIL | 20 | IPODS R EVIL! Nearly all the otha mp3 playas cood kick iPods @$$ easily! Most ipoders cant even name 5 dif mp3's, they don't even research be4 jumping to iPod! iPod doesnt: come in as many colors as zen, or hav voice recording lik almost all the others do, come w/ a charger or a case. Its bigger than the rio, or zen(w/ the same amount of gigs). It: only has one kinda case, takes longer to charge+has less battery life, costs more+u get less gigs. Oh, + THEY ALL HAV LIL DEMONS IN THEM. iPods SUCK! | Negative |
| I <3 my hawt iPod | 19 | ipod owners! this is the place for you!!! | Positive |

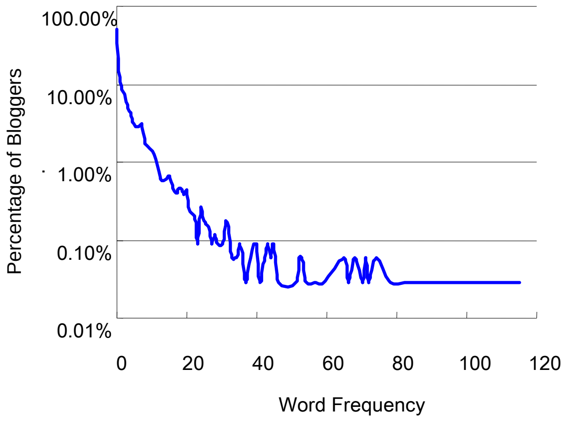

**Figure 2. Percentage of Bloggers (in Log Scale) Versus Word Frequency in the iPod Data Set**

**Table 2. The Statistics of the Interaction Networks in the iPod Data set**

| Network | Subscription | Comment | Combined |
|---|---|---|---|
| Number of Nodes | 1,103 | 1,108 | 1,282 |
| Number of Links | 948 | 1,028 | 1,302 |
| Average Degree | 1.72 | 1.86 | 2.38 |
| Highest In-Degree | 7 | 16 | 23 |
| Highest Out-Degree | 9 | 26 | 26 |

5 hours listening to my ipod," "bought a new ipod case," "I just downloaded some new songs on my iPod," and "my ipod was left on the bus!"). These sentences provide some information on how users interact with their iPods, and it is often useful for analysts to keep track of this type of posting (Pikas 2005). On the other hand, only 4 of the 300 sentences explicitly showed the blogger's review or opinion of the iPod. Examples of these sentences include "ipods are awesome," "we decided that ipods r cool," and "I hate ipods." Given the small percentage of sentences, we observed that the data set did not provide enough cases for conducting opinion mining, which is often used for other types of user-generated content (e.g., online customer reviews; Glance et al. 2005; Liu et al. 2005). Our results indicated that opinion mining may be more appropriate and feasible for review-oriented blogs, which provide consumer evaluations and opinions of particular products, but less suitable for personal blogs, which focus on diaries and personal content (Ip and Wagner 2008).

## Interaction Networks

The interaction networks are the networks constructed by the interaction relationships between bloggers extracted from the collected blogs. As mentioned in the previous section, we identified two types of interactions among bloggers: subscription, which occurs when one blogger subscribes to another blog, and comment, which occurs when one blogger makes a comment on another person's blog entry. Blogger interactions are not constrained by one single type of activity. It is thus natural to study the network combining both subscription and comment relationships; therefore, we analyzed the following three networks of bloggers: (1) subscription network, (2) comment network, and (3) a combined network of subscriptions and comments. Table 2 presents the basic statistics of these interaction networks. Note that none of these three networks contains all of the 3,493 bloggers in our data set. This is because many bloggers have neither subscription nor commenting relationships with anyone else in this particular data set. They are isolated nodes and are not included in the networks.

We also studied the characteristics of the cumulative degree distributions of the networks. The cumulative degree distribution, $P(k)$, is defined as the probability that an arbitrary node in the network has at least $k$ links (Albert and Barabási 2002). We found that all these networks show power–law degree distributions, indicating scale-free topology (subscription: $R^2 = 0.94$; comment: $R^2 = 0.92$; combined: $R^2 = 0.90$). As an example, we plotted the cumulative degree distributions of the combined network in logarithmic scales in Figure 3. The curve for the network is roughly straight with only a small bump. The power–law degree distribution indicates that a small number of bloggers are involved with a larger number of interactions and they are the "centers" of the network.

## Central Bloggers

It is important to identify the central bloggers who play important roles in information dissemination in the implicit communities. We used three popular centrality measures, namely degree, betweenness, and closeness (Freeman 1979), to identify the key bloggers.

We first examined the degree centrality of the bloggers. In the three types of networks, each blogger has three degrees: in-degree (the number of incoming links), out-degree (the number of outgoing links), and degree (the total number of links). A blogger with a high in-degree usually is popular or "authoritative" (Kleinberg 1999). For example, a large number of incoming subscription links implies that the blogger is somehow liked or endorsed by others who subscribe to his/her blog. A blogger becomes popular if the blog content is interesting and many people like to keep track of the updates. We found that the highest value of in-degree in the subscription network was 7 (see Table 2). The high out-degree subscribers, in contrast, may not be popular. Instead they may be "hubs" who can direct their visitors to many other interesting, popular blogs, and thus are also quite important to identify. The highest out-degree value is 9 in this network.

Similarly, in the comment network, a blogger with a high in-degree is one who attracts comments from many other bloggers, and a blogger with a high out-degree is one who

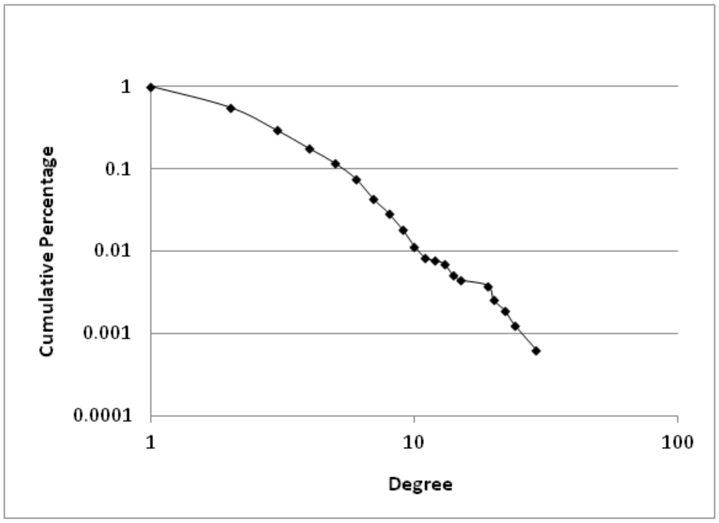

**Figure 3. Cumulative Degree Distributions of the Combined Network in the iPod Data Set**

posts comments on many other blogs. In our data, we found that the highest in-degree value in the comment network is 16 while the highest out-degree value is 26.

Note that the key bloggers with high degrees may not necessarily blog often about the iPod and vice versa. We compared the list of the top bloggers who frequently mentioned "iPod" as identified from the content analysis and the top bloggers with high degree centrality. There is no overlap between the two lists. The degrees of the bloggers who frequently blog about the iPod are rather low, ranging between 0 and 3. This implies that those bloggers may not communicate extensively with other bloggers.

The top 10 bloggers in each of the three networks as defined by the degree centrality as well as the other two measures (i.e., betweenness and closeness) are shown in Tables 3, 4, and 5. This allows us to compare the top bloggers in the different networks. In Table 3, we can see that there is only one blogger (#1015) appearing in the top 10 in both the subscription network and the comment network. In other words, bloggers who have the most subscription relationships are not the ones who have the most comments. In the combined network, eight of the top 10 bloggers are found in the list of top 10 bloggers in the comment network, while only two are present in the top bloggers in the subscription network.

Table 4 shows the top bloggers in terms of betweenness. There is no overlap between the top bloggers in the subscription network and the comment network. In the combined network, five of the top 10 bloggers are from the top 10 in the comment network, and only one is from the top 10 in the subscription network. This is similar to the pattern shown in the analysis of degree centrality, where the top bloggers in the combined network mostly come from the comment network.

Table 5 shows the list of bloggers with the highest closeness centrality. There is no overlap between the top bloggers in the subscription and the comment networks. In the list for the comment network, only two bloggers are from the subscription network and one is from the comment network. The small overlap indicates that the network structure changes significantly when the subscription network and the comment network are combined.

The top bloggers with high centrality may play an important role in spreading information, ideas, and opinions to the rest of the network. These bloggers may be opinion leaders whose views and opinions have influential effects on others. The process of information dissemination in social networks is often referred to as "word-of-mouth" (Engel et al. 1969) or "information cascade" (Watts 2002), a social phenomenon frequently studied in marketing research (Kozinets et al. 2010; Trusov et al. 2009).

To verify the role of central bloggers in disseminating information, we performed a simple study to simulate the word-of-mouth process in the blogosphere following a widely cited information cascading model in marketing (Watts and Dodds 2007). Specifically, we simulated the situation where the top bloggers are selected for one-to-one "seeding" in a marketing campaign (for instance, by giving them a free iPod), in the

**Table 3. Top 10 Bloggers with Highest Degree Centrality in the iPod Networks**

| | Subscription Network | | Comment Network | | Combined Network | |
|---|---|---|---|---|---|---|
| **Rank** | **Id** | **Blogger Name** | **Id** | **Blogger Name** | **Id** | **Blogger Name** |
| 1 | 2210 | patrickstar_4224 | 2016 | Mybearjana | 2016 | Mybearjana |
| 2 | 2284 | power_from_within | 1342 | Ipodvideo | 1015 | getcha__FREAKx3on |
| 3 | 948 | Free_ipod_lover | 1015 | Getcha__FREAKx3on | 2968 | trendy__xbarbie |
| 4 | 1041 | Gnell_Wallace_54 | 1954 | MIZZ_BROKEN_ONE | 1954 | MIZZ_BROKEN_ONE |
| 5 | 1181 | HsXeC__BaSsIsT | 2968 | Trendy__xbarbie | 1342 | Ipodvideo |
| 6 | 312 | Blondebebe_1 | 394 | brunettegirlie__x3 | 394 | brunettegirlie__x3 |
| 7 | 683 | Dev_Lee | 3207 | x__retr0thrills | 3207 | x__retr0thrills |
| 8 | 1015 | getcha__FREAKx3on | 412 | C0ACH__xBARBiE | 2534 | shadowed_star |
| 9 | 1420 | Jennnnyyyyy | 2688 | Sporty__xbarbie | 2489 | save_me_from_myself_6904 |
| 10 | 2489 | save_me_from_myself_6904 | 2213 | PeaceLoveSummerx3 | 412 | C0ACH__xBARBiE |

