Available online at www.sciencedirect.com
ScienceDirect
Int. J. Human-Computer Studies 65 (2007) 57–70
International Journal of Human-Computer Studies
www.elsevier.com/locate/ijhcs

Mining communities and their relationships in blogs: A study of online hate groups

Michael Chau[a,*], Jennifer Xu[b]

[a]School of Business, The University of Hong Kong, Pokfulam, Hong Kong
[b]Department of Computer Information Systems, Bentley College, Waltham, MA 02452, USA

Available online 10 October 2006

Abstract

Blogs, often treated as the equivalence of online personal diaries, have become one of the fastest growing types of Web-based media. Everyone is free to express their opinions and emotions very easily through blogs. In the blogosphere, many communities have emerged, which include hate groups and racists that are trying to share their ideology, express their views, or recruit new group members. It is important to analyze these virtual communities, defined based on membership and subscription linkages, in order to monitor for activities that are potentially harmful to society. While many Web mining and network analysis techniques have been used to analyze the content and structure of the Web sites of hate groups on the Internet, these techniques have not been applied to the study of hate groups in blogs. To address this issue, we have proposed a semi-automated approach in this research. The proposed approach consists of four modules, namely blog spider, information extraction, network analysis, and visualization. We applied this approach to identify and analyze a selected set of 28 anti-Blacks hate groups (820 bloggers) on Xanga, one of the most popular blog hosting sites. Our analysis results revealed some interesting demographical and topological characteristics in these groups, and identified at least two large communities on top of the smaller ones. The study also demonstrated the feasibility in applying the proposed approach in the study of hate groups and other related communities in blogs.

Keywords: Blogs; Social network analysis; Hate groups; Web mining

1. Introduction

Blogs, or Weblogs, have become increasingly popular in recent years. Blog is a Web-based publication that allows users to add content periodically, normally in reverse chronological order, in a relatively easy way. Blogs also combine personal Web pages with tools to make linking to other pages easier, as well as to post comments and afterthoughts (Blood, 2004). Instead of having a few people being in control of the threads on traditional Internet forums, blogs basically allow anyone to express their ideas and thoughts. Many free blog hosting sites are available. Sites like www.blogger.com, www.xanga.com, and www.livejournal.com, are a few examples.

Many communities have emerged in the blogosphere. These could be support communities such as those for technical support or educational support (Nardi et al., 2004), or groups of bloggers who already knew each other in other context, such as a group for a high school or a company. In addition, there are also communities formed by people who share common interests or opinions. Many free blog hosting sites have the function to allow bloggers to link to each other to form explicit groups. Similar to other Web-based media such as Web sites, discussion forums, or chat rooms where hate groups are present (Anti-Defamation League, 2001; CNN, 1999; Glaser et al., 2002), there are also hate groups in blogs that are formed by bloggers who are racists or extremists. The consequences of the formation of such groups on the Internet cannot be underestimated. Hate groups or White supremacist groups like the Ku Klux Klan have started to use the Internet to spread their beliefs, recruit new members, or even advocate hate crimes with considerable success (Anti-Defamation

*Corresponding author. Tel.: +852 2859 1014; fax: +852 2858 5614.
E-mail addresses: mchau@business.hku.hk (M. Chau), jxu@bentley.edu (J. Xu).

1071-5819/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ijhcs.2006.08.009

League, 2001). The Web has allowed these groups to reach much further into society than ever before. Young people, the major group of bloggers, are more likely to be affected and even "brainwashed" by ideas propagated through the Web as a global medium. Hatred and extremism ideas could easily be embedded into their minds to make them become members of these hate groups or even conduct hate crimes.

Facing the new trend in the cyberspace, our study has two objectives. First, we propose a semi-automated approach that combines blog spidering and social network analysis techniques to facilitate the monitoring, study, and research on the networks of bloggers, especially those in hate groups. To study the cyber activities of hate groups in blogs, it is important to devise an efficient and effective way to identify these groups, extract the information of their members, and explore their relationships. In recent years, advanced techniques such as text mining, Web mining, and social network analysis have been widely used in studies on cyber crimes, online extremist groups, and terrorist organizations (e.g., Burris et al., 2000, Chen et al., 2004, Zhou et al., 2005). However, the application of these techniques to blog analysis on the Web is a new area and no prior research has been published. Our approach consists of a set of techniques for identifying and analyzing hate groups in the blogosphere on the Web. Moreover, we include in our approach topological analysis, a new network analysis methodology that has not been employed in prior hate group and blog-related research, to study the relationships between bloggers.

Second, our study seeks insights into the organization and movement of online hate groups. We reveal the structural properties of social networks of bloggers through a case study and compare our findings with those from previous studies. Like many other extremist organizations, hate groups and their activities are a type of social movement, which has significant social and political implications (Burris et al., 2000; Douglas et al., 2005). Because of their tendency toward violence, association with crimes, and especially, potential influence on youths, hatred and hate groups have attracted much attention from academics recently. Moreover, since such a movement is dynamic and keeps changing over time, it is desirable to constantly monitor the changes, compare findings across studies, and continuously update our understanding of hate groups.

The remainder of the paper is organized as follows. In Section 2 we review the research background of hate group analysis and related research in text mining, Web mining, and social network analysis. We pose our research questions in Section 3, and a semi-automated approach to hate group analysis in blogs is presented in Section 4. In Section 5 we present a case study that we have performed on a popular blog hosting site based on the proposed approach and discuss our analysis findings. Lastly, we conclude our research in Section 6 and suggest some directions for future work.

2. Related work

In this section we will review the background of the increasingly popular blogging phenomena. We also review relevant techniques in Web mining and social network analysis that have been applied in analyzing Web contents.

2.1. Blogging

Blogs have become increasingly popular in the past few years. In the early days, blogs, a short form of weblogs, were used mainly for pages where links to other useful resources were periodically "logged" and posted. At that time blogs were mostly maintained by hand (Blood, 2004). After easy-to-use blogging software became widely available in the early 2000s, the nature of blogs has changed and many blogs are more like personal Web sites that contain various types of content (not limited to links) posted in reverse-chronological order. Bloggers often make a record of their lives and express their opinions, feelings, and emotions through writing blogs (Nardi et al., 2004). Many bloggers consider blogging as an outlet for their thoughts and emotions. Besides personal blogs, there are also blogs created by companies. For example, ice.com, an online jewelry seller, has launched three blogs and reported that thousands of people linked to their Web site from these blogs (Hof, 2005).

One of the most important features in blogs is the ability for any reader to write a comment on a blog entry. On most blog hosting sites, it is very easy to write a comment, in a way quite similar to replying to a previous message in traditional discussion forums. The ability to comment on blogs has facilitated the interaction between bloggers and their readers. On some