**Table 4. Top 10 Bloggers with Highest Betweenness Centrality in the iPod Networks**

| | Subscription Network | | Comment Network | | Combined Network | |
|---|---|---|---|---|---|---|
| **Rank** | **Id** | **Blogger Name** | **Id** | **Blogger Name** | **Id** | **Blogger Name** |
| 1 | 1015 | getcha__FREAKx3on | 1342 | Ipodvideo | 1954 | MIZZ_BROKEN_ONE |
| 2 | 1954 | MIZZ_BROKEN_ONE | 3368 | xX_YourPen1sIsEmo_Xx | 1342 | Ipodvideo |
| 3 | 1181 | HsXeC__BaSsIsT | 1608 | Kylee09 | 2016 | Mybearjana |
| 4 | 2 | a____MURDER__divinex | 2016 | Mybearjana | 2161 | OrangeBabeMLE |
| 5 | 2002 | Music_Madness0X | 1449 | Joelheflin | 1608 | kylee09 |
| 6 | 2681 | spontaneous_niknak | 1929 | Mind_Games_Infinity | 3368 | xX_YourPen1sIsEmo_Xx |
| 7 | 2003 | music_yox3 | 1733 | Livinginaparadox | 3470 | youXareXaXbutterflyXprincess |
| 8 | 840 | european_sheekX3 | 1333 | Ipodman | 3385 | Xxhilljoxx |
| 9 | 1113 | Heartbroken_and_confused | 2884 | the_temporary_solution | 1041 | Gnell_Wallace_54 |
| 10 | 2050 | needashuffle | 1936 | Miss_HelloKitty21 | 1449 | Joelheflin |

**Table 5. Top 10 Bloggers with Highest Closeness Centrality in the iPod Networks**

| | Subscription Network | | Comment Network | | Combined Network | |
|---|---|---|---|---|---|---|
| **Rank** | **Id** | **Blogger Name** | **Id** | **Blogger Name** | **Id** | **Blogger Name** |
| 1 | 1954 | MIZZ_BROKEN_ONE | 1342 | Ipodvideo | 1954 | MIZZ_BROKEN_ONE |
| 2 | 2 | a____MURDER__divinex | 3368 | xX_YourPen1sIsEmo_Xx | 2161 | OrangeBabeMLE |
| 3 | 2002 | Music_Madness0X | 1608 | kylee09 | 1342 | Ipodvideo |
| 4 | 1181 | HsXeC__BaSsIsT | 1929 | Mind_Games_Infinity | 1015 | getcha__FREAKx3on |
| 5 | 1015 | getcha__FREAKx3on | 1449 | Joelheflin | 3207 | x__retr0thrills |
| 6 | 2681 | spontaneous_niknak | 1733 | Livinginaparadox | 2213 | PeaceLoveSummerx3 |
| 7 | 2210 | patrickstar_4224 | 1936 | Miss_HelloKitty21 | 3385 | Xxhilljoxx |
| 8 | 2284 | power_from_within | 1494 | justin__tyner | 2000 | Music___0____URLS |
| 9 | 112 | AndIfIHadTheGuts_Caleb | 1324 | ipod_lover | 1105 | HC0Bl0NdiEx3 |
| 10 | 10 | A_T_B | 3451 | you_cant_bring_me_down1010 | 2968 | trendy__xbarbie |

**Table 6. Effectiveness in Information Dissemination in the iPod Networks**

| Effectiveness | | Bloggers selected from: | | |
|---|---|---|---|---|
| | | Subscription Network | Comment Network | Combined Network |
| Bloggers selected based on: | Degree centrality | 20.07 | 22.07 | 25.95 |
| | Betweenness centrality | 17.95 | 16.74 | 23.95 |
| | Closeness centrality | 15.29 | 14.24 | 20.34 |

hope that they will review the product in the blog and spread the message to other bloggers in the network (Kozinets et al. 2010). In our simulation, we assumed that after a blogger A posts a message, it will be read by blogger B with a probability $p(A, B)$ which is calculated based on the previous interactions of these two bloggers. In this study we calculated the probability as a linear combination of two Boolean values which indicate the presence of the subscription link and the comment link between the bloggers.[3] In each simulation, a set of top 10 bloggers were selected and activated (simulating the seeding of a marketing message) and we studied how many bloggers the message could reach, based on the probability calculated above. We selected 3 × 3 = 9 sets of bloggers, where in each set the top 10 bloggers were selected based on one of three the measures from one of the three networks. The effectiveness was then calculated as the number of bloggers receiving the message divided by the number of seed bloggers in each simulation. For each set of bloggers, the simulation was conducted 100 times, and the average effectiveness is shown in Table 6.

The results revealed that the top bloggers selected from the combined network (the rightmost column in the table) achieved higher effectiveness in disseminating messages than the top bloggers in the individual networks. When comparing the sets of bloggers selected based on the different centrality measures, we found that the set of bloggers selected based on degree centrality was the most effective. The top bloggers selected based on betweenness were less effective, and those based on closeness were the least effective.

## Implicit Communities

We performed community analysis on the three networks (i.e., subscription network, comment network, and the combined network). Our data revealed that both the subscription network and the comment network were very disconnected and had more nodes than links (as shown in Table 2), while the combined network was better connected. Clustering analysis was performed on the three networks and the related statistics are shown in Table 7.

The subscription network contained 289 separate clusters. The visualization of the top 15 clusters with a size greater than 10 is shown in Figure 4. In the figure, each circle represents a blogger and an arrow represents the relationship between two bloggers. The colors of the circles represent different attitudes of bloggers. Bloggers with a positive attitude are colored in red (dark grey), negative in blue (medium grey), and neutral in gray (light gray).[4] The key bloggers with a high degree are represented by larger nodes. The arrow between two nodes represents the direction of the relationship. For example, an arrow from A to B indicates that A subscribes to B's blog, meaning B's blogs are likely to be read by A.

The largest cluster, located in the middle of Figure 4, contains 71 members, which is 6.4 percent of the members in the network. This cluster consists mostly of bloggers with a positive attitude towards iPods. Bloggers #2489, #1015, and #2968 have many subscribers. Most of the other smaller clusters are also dominated by positive bloggers, except for two clusters that contain mostly neutral bloggers.

In the comment network, there are 292 clusters, with 9 clusters having at least 10 members. The visualization of these nine clusters is shown in Figure 5. The largest cluster, shown in the upper right, contains 124 members, constituting 11.2 percent of the total members in the network. In this cluster, we can see a group of blue circles consisting of bloggers who are negative toward the iPod, headed by Blogger #2016. This blogger has left comments on the blogs of many other negative bloggers. Further examination revealed that this blogger is the leader of a group called "I hate iPods." The rest of this cluster consists mainly of neutral bloggers, with a few positive bloggers.

[3] There are many possible alternative probability models. We experimented with several other models and obtained similar results. Therefore only the results obtained with the simplest model are presented here.

[4] Figures 4, 5, and 6 showing the colored nodes are available in the "Online Supplements" section of the *MIS Quarterly*'s website (http://www.misq.org).

**Table 7. The Statistics of the Clusters in the iPod Networks**

| Network | Subscription | Comment | Combined |
|---|---|---|---|
| Number of Clusters | 289 | 292 | 201 |
| Number of Nodes in the Largest Cluster | 71 | 127 | 717 |
| Number of Links in the Largest Cluster | 85 | 147 | 1,033 |

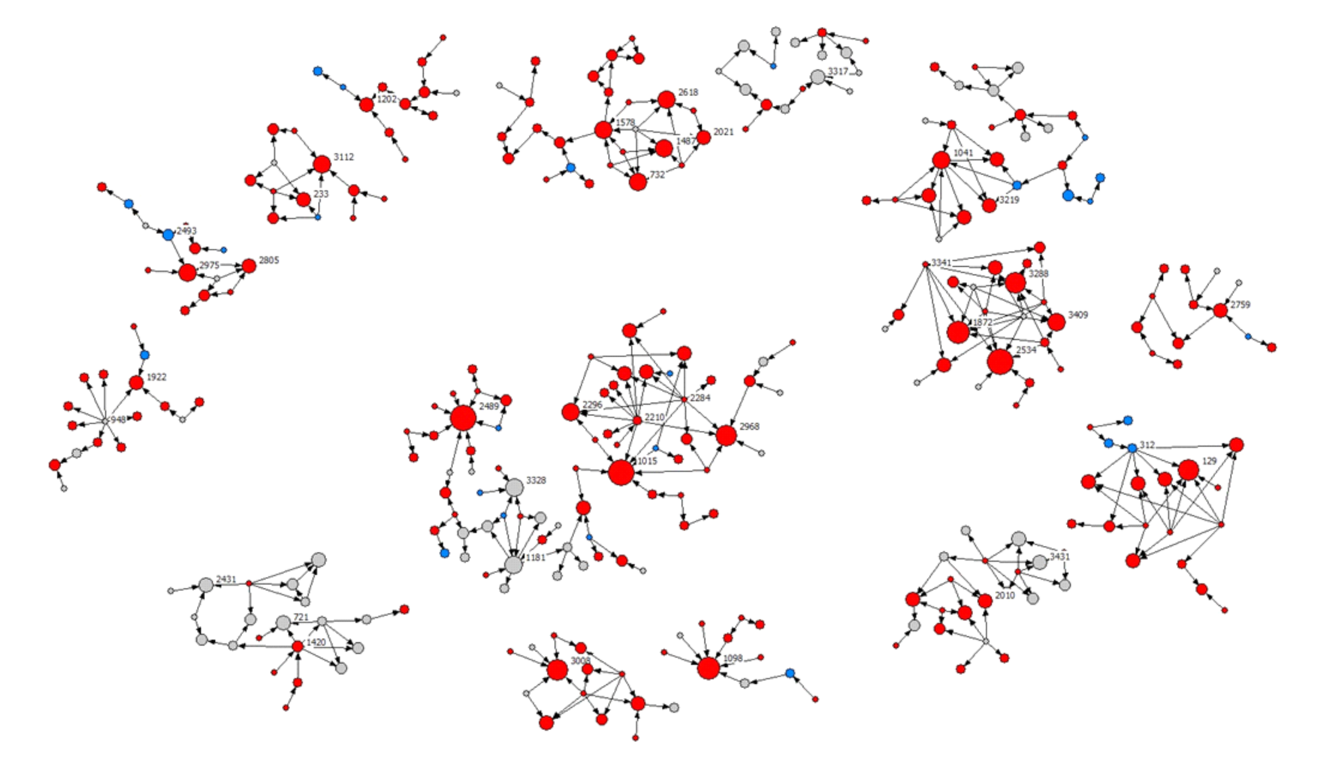

**Figure 4. Major Clusters in the Subscription Network in the iPod Data Set**

There are 79 members in the second largest cluster in the comment network, located on the left in Figure 5. This is a group of positive bloggers, including several who are active in posting comments, such as #2968, #1015, and #394.

We constructed the combined network by considering both types of relationships. This combined network contains 1,571 bloggers and 1,866 links. There are 201 clusters; the largest one has 717 members connected by 1,033 links. This cluster represents 45.6% of the total members in the combined network. This cluster is much larger than the largest clusters in the subscription network and the comment network, meaning that some bloggers connect those two networks.

We performed topological analysis on this largest cluster in the combined network. The cluster has a diameter of 30. The average shortest path length of the cluster, which is the mean of all-pairs shortest paths in a network, is 10.83. This means that, on average, a blogger in this cluster has to take more than 10 steps to reach another arbitrary blogger in the same cluster. The global efficiency, defined as the average of the inverses of shortest path lengths over all pairs of nodes in a network (Crucitti et al. 2003), is 0.12. All of these numbers suggest that the implicit communities (defined by subscription and comment) appear to be less efficient than the explicit communities (defined by the blogrings membership) which form a very dense network (Chau and Xu 2007). However, one should note that subscription and comment relationships are evidence of direct interaction between two bloggers, while blogring membership only represents indirect interaction. In other words, although the paths between bloggers in the subscription and comment networks are longer, information is more certain to pass through.

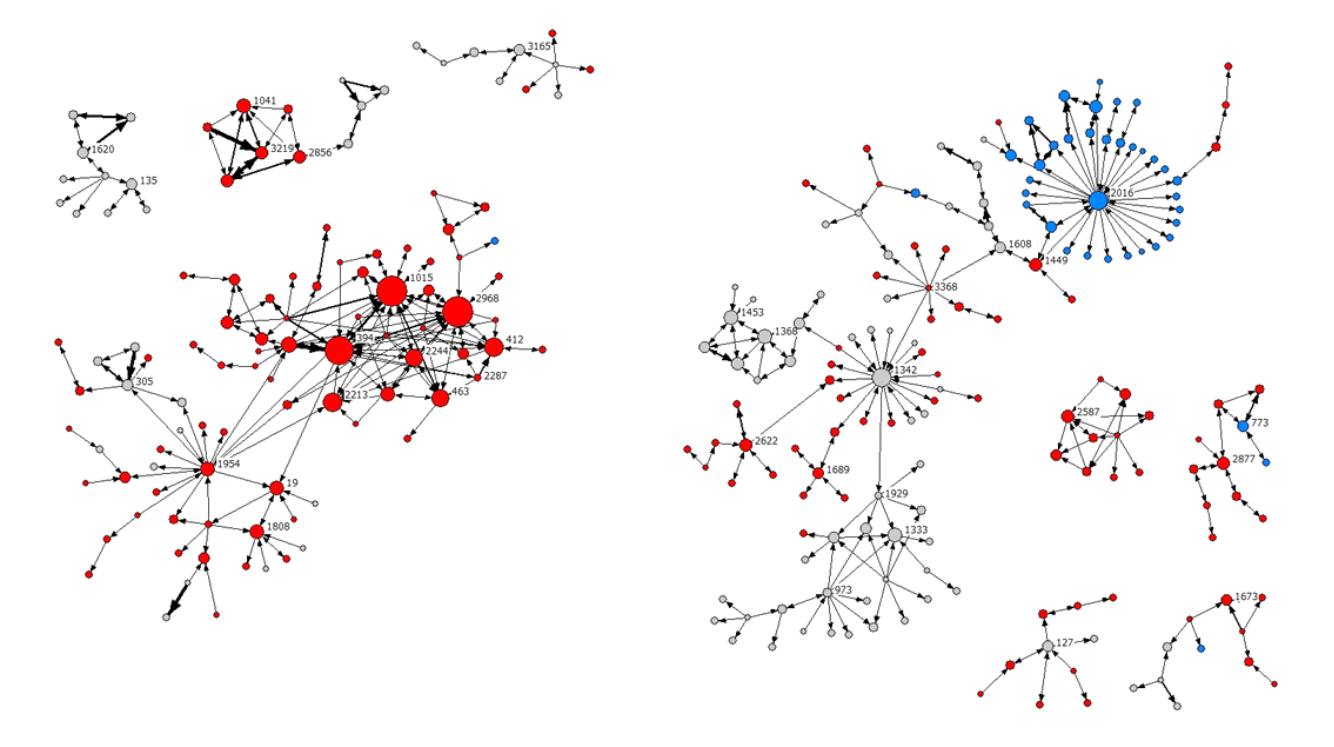

**Figure 5. Major Clusters in the Comment Network in the iPod Data Set**

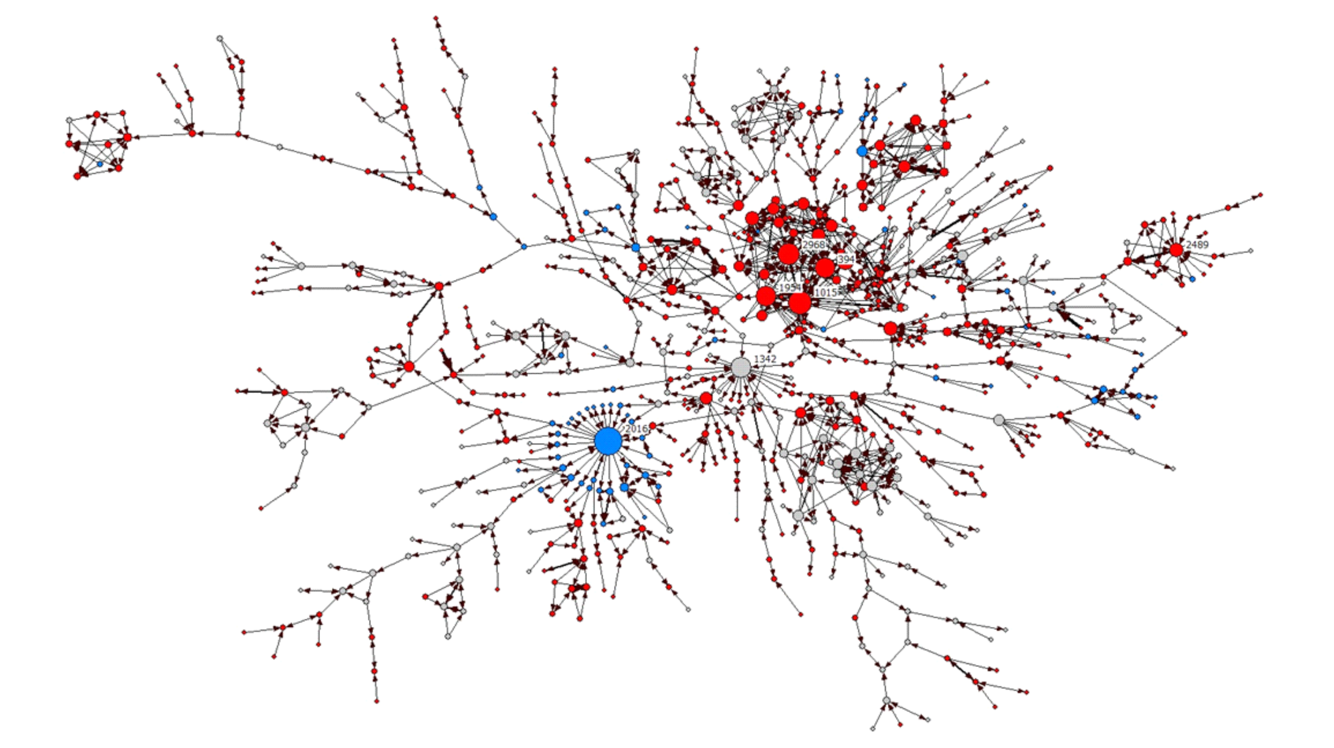

**Figure 6. Major Clusters in the Combined Network in the iPod Data Set**

A visualization of this largest cluster is shown in Figure 6. It is possible to identify a number of smaller communities in this cluster. These communities consist of bloggers who interact more frequently with members within the community than with others. From the figure we can see that bloggers with the same attitude toward iPods are clustered together in general, as shown by the color of the circles. A community of positive bloggers (red circles) can be found near the upper right of the center of the figure. This community includes a number of bloggers, such as #2968, #394, #1954, and #1015 represented by the big red circles, who are active in interacting with other bloggers.

Another major community can be found near the lower left center of the figure. This community consists of a number of negative bloggers (blue circles), with blogger #2016 being the most active. A third subcommunity, with neutral blogger #1342 being the most prominent node, can be found in the center of the figure. Other small communities also exist and can be found by identifying the major branches of the network.

All three major communities in the combined network are also found in the comment network, but only one (the positive community) can be found in the subscription network. This indicates that the comment network is more important, as it forms the base of the combined network and represents the major communities.

We can also see that while most interactions happen between bloggers with the same attitude toward iPods, bloggers with different attitudes also interact with each other. This is shown by the links between red circles and blue circles in the figure. In other words, the bloggers' differing opinions did not stop them from interacting.

## Case Study 2: Starbucks

The coffee chain store Starbucks was selected as the topic for our second case study. Starbucks provides people a place to have a drink and relax, and many people have become supporters of Starbucks and formed online communities about this company.

### Data Set

The data collection process was similar to that employed in the first case study. Using the onsite search engine, we searched for all the Xanga blogrings that contained the word "Starbucks" in their titles and descriptions. We found 873 groups and retrieved relevant data about them. Irrelevant or invalid groups were removed, resulting in 440 groups with a total of 15,360 bloggers and 694,645 blog entries in the final data set. We manually labeled each group as positive, negative, or neutral toward Starbucks based on the group description. Table 8 presents the information about the 20 largest groups in this study. Individual blogger's attitudes were determined based on the attitudes of the groups they joined. We found that the vast majority of bloggers (92.1%) have a positive attitude toward Starbucks, and only a small percentage of bloggers have a negative (1.5%) or a neutral (6.4%) attitude.

### Content Analysis

For each blog collected, we measured the word frequency (i.e., the number of times that the word Starbucks was mentioned). We found that the word was mentioned 20,814 times by 4,948 bloggers (32.2%). The blogger with the highest word frequency (333 times) mentioned Starbucks in almost every entry, mostly talking about having Starbucks coffee or spending time at Starbucks stores. The remaining 10,412 bloggers (67.8%) did not mention the word Starbucks in any of their blog entries and had a word frequency of 0. This percentage is even higher than the 55.0 percent found in the iPod data set. This number conforms to the finding in Case Study 1 that, although they have joined relevant interest groups, many bloggers do not blog about the company or product explicitly. These bloggers cannot be easily found by a traditional keyword search based on blog contents. We plot the percentage of bloggers (in logarithm scales) against the word frequency in Figure 7.

Similar to Case Study 1, we also extracted all sentences that contain the word Starbucks and randomly selected 300 sentences for further analysis. We found that most of these sentences (297 out of 300) are neutral and are about the blogger going to Starbucks or buying Starbucks coffee. Examples of these sentences are "Then we piled in his car and went to Starbucks," "I went to the deli after Starbucks," and "At Starbucks we met up with Paul." Only 3 of the 300 sentences explicitly indicated the blogger's attitude toward Starbucks, such as "I love Starbucks late at night," "Starbucks is awesome," and "I hate Starbucks, anyone who likes Starbucks can go shoot there [sic] brains out." Such opinion statements may be useful for business analysts to understand why people have particular attitudes toward the company.

### Interaction Networks

Three types of networks were analyzed to find the characteristics of the blog interactions (subscription, commenting,

**Table 8. The 20 Largest Groups for Starbucks**

| Group Title | Number of Members | Group Description | Category |
|---|---|---|---|
| *~*~Starbucks Frappachino Addicts~*~* | 1,937 | this blogring is for all of u that are as addicted to fraps as i am! all flavors welcome! | Positive |
| *Starbucks Addiction* | 1,601 | STARBUCKS...woOoO--need I say more?! | Positive |
| ! ~Starbucks Anonymous~ ! | 1,384 | If you love STARBUCKS, especially frappachinos, then this is for you! Support it, or drink Starbucks, this is your Blogring!! | Positive |
| <3 StarBucks | 1,134 | if you love starbucks then this blogring is right for you! | Positive |
| I LUV STARBUCKS | 984 | STARBUCKS STARBUCKS STARBUCKS STARBUCKS STARBUCKS STARBUCKS STARBUCKS STARBUCKS STARBUCKS! | Positive |
| STARBUCKS | 611 | Starbucks is awesome. Need I say more?! | Positive |
| i live at starbucks ♥ | 592 | Blogring leader: x_tanlines_x | Positive |
| !_*STARBUCKS*_! | 482 | THIS IS TO ALL U PEOPLE WHO LOVE STARBUCKS!!! IF U FEEL THE CRAVING FOR HYPERNESS THEN THIS IS TO YOU! LOVE FRAPS, BROWNIES, COFFEE, OR N E THING THAT STARBUCKS HAS TO OFFER? STARBUCKS...RULES! | Positive |
| ! ¤ Starbucks Frappachino Addicts ¤ ! | 409 | !Thiz Bloqq iz fo tha ONES who are addicted to tha STARBUCKS Frappachino!!! Any Flavor... Any size! JoiN iF you OnE oF Tha AddiCts! | Positive |
| Diet Coke and Starbucks | 399 | so f**king good. | Positive |
| starbucks//my//addiction | 278 | I LOVE STARBUCKS! I LOVE STARBUCKS! I LOVE STARBUCKS! I LOVE STARBUCKS! I LOVE STARBUCKS! I LOVE STARBUCKS! I LOVE STARBUCKS! I LOVE STARBUCKS! I LOVE STARBUCKS!I LOVE STARBUCKS!I LOVE STARBUCKS!I LOVE STARBUCKS!I LOVE STARBUCKS!I LOVE STARBUCKS!I LOVE STARBUCKS!I LOVE STARBUCKS!I LOVE STARBUCKS! | Positive |
| * StArBuCk*s_LoVeRs* | 251 | | Positive |
| i don't need drugs...I have Starbucks® | 240 | for those who a. want to smack people who accidenitly bump into you which makes you spill you starbucks b. would chose starbucks as your last meal on death row c. have your allowance paid in starbucks's gift cards d. have on the top of b-day and x-mas gift lists starbuck gift cards e. even have the thought of having the symbol of starbucks tatooed on you f. like starbucks....... immensly g. would kill for starbucks... well not really....well then again | Positive |
| StArBUcKs iS LoVE!<3 | 223 | CoFFee iS tHe c00LeSt ThiNg.. i ThiNk StaRbuCkS DesERveS A weBriNg foR dEv0tEd sTaRbuCkS L0vErs!!! <33 ~*~starbucks is love~*~ much l0ve~ .::aBBeRgAiLy::. p.s. ~ thanks for the nickname, whit<33 ALSO!! my friend and i kind of made up this thing, "starbucks = love" and now we use it all the time, like "i starbucks you!" so this webring is also dedicated to our cool little thing we made up.. heehee! :-D | Positive |
| STARBUCKS BARISTAS!! | 222 | Green apron wearing.....condiment bar checking.....hate making frappuchinos all the time....pastry stealing....eating left over sandwiches....only @ SBUX | Neutral |
| starbucks; the new york stop sign. | 217 | whether you're from the city or not. you love starbucks. join b***hes. | Positive |
| starbucks <3 | 167 | ..caramel macchiatos, coconut cremes, mocha frappucinos, vanilla espressos, hazelnut cappucinos, & toffee nut lattes sound dreamy? ..do you <3 walk`n into a starbucks store anytime & anywhere? if so, you must be a lover of starbux | Positive |
| ♥,Starbucks is love ♥, | 167 | Alicia's rps | Positive |
| *****Starbucks***** | 160 | come take a break, come chill at Starbucks , meet cool people, be social , have fun , be cool , be real , visit other people's xangas , and Give/Recieve LOTS of PROPS | Neutral |
| fall out boy, making out and starbucks <3 | 157 | the name explains it all | Positive |

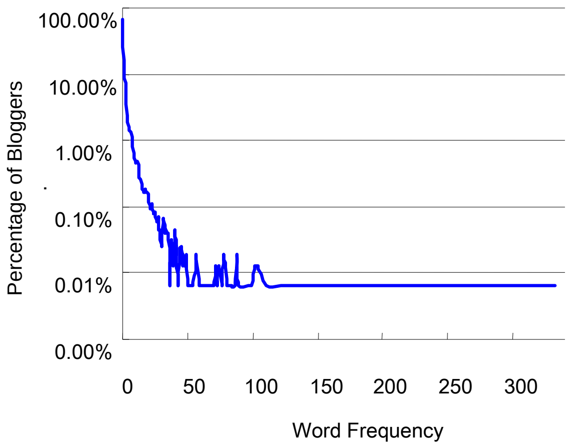

**Figure 7. Percentage of Bloggers (in Log Scale) Versus Word Frequency in the Starbucks Data Set**

**Table 9. The Statistics of the Interaction Networks in the Starbucks Data Set**

| Network | Subscription | Comment | Combined |
|---|---|---|---|
| Number of Nodes | 5,222 | 3,659 | 6,209 |
| Number of Links | 4,553 | 3,064 | 5,964 |
| Average Degree | 1.74 | 1.67 | 1.92 |
| Highest In-Degree | 20 | 15 | 26 |
| Highest Out-Degree | 69 | 24 | 69 |

and combined activities) between Starbucks consumers. Approximately 34 percent of the bloggers (5,222) in our data set appear in the subscription network (i.e., these bloggers have subscribed to or received subscriptions from others). The comment network is smaller than the subscription network with 3,659 bloggers. Combined, the subscription and comment networks include 40.4 percent of all the bloggers in the data. Table 9 presents the basic statistics of the three interaction networks. The remaining 9,151 nodes are isolated bloggers who do not have subscription and commenting interactions with other bloggers in this particular data set.

The three networks display scale-free characteristics with power–law degree distributions. The comment network, for example, is a rather strong scale-free network, whose cumulative degree distribution fits fairly well with the power–law distribution ($R^2$ = 0.97) in the log–log plot presented in Figure 8. This implies that while many bloggers have only a few comment links with others, a small number of bloggers are very interactive, making comments on others' blogs or attracting comments and stimulating discussion. The subscription network, on the other hand, has a lower goodness-of-fit score ($R^2$ = 0.86), which also causes the degree distribution of the combined network to fit the power–law distribution less well ($R^2$ = 0.88) than that of the comment network.

### Central Bloggers

The three centrality measures (degree, betweenness, and closeness) were used to identify the key bloggers in the Starbucks communities.

The highest in-degree in the subscription network is 20 while the highest in-degree in the comment network is 15. In the combined network, an in-link that appeared in either or both the subscription and the comment network was counted as one link. The highest in-degree in the combined network is 26 (see Table 9). The highest out-degrees in the three networks are 69, 24, and 69, respectively. As in Case Study 1, the in-degrees and out-degrees help identify the "authorities" who attract many subscriptions or comments and "hubs" who subscribe to or make comments on others' blogs, respectively. The hubs and authoritative bloggers may not necessarily be the same. For example, blogger #6742 has the highest out-degree (69) in the subscription network. However, the in-

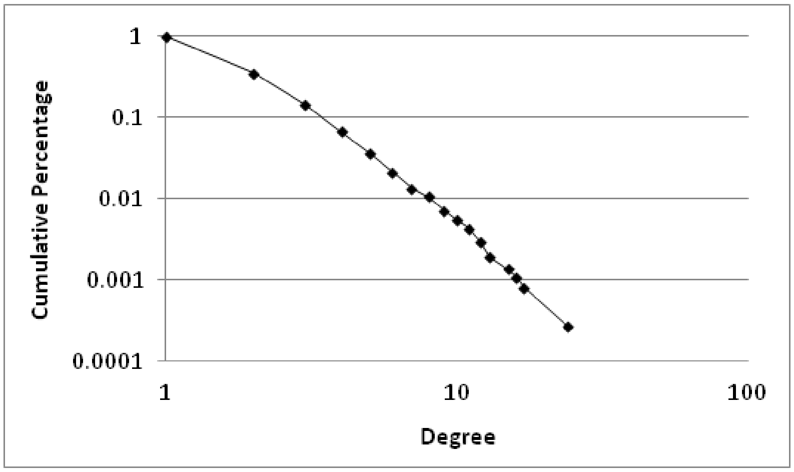

**Figure 8. Cumulative Degree Distributions of the Comment Network in the Starbucks Data Set**

degree is only one. This implies that this blogger likes to read blogs written by many other bloggers, but his/her own blog has not attracted many subscriptions. In contrast, blogger #6079 is the top authoritative node with the highest in-degree (20) but a low out-degree (4) in the subscription network. Thus, this blog has attracted many subscriptions, but the blogger is not interested in subscribing to many others' blogs.

Table 10 lists the top 10 bloggers with the highest degrees in the three networks. We found that five bloggers (#5873, #9629, #6079, #13699, and #335) appear in all three lists, indicating that they are quite actively involved in both subscription and comment interactions with other bloggers. Unlike Case Study 1 in which the top bloggers in the combined network mostly came from the comment network, the list of the top bloggers in the combined network in Table 10 overlaps to a large extent with both the subscription network list (sevem overlaps) and the comment network list (six overlaps).

We compared these lists with the top 10 list from our content analysis and found that there is no overlap between the bloggers with the top 10 word frequency and the bloggers who are top 10 in any one of the centrality measures. This concurs with the results obtained in Case Study 1. The bloggers who frequently blog about the company do not tend to have high degrees, and those who interact a lot with others do not necessarily blog a lot about the company.

Table 11 shows the top 10 bloggers with the highest betweenness scores in the three networks. Similar to the patterns found for degrees, the combined network list overlaps slightly more with the subscription network list (three overlaps) than with the comment network list (two overlaps).

The top 10 bloggers with the highest closeness scores are listed in Table 12. Again, the subscription network provides more top bloggers (six overlaps) in the combined network than does the comment network (three overlaps).

Similar to Case Study 1, we conducted simulation studies to find which types of central nodes are most effective in information dissemination in the network. Table 13 reports the average number of bloggers receiving the message divided by the number of seed bloggers in each simulation. By comparing the sets of bloggers who were the top in different centrality measures, we found that the set with the top degree centrality was the most effective, while the set with top betweenness was less effective and the set with top closeness was the least effective. The results also showed that the bloggers with the highest degree centrality selected from the combined network achieved higher effectiveness in disseminating messages than the other two networks. This is consistent with the results in Case Study 1.

### Implicit Communities

Similar to the community analysis results from Case Study 1, the networks in the Starbucks data set are rather disconnected with more nodes than links (see Table 9). Table 14 presents the statistics of the clusters in the three interaction networks of Starbucks consumers. All three networks contain many isolated clusters. The largest clusters in the subscription and

**Table 10. Top 10 Bloggers with Highest Degree Centrality in the Starbucks Networks**

| | Subscription Network | | Comment Network | | Combined Network | |
|---|---|---|---|---|---|---|
| **Rank** | **Id** | **Blogger Name** | **Id** | **Blogger Name** | **Id** | **Blogger Name** |
| 1 | 6742 | LaurenLC_Conrad | 9629 | PlatinumFashion | 6742 | LaurenLC_Conrad |
| 2 | 5813 | Jess_LagunaBeach | 5873 | Jjuicy | 5813 | Jess_LagunaBeach |
| 3 | 5873 | Jjuicy | 6079 | Juthena | 5873 | Jjuicy |
| 4 | 9629 | PlatinumFashion | 13087 | Viintagee | 9629 | PlatinumFashion |
| 5 | 6079 | Juthena | 335 | Alwaiz | 13699 | x_pinupp |
| 6 | 13699 | x_pinupp | 1143 | Bellacouturex | 6079 | Juthena |
| 7 | 335 | Alwaiz | 473 | Angelfiedpinay | 335 | alwaiz |
| 8 | 6712 | Late_Nite_Wonderz | 10192 | Retroflavoured_flats | 13087 | viintagee |
| 9 | 11321 | somethingele9ant | 9615 | Pixieme | 12413 | thenextnicolerichiex3 |
| 10 | 7167 | Littlemoules | 13699 | x_pinupp | 9266 | parkavenue___xx |

**Table 11. Top 10 Bloggers with Highest Betweenness Centrality in the Starbucks Networks**

| | Subscription Network | | Comment Network | | Combined Network | |
|---|---|---|---|---|---|---|
| **Rank** | **Id** | **Blogger Name** | **Id** | **Blogger Name** | **Id** | **Blogger Name** |
| 1 | 13145 | vogue__conspiracy | 13531 | x__alixandra | 6742 | LaurenLC_Conrad |
| 2 | 6742 | LaurenLC_Conrad | 6782 | le_cerise | 6522 | krissi6irl |
| 3 | 6522 | krissi6irl | 14539 | XoXglamourbarbieXOX | 9842 | prisci_6irl |
| 4 | 12733 | trashy_vogue | 13699 | X_pinupp | 7347 | love_music_xx |
| 5 | 9842 | prisci_6irl | 15271 | youre_my_wonderwall_x3 | 14539 | XoXglamourbarbieXOX |
| 6 | 5873 | Jjuicy | 4293 | GLAMOROUS_x_SHOPAHOLIC | 3035 | delicatex__romance |
| 7 | 13087 | Viintagee | 3772 | FallenFairee | 6184 | katie_the_kangaroo |
| 8 | 9312 | peacelovecouture | 14472 | xopink_is_the_new_orangexo | 7844 | mEgOlOvEsRoBbY |
| 9 | 5451 | invisibleaddiction | 14399 | XoCherriesXoX | 4885 | HollisterxBabe23 |
| 10 | 4545 | handbagcouturee | 6845 | Letmaxkno | 13531 | x__alixandra |

**Table 12. Top 10 Bloggers with Highest Closeness Centrality in the Starbucks Networks**

| | Subscription Network | | Comment Network | | Combined Network | |
|---|---|---|---|---|---|---|
| **Rank** | **Id** | **Blogger Name** | **Id** | **Blogger Name** | **Id** | **Blogger Name** |
| 1 | 13145 | vogue__conspiracy | 13531 | x__alixandra | 6742 | LaurenLC_Conrad |
| 2 | 12733 | trashy_vogue | 6782 | le_cerise | 9842 | prisci_6irl |
| 3 | 13087 | Viintagee | 14539 | XoXglamourbarbieXOX | 6522 | krissi6irl |
| 4 | 6522 | krissi6irl | 14399 | XoCherriesXoX | 13531 | x__alixandra |
| 5 | 9842 | prisci_6irl | 13699 | x_pinupp | 5873 | jjuicy |
| 6 | 9312 | peacelovecouture | 5873 | Jjuicy | 12733 | trashy_vogue |
| 7 | 5873 | jjuicy | 9376 | peruvian_royalty | 9189 | OxSuGa_KiSsEs1Xo |
| 8 | 6742 | LaurenLC_Conrad | 6845 | Letmaxkno | 13087 | viintagee |
| 9 | 10645 | SEQUiiNED | 9313 | Peacelovecoutureex | 5813 | Jess_LagunaBeach |
| 10 | 7400 | lovely_intuition | 3245 | Dooneyndbourke | 14399 | XoCherriesXoX |

**Table 13. Effectiveness in Information Dissemination in the Starbucks Networks**

| Effectiveness | | Bloggers Selected From: | | |
|---|---|---|---|---|
| | | Subscription Network | Comment Network | Combined Network |
| Bloggers selected based on: | Degree centrality | 58.05 | 56.80 | 63.90 |
| | Betweenness centrality | 50.71 | 47.95 | 55.85 |
| | Closeness centrality | 50.69 | 45.96 | 52.05 |

**Table 14. The Statistics of the Clusters in the Starbucks Networks**

| Network | Subscription | Comment | Combined |
|---|---|---|---|
| Number of Clusters | 1,493 | 995 | 1,486 |
| Number of Nodes in the Largest Cluster | 493 | 432 | 1,586 |
| Number of Links in the Largest Cluster | 655 | 565 | 2,236 |

comment networks each contains about 3 percent (subscription network: 493; comment network: 432) of the total members in the data set. The combined network, on the other hand, is better connected and its largest cluster connects slightly more than 10 percent (1,586) of all members. This implies that the subscription links and comment links are complementary to some extent and they join disconnected clusters together when combined.

Figures 9, 10, and 11 present the visualizations of the largest clusters in the subscription, comment, and combined networks, respectively. Again, the nodes represent the bloggers and are color coded based on the bloggers' attitudes toward Starbucks (positive: red; negative: blue; neutral: gray).[5] The size of a node is proportional to the degree of the node.

The largest cluster in the subscription network is dominated by bloggers with positive attitudes toward Starbucks. However, unlike the iPod clusters which also contain blue nodes, this cluster includes no bloggers with a negative attitude (blue) and only a few with a neutral attitude (gray). This is because the entire Starbucks data set is dominated by positive bloggers, as discussed in the description of the data set in the previous section.

In Figure 9, blogger #9629 has the highest in-degree (19) and thus is the most authoritative blogger in this cluster. Note that the blogger #6079 with an in-degree of 20 mentioned in the previous section regarding central bloggers is not contained in this particular cluster. The two big nodes (#6742 and #5813) in the lower right corner in this figure have very high out-degrees (69 and 57, respectively), indicating that the two bloggers subscribe to many other blogs. One can see in the visualization that the entire cluster can be roughly divided into three parts with blogger #13145 in the center. This blogger has a low in-degree (five) but is the highest one in the betweenness list (see the first column in Table 11). By receiving subscriptions from five bloggers from different parts of the cluster, blogger #13145 becomes the central "bridge" in this cluster.

The largest cluster in the comment network in Figure 10 does not contain any bloggers with a negative attitude toward Starbucks either. There are a few neutral bloggers grouped closely together in the lower right corner of the visualization. The bloggers with high degrees are labeled with their IDs (e.g., #13087, #5873, and #9615). The thickness of an arrow is proportional to the number of comments associated with the link. Compared with the major clusters in the comment network in Case Study 1, this cluster is relatively dense and well connected.

To avoid making the visualization of the largest cluster in the combined network too cluttered, we omitted arrow heads of the links in Figure 11. Interestingly, although this cluster is again dominated by positive bloggers, it includes bloggers with negative attitudes toward Starbucks (see the middle left of the visualization). It turns out that this small set of bloggers are all members of the blogring whose group description states that Starbucks is for old people and drinking coffee in Starbucks is out of fashion. The two largest nodes (#6742 and #5813) gain their centrality status because they subscribe to a large number of others' blogs.

[5] Figures 9, 10, and 11 showing the colored nodes are available in the "Online Supplements" section of the *MIS Quarterly*'s website (http://www.misq.org).

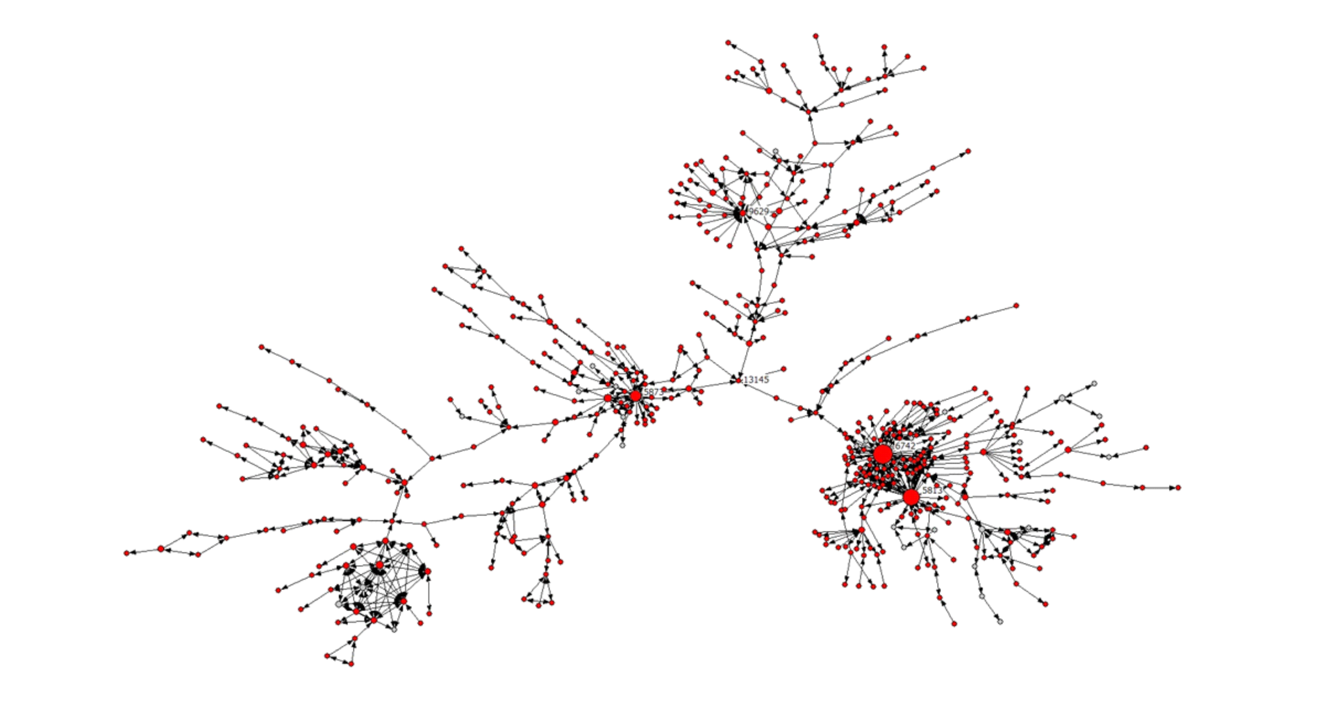

**Figure 9. The Largest Cluster in the Subscription Network in the Starbucks Data Set**

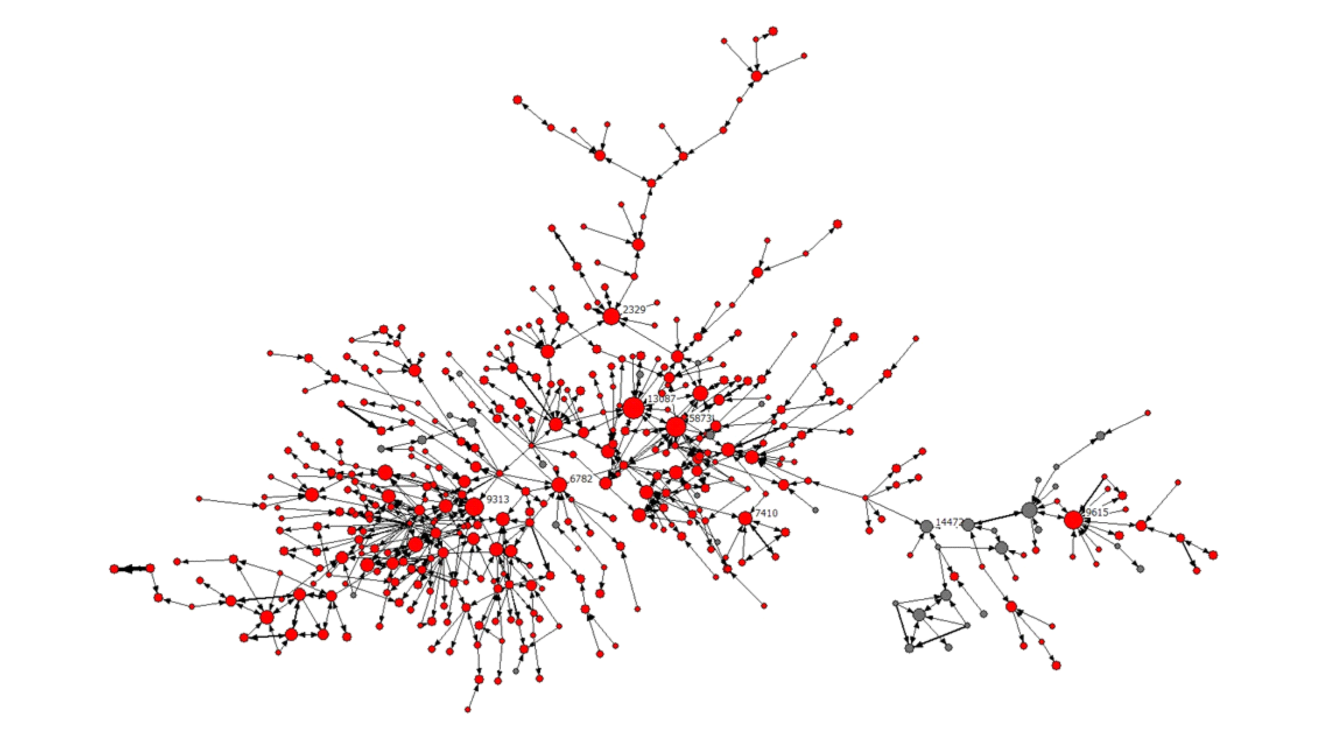

**Figure 10. The Largest Cluster in the Comment Network in the Starbucks Data Set**

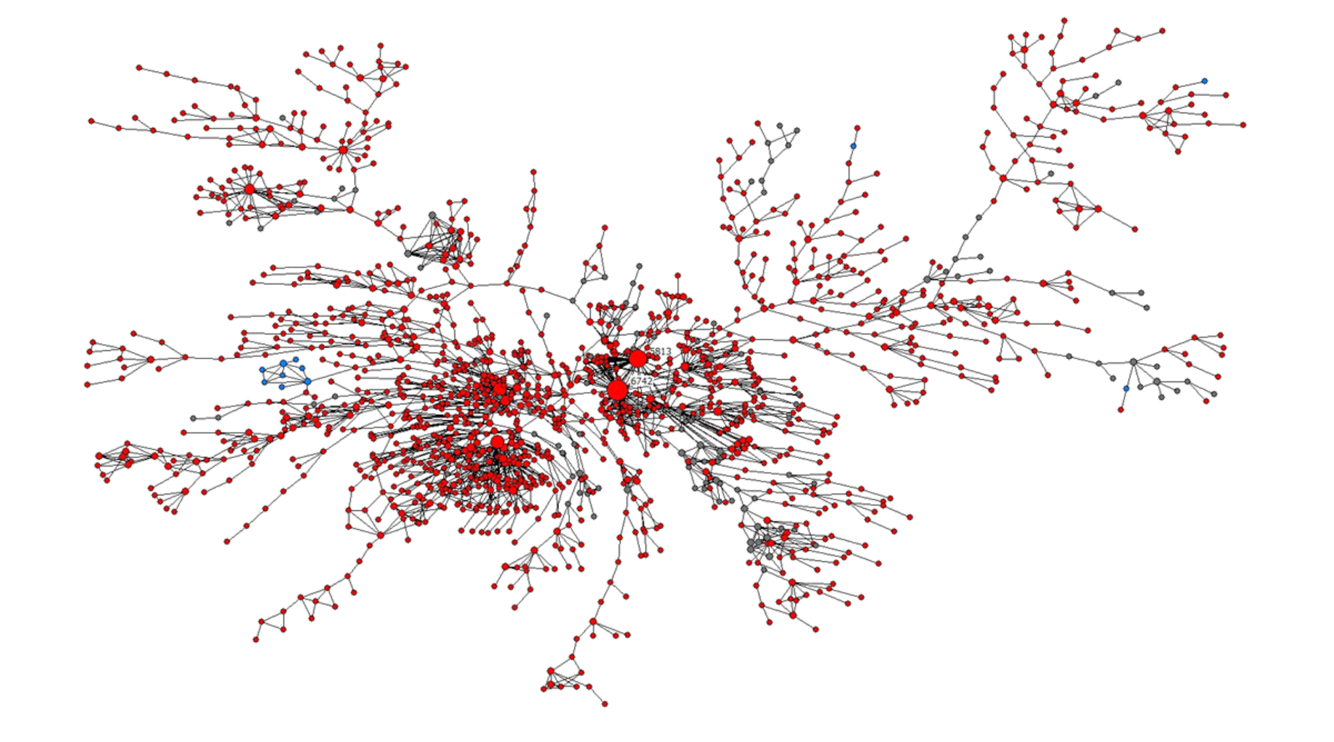

**Figure 11. The Largest Cluster in the Combined Network in the Starbucks Data Set**

## Discussion

In this section we discuss our findings with an aim to answer the questions raised regarding the contents, interaction networks, central bloggers, and implicit communities of the bloggers.

1. ***What are the characteristics of the blog contents of the consumers? How are they related to the product or company of interest?***

   Our results show that only 45.0 percent and 32.2 percent of bloggers identified in the explicit communities mentioned iPod and Starbucks, respectively, in their blogs. Our results verified that a simple keyword search method would miss a large portion of users who relate to the product or company. As suggested by our framework, analysis of blogger communities is necessary to reach these bloggers.

   Out of those contents containing the "iPod" keyword in Case Study 1, a great majority were about how the bloggers used the product. Similarly, most blogs mentioning Starbucks in Case Study 2 were about the bloggers buying Starbucks coffee or patronizing Starbucks. A very small portion (less than 2%) of the extracted contents explicitly revealed the blogger's feelings and opinions toward the product or company. This result is different than observations from other studies on opinion mining based on blog contents (Macdonald et al. 2010). One possible reason is that Xanga is a blog hosting site mostly used for personal contents and diaries and is not focused on product or company reviews (Ip and Wagner 2008). It is common to find review articles dedicated to particular products or companies on review-oriented blog sites. However, a blogger who likes iPod, for instance, may not write posts reviewing the pros and cons of iPod in depth in his/her diary-oriented blog on Xanga.

2. ***What are the characteristics of the interaction networks formed by bloggers?***

   In the two case studies, we analyzed two types of interactions between bloggers: subscription and comment. Both the subscription and comment networks found in the iPod data set are still rather disconnected. Only a small percentage of the total members are connected and have formed relatively small clusters. A similar pattern of low connectivity was also found in the

Starbucks data set. However, when the two types of interactions are combined, a much larger, dominating cluster would be formed. As discussed previously, bloggers are not constrained by any single type of interactions and may be involved in both subscription and commenting communication. Thus, it is very likely that the combined networks are closer to the real structure of blogger communities.

We also noted that subscription and comment relationships might not be equally important in shaping the combined network in different contexts. For example, in Case Study 1 we found that the combined network was dominated by comment relationships. However, Case Study 2 revealed that the subscription relationships affected the structure of the combined network more strongly than did the comment relationships.

Both types of interaction networks are scale-free and their degree distributions follow a power–law distribution in the two cases. The results verified that there exists a small number of bloggers who are involved in a larger number of interactions, and they are the central or key bloggers of the network. This is consistent with most other social network studies (e.g., Albert and Barabási 2002; Barabási and Albert 1999; Jeong et al. 2001) and blog network analysis studies (e.g., Shi et al. 2007; Yang and Counts 2010).

The blogger communities make it possible for new information and ideas to spread among bloggers through the channels facilitated by the subscription and comment links. In this sense, although consumers may not necessarily write often about a product or company in their personal blogs, their communities and social interactions within these communities can provide potential channels for disseminating product- or service-related information for marketing and customer relationship management purposes.

3. ***Who are the central bloggers in these interactions? Are these bloggers effective in disseminating information?***

Our findings in both cases suggested that the networks of bloggers have different centers of influence. This is consistent with the findings in previous studies (e.g., Burris et al. 2000; Chau and Xu 2007). Central bloggers, who interact with many others through subscription and comment activities, may be authoritative bloggers, opinion leaders, or hubs of communication. Moreover, some new, active bloggers may quickly attract considerable attention from other bloggers. These new bloggers frequently update their blogs and post new entries, thereby receiving many comments and subscriptions. They may become rising stars in the blogosphere and can potentially influence others.

In our simulation studies, we found that bloggers with a high degree in the combined network were most effective in disseminating messages to other bloggers. The bloggers identified by betweenness and closeness measures are slightly less effective. This result is consistent with the findings in the information cascading model study (Watts and Dodds 2007), which reveals that "hyper-influential" individuals, whose influential ability is a function of their degree and personal characteristics (e.g., expertise), play a much stronger role in enabling large-scale information cascades than do average individuals. These findings have important implications for companies to strategically select seed bloggers in their Web 2.0 marketing programs.

4. ***Do the implicit communities formed by different types of interactions demonstrate different properties? Which types of interactions are more important in shaping the communities?***

In both the iPod and Starbucks cases, we found that the networks of bloggers are decentralized. In each network we found a large number of isolated clusters with several larger ones. We did not find centralized structures such as star and hierarchical structures. This result was not surprising because these implicit communities have been formed spontaneously. Nevertheless, the absence of formal organizational structure does not eliminate the possibility that these groups may help prepare future members for more formal organizations, such as product fan clubs.

However, densely knit clusters of bloggers do exist in all networks in the study. In particular, by combining subscription and comment relationships, bloggers who were not previously connected were brought together and formed newer or bigger communities. However, no giant component (clusters containing more than 60% of the members; Bollobás 1985) was found in these networks. Our results show that blogger communities for a product or company are less centralized and have fewer interactions than those communities in other domains, where bloggers share common interests, beliefs, and social values (Chau and Xu 2007).

Indeed, the role of attitude and interest in bringing bloggers together manifested itself, to some extent, in our case studies. We found that within the largest cluster in the combined network, there may exist several subcommunities of bloggers who share the same attitude toward the product or company of interest. In these subcommunities, bloggers frequently interact with each other by reading one another's blogs and making comments. These communities may play an important role in reinforcing the interests and beliefs of their members and help create a "collective identity" (Gerstenfeld et al. 2003). We found that members from one subcommunity also interact with members from other subcommunities. In both case studies, we found that bloggers interacted with other bloggers with different attitudes.

Note that the implicit communities formed based on bloggers' interactions (subscription and comment) do not necessarily overlap with the blogrings, which are the explicit communities. Bloggers have their own cliques in which they have their own interests, opinions, and even leaders. These implicit communities are more meaningful because bloggers actually have exchanged and spread their opinions and messages through these interactions.

## Limitations

Our research has several limitations. First, we collected blogs in only one blog hosting site, Xanga, which contains primarily personal blogs. There are many other major blog hosts with different focuses (e.g., consumer reviews). While we found in the case studies that specific opinions toward the topics of interest were not abundant in the blog contents, this may not be applicable to other review-oriented blogs. Therefore, caution should be taken when applying the results of the current study to other blog hosting sites.

Second, as different sites have different technical features for organizing blogs, the procedures for blog data collection from other sites may be different depending on the specifics of those sites. Moreover, these technical features may to some extent affect the ways in which bloggers interact with one another and further affect the structure of the interaction networks. For example, subscribing to blogs at some blog hosting sites is not as easy as it is at Xanga. This will significantly reduce the chance of subscription links between bloggers in those sites.

In addition, our data sets are relatively small compared with the large collections of data used in some other studies (Macdonald et al. 2010). When a very large volume of blogs are available, it is possible that more interesting opinions relevant to the topics of interest can be found, as has been done in many other blog opinion mining studies. While some text mining techniques can be applied to these large data sets to obtain more relevant blogs, it is time-consuming to collect the data and the data sets often are outdated, as they are constrained to the given snapshot provided by the creator (for example, it took three months in 2005 and 2006 to collect the TREC Blogs06 data set and the resulting collection is 148GB in size; the TREC Blogs08 data set collection took even longer). We suggest that these data sets are more suitable for academic research than timely business intelligence analysis for practical purposes. Therefore, as our proposed framework covers the whole blog data collection and analysis process, we think it is more suitable and less costly for business intelligence purposes.

Last but not least, we used only a small set of existing techniques to mine the contents and network structures in these two case studies. Many other techniques for content analysis, opinion mining, influential blogger identification, community analysis, and information dissemination analysis can fit into our framework and are all potentially useful in generating business intelligence.

## Conclusion and Future Research

In this paper, we present our design of a framework and a system that automatically collects blogs and analyzes the blog contents and underlying social networks of bloggers. Using the system, we study bloggers' interaction patterns and communities in two case studies centered on a consumer product and a company. Our case studies demonstrate the usefulness of the framework and reveal interesting patterns, which answer important questions in the domain of business intelligence in blogs.

Our research has several implications. First, our framework and system can be used for extracting information such as demographic data and various relationships from blogs and for performing business intelligence analysis on the data collected. The content analysis and network analysis methods applied on the blog collections can be extended to other studies. As the framework is generic, it can be easily applied in other applications. The techniques are not constrained by those used in these two cases. New techniques can be readily plugged into the framework for identifying novel patterns and useful knowledge. In practice, business and marketing managers can apply the framework for business intelligence analysis on a wide range of organizations, products, and topics.

Second, we found that different types of interactions resulted in different clusters of bloggers, and by combining subscription and comment relationships we were able to connect more bloggers together. Comment relationships are more important in forming the largest cluster in the combined network. As discussed earlier, a comment relation provides strong evidence that a reader has read a blogger's entry, formed an opinion, and communicated to the blogger. The current study has provided some insights into the nature of these links and offered a methodology for possible further research.

Third, we believe our study is timely and important for research and practice in the area of business intelligence. While many previous studies recognized the potential of blog mining for business intelligence, very few have provided a viable methodology, coupled with two in-depth case studies, describing how it should be conducted. As business intelligence for Web 2.0 content is becoming increasingly important, our study has provided the foundation for future work on this topic.

We have several directions for our future research. One possible future study will be to evaluate our design and compare the analysis results across different blog hosting sites, as in the current study we only collected data on one single blog site. Similarly it will be useful to perform analysis on other products and companies, which may result in different interaction patterns. This will improve the generalizability of our results. Another direction is to apply our framework and methodology to other domains. For example, the communities of bloggers who have particular political views can be studied using our framework.

## Acknowledgments

This research has been supported in part by a Seed Funding for Basic Research grant from the University of Hong Kong. We would like to thank Porsche Lam, Keith Tseung, Frances Law, Yuanyuan Hong, and Han Zhou of the University of Hong Kong for their help in data collection and analysis. We are also thankful to Ahmed Abbasi of Virginia Tech for his comments on the paper. Last but not least, we are grateful to the guest senior editors, associate editor, and the three anonymous reviewers for their invaluable comments during the review process.

## References

Abbasi, A., and Chen, H. 2008. "CyberGate: A Design Framework and System for Text Analysis of Computer-Mediated Communication," *MIS Quarterly* (32:4), pp. 811-837.

Abbasi, A., Chen, H., Thoms, S., and Fu, T. 2008. "Affect Analysis of Web Forums and Blogs Using Correlation Ensembles," *IEEE Transactions on Knowledge and Data Engineering* (20:9), pp. 1168-1180.

Adar, E., and Adamic, L. A. 2005. "Tracking Information Epidemics in Blogspace," in *Proceedings of the 2005 IEEE/WIC/ ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, Compiègne, France, September 19-22.

Agarwal, N., Galan, M., Liu, H., and Subramanya, S. 2010. "WisColl: Collective Wisdom Based Blog Clustering," *Information Sciences* (180:1), pp. 39-61.

Agarwal, N., Liu, H., Tang, L., and Yu., P. S. 2008. "Identifying the Influential Bloggers in a Community," in *Proceedings of the Conference on Web Search and Data Mining,* Palo Alto, CA, February 11-12, pp. 207-218.

Albert, R., and Barabási, A.-L. 2002. "Statistical Mechanics of Complex Networks," *Reviews of Modern Physics* (74:1), pp. 47-97.

Albert, T. C., Goes, P. B., and Gupta, A. 2004. "GIST: A Model for Design and Management of Content and Interactivity of Customer-Centric Web Sites," *MIS Quarterly* (28:2), pp. 161-182.

Ali-Hasan, N. F., and Adamic, L. A. 2007. "Expressing Social Relationships on the Blog through Links and Comments," in *Proceedings of International Conference on Weblogs and Social Media,* Boulder, CO, March 26-28.

Baker, S., and Green, H. 2005. "Blogs Will Change Your Business," *BusinessWeek,* May 2, pp. 44-53.

Barabási, A.-L., and Albert, R. 1999. "Emergence of Scaling in Random Networks," *Science* (286:5439), pp. 509-512.

Bautin, M., Vijayarenu, L., and Skiena, S. 2008. " International Sentiment Analysis for News and Blogs," in *Proceedings of the Second International Conference on Weblogs and Social Media* Seattle, WA, March 30-April 2.

Blood, R. 2004. "How Blogging Software Reshapes the Online Community," *Communications of the ACM* (47:12), pp. 53-55.

Bollobás, B. 1985. *Random Graphs,* London: Academic Press.

Bulters, J., and de Rijke, M. 2007. "Discovering Weblog Communities," in *Proceedings of International Conference on Weblogs and Social Media*, Boulder, CO, March 26-28.

Burris, V., Smith, E., and Strahm, A. 2000. "White Supremacist Networks on the Internet," *Sociological Focus* (33:2), pp. 215-235.

Chau, M., Shiu, B., Chan, I., and Chen, H. 2007. "Redips: Backlink Search and Analysis on the Web for Business Intelligence Analysis," *Journal of the American Society for Information Technology* (58:3), pp. 351-365.

Chau, M., Qin, J., Zhou, Y., Tseng, C., and Chen, H. 2005. "SpidersRUs: Automated Development of Vertical Search Engines in Different Domains and Languages," in *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries,* Denver, CO, June 7-11, pp. 110-111.

Chau, M., and Xu, J. 2007. "Mining Communities and Their Relationships in Blogs: A Study of Hate Groups," *International Journal of Human–Computer Studies* (65), pp. 57-70.

Chau, M., Xu, J., Cao, J., Lam, P., and Shiu, B. 2009. "Blog Mining: A Framework and Example Applications," *IEEE IT Professional* (11:1), pp. 36-41.

Chen, C., Paul, R. J., and O'Keefe, B. 2001. "Fitting the Jigsaw of Citation: Information Visualization in Domain Analysis," *Journal of American Society of Information Science and Technology* (52:4), pp. 315-330.

Chevalier, J. A., and Mayzlin, D. 2006. "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research* (43), pp. 345-354.

Chung, W., Chen, H., and Nunamaker, J. F. 2005. "A Visual Knowledge Map Framework for the Discovery of Business Intelligence on the Web," *Journal of Management Information Systems* (21:4), pp. 57-84.

Cooley, R., Mobasher, B., and Srivastava, J. 1997. "Web Mining: Information and Pattern Discovery on the World Wide Web," in *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence*, Newport Beach, CA, November 3-8, pp. 558-567.

Crucitti, P., Latora, V., Marchiori, M., and Rapisarda, A. 2003. "Efficiency of Scale-Free Networks: Error and Attack Tolerance," *Physica A* (320), pp. 622-642.

Engel, J. F., Kegerreis, R. J., and Blackwell, R. D. 1969. "Word-of-Mouth Communication by the Innovator," *Journal of Marketing Research* (33), pp. 15-19.

Erickson, T. 1997. "Social Interaction on the Net: Virtual Community as Participatory Genre," in *Proceedings of the 30th Hawaii International Conference on System Sciences*, Los Alamitos, CA: IEEE Computer Society Press.

Feng, S., Wang, D., Yu, G., Yang, C., and Yang, N. 2009. "Sentiment Clustering: A Novel Method to Explore in the Blogosphere," in *Proceedings of the Joint International Conferences, APWeb, WAIM: Advances in Data and Web Management*, Q. Li, L. Feng, S. Wang, X. Zhou, and Q. Zhu (eds.), New York: Springer, pp. 332-344.

Freeman, L. C. 1979. "Centrality in Social Networks: Conceptual Clarification," *Social Networks* (1), pp. 215-240.

Gerstenfeld, P. B., Grant, D. R., and Chiang, C. P. 2003. "Hate Online: A Content Analysis of Extremist Internet Sites," *Analyses of Social Issues and Public Policy* (3), pp. 29-44.

Gill, A. J., Nowson, S., and Oberlander, J. 2009. "What Are They Blogging About? Personality, Topic and Motivation in Blogs," in *Proceedings of the Third International Conference on Weblogs and Social Media*, San Jose, CA, May 17-20.

Glance, N., Hurst, M., Nigam, K., Siegler, M., Stockton, R., and Tomokiyo, T. 2005. "Analyzing Online Discussion for Marketing Intelligence," in *Proceedings of the 14th International World Wide Web Conference*, Chiba, Japan, May 10-14.

Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A. 2004. "Information Diffusion through Blogspace," in *Proceedings of the 13th International World Wide Web Conference*, New York, May 17-20, pp. 491-501.

He, B., Macdonald, C., and Ounis, I. 2008, "Ranking Opinionated Blog Posts Using OpinionFinder," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Singapore, July 20-24, pp. 727-728.

Hevner, A. R., March, S. T., Park, J., and Ram, S. 2004. "Design Science in Information Systems Research," *MIS Quarterly* (28:1), pp. 75-105.

Ip, R. K. F., and Wagner, C. 2008. "Weblogging: A Study of Social Computing and its Impact on Organizations," *Decision Support Systems* (45), pp. 242-250.

Jeong, H., Mason, S. P., Barabási, A.-L., and Oltvai, Z. N. 2001. "Lethality and Centrality in Protein Networks," *Nature* (411:6833), pp. 41-42.

Kleinberg, J. 1999. "Authoritative Sources in a Hyperlinked Environment," *Journal of the ACM* (46:5), pp. 604-632.

Kozinets, R. V., de Valck, K., Wojnicki, A. C., and Wilner, S. J. S. 2010. "Networked Narratives: Understanding Word-of-Mouth Marketing in Online Communities," *Journal of Marketing Research* (74:2), pp. 71-89.

Kumar, R., Novak, J., Raghavan, P., and Tomkins, A. 2005. "On the Bursty Evolution of Blogspace," *World Wide Web: Internet and Web Information Systems* (8), pp. 159-178.

Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. 1999. "Trawling the Web for Emerging Cyber-Communities," *Computer Networks* (31:11-16), pp. 1481-1493.

Kumar, S., Zafarani, R., Abbasi, M. A., Barbier, G., and Liu, H. 2010. "Convergence of Influential Bloggers for Topic Discovery in the Blogosphere," in *Proceedings of International Conference on Social Computing, Behavioral Modeling and Prediction*, Bethesda, MD, March 30-31, pp. 406-412.

Lakshmanan, G. T., and Oberhofer, M. A. 2010. "Knowledge Discovery in the Blogosphere: Approaches and Challenges," *IEEE Internet Computing* (14:2), pp. 24-34.

Lee, Y., Na, S. H., Kim, J., Nam, S. H., Jung, H. Y., and Lee, J. H. 2008. "KLE at TREC 2008 Blog Track: Blog Post and Feed Retrieval," in *Proceedings of the 17th Text REtrieval Conference*, Gaithersberg, MD, November 16-19, pp. 500-277.

Lewis, S. 2008. "Using Online Communities to Drive Commercial Product Development," in *Proceedings of the 26th Conference on Human Factors in Computing Systems*, Florence, Italy.

Liang, H., Tsai, F. S., and Kwee, A. T. 2009. " Detecting Novel Business Blogs," in *Proceedings of the 7th International Conference on Information, Communications and Signal Processing*, Macau, China, December 8-10.

Lin, C. L., and Kao, H. Y. 2010. "Blog Popularity Mining Using Social Interconnection Analysis," *IEEE Internet Computing* (14:4), pp. 41-49.

Lin, Y.-R., Sundaram, H., Chi, Y., Tatemura, J., and Tseng, B. 2006. "Discovery of Blog Communities Based on Mutual Awareness," in *Proceedings of the 3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, Edinburgh, Scotland, May 23.

Liu, B., Hu, M., and Cheng, J. 2005. "Opinion Observer: Analyzing and Comparing Opinions on the Web," in *Proceedings of the 14th International World Wide Web Conference*, Chiba, Japan, May 10-14.

Liu, X., Wang, Y., Li, Y., and Shi, B. 2011. "Identifying Topic Experts and Topic Communities in the Blogspace," in *Proceedings of the16th International Conference on Database Systems for Advanced Applications*, Hong Kong, April 22-25, pp. 68-77.

Macdonald, C., Santos, R. L. T., Ounis, I., and Soboroff, I. 2010. "Blog Track Research at TREC," *SIGIR Forum* (44:1), pp. 58-75.

Mathioudakis, M., and Koudas, N. 2009. "Efficient Identification of Starters and Followers in Social Media," in *Proceedings of the 12th International Conference on Extending Database Technology*, Saint Petersburg, Russian Federation, March 23-26, pp. 708-719.

McGonagle, J. J., and Vella, C. M. 1999. *The Internet Age of Competitive Intelligence*, Westport, CT: Quorum.

Nahon, K., Hemsley, J., Walker, S., and Hussain, M. 2011. "Blogs: Spinning a Web of Virality," *Proceedings of the iConference*, Seattle, WA, February 8-11.

Nardi, B. A., Schiano, D. J., Gumbrecht, M., and Swartz, L. 2004. "Why We Blog," *Communications of the ACM* (47:12), pp. 41-46.

O'Leary, D. E. 2011. "Blog Mining-Review and Extensions: 'From Each According to His Opinion,'" *Decision Support Systems* (51:4), pp. 821-830.

Papagelis, M., Bansal, N., and Koudas, N. 2009. "Information Cascades in the Blogosphere: A Look Behind the Curtain," in *Proceedings of the Third International Conference on Weblogs and Social Media*, San Jose, CA, May 17-20.

Pikas, C. K. 2005. "Blog Searching for Competitive Intelligence, Brand Image, and Reputation Management," *Online* (29:4), pp. 16-21.

Roberts, T. L. 1998. "Are Newsgroups Virtual Communities?," in *CHI'98 Conference Proceedings: Human Factors in Computing Systems*, C-M. Karat and A. Lund (eds.), Los Angeles, CA,

Sack, W. 2000. "Conversation Map: An Interface for Very Large-Scale Conversations," *Journal of Management Information Systems* (17:3), pp. 73-92.

Shi, X., Tseng, B., and Adamic, L. A. 2007. "Looking at the Blogosphere Topology through Different Lenses," *Proceedings of the International Conference on Weblogs and Social Media*, Boulder, CO, March 26-28.

Tang, L., Liu, H., and Zhang, J. 2012. "Identifying Evolving Groups in Dynamic Multi-Mode Networks," *IEEE Transactions on Knowledge and Data Engineering* (24:1), pp. 72-85.

Tang, L., Wang, X., and Liu, H. 2009. "Uncovering Groups Via Heterogeneous Interaction Analysis," in *Proceedings of the IEEE International Conference on Data Mining*, Miami, FL, December 6-9.

Trusov, M.,Bucklin, R. E., and Pauwels, K. 2009. "Effects of Word-of-Mouth Versus Traditional Marketing: Findings from an Internet Social Networking Site," *Journal of Marketing Research* (73), pp. 90-102.

Tsai, F. S. 2011. "A Tag-Topic Model for Blog Mining," *Expert Systems with Applications* (38), pp. 5330-5335.

Tsai, F. S., Chen, Y., and Chan, K. L. 2007. "Probabilistic Techniques for Corporate Blog Mining," in *Proceedings of the 11th Pacific Asia Conference on Knowledge Discovery and Data Mining*, Nanjing, China, May 22-25, pp. 35-44.

Viegas, F. B., and Smith, M. 2004. "Newsgroup Crowds and AuthorLines: Visualizing the Activity of Individuals in Conversational Cyberspaces," in *Proceedings of the 45th Hawaii International Conference on System Sciences*, Los Alamitos, CA: IEEE Computer Society Press.

Wasserman, S., and Faust, K. 1994. *Social Network Analysis: Methods and Applications*, Cambridge, England: Cambridge University Press.

Watts, D. J. 2002. "A Simple Model of Information Cascades on Random Networks," in *Proceedings of the National Academy of Science* (99), pp. 5766-5771.

Watts, D. J., and Dodds, P. S. 2007. "Influentials, Networks, and Public Opinion Information," *Journal of Consumer Research* (34), pp. 441-458.

Watts, D. J., and Strogatz, S. H. 1998. "Collective Cynamics of 'Small-World' Networks," *Nature* (393), pp. 440-442.

Xu, J. J., and Chau, H. 2006. "The Social Identity of IS: Analyzing the Collaboration Network of the ICIS Conferences (1980–2005)," in *Proceedings of the 27th International Conference on Information Systems*, Milwaukee, WI, November 10-13.

Xu, J. J., and Chen, H. 2005. "CrimeNet Explorer: A Framework for Criminal Network Knowledge Discovery," *ACM Transactions on Information Systems* (23:2), pp. 201-226.

Yang, J., and Counts, S. 2010. "Comparing Information Diffusion Structure in Weblogs and Microblogs," in *Proceedings of the 4th International Conference on Weblogs and Social Media*, W. W. Cohen nd S. Gosling (eds.), Washington, DC, May 23-26.

Zhang, X., Zhou, Z., and Wu, M. 2009. "Positive, Negative, or Mixed? Mining Blogs for Opinions," in *Proceedings of the 14th Australasian Document Computing Symposium*, J. Kay, P. Thomas, and A. Trotman (eds.), Sydney, Australia, December 4, pp. 141-145.

Zhu, J. Y., and Tan, B. C. Y. 2007. "Effectiveness of Blog Advertising: Impact of Communicator Expertise, Advertising Intent and Product Involvement," in *Proceedings of the 28th Annual International Conference on Information Systems*, Montreal, Canada, December 9-12.

## About the Authors

**Michael Chau** is an associate professor in the School of Business at the University of Hong Kong. He received a Ph.D. in management information systems from the University of Arizona and a B.Sc. in computer science and information systems from the University of Hong Kong. His current research interests include information retrieval, web mining, data mining, social media, electronic commerce, and security informatics. His research has appeared in journals such as *ACM Transactions on MIS*, *ACM Transactions on IS*, *Communications of the ACM*, *Decision Support Systems*, *IEEE Computer*, *IEEE Transactions on Knowledge and Data Engineering*, *Journal of the Association for Information Systems*, and *Journal of American Society for Information Science and Technology*. He is the author of more than 100 articles and ranked as the 14th most productive researcher in the field of information science in the period 1998-2007 in a research productivity study.

**Jennifer J. Xu** is an associate professor of Computer Information Systems at Bentley University. She received her Ph.D. in Management Information Systems from the University of Arizona. Her research interests include knowledge discovery and data mining, information visualization, human–computer interaction, enterprise systems, and social network analysis, with a particular interest in mining networks for knowledge management and business intelligence purposes. Her research has appeared in various journals, including *Journal of the Association for Information Systems*, *ACM Transactions on Information Systems*, *IEEE Transactions on Systems, Man and Cybernetics*, *Communications of the ACM*, and *Journal of American Society for Information Science and Technology*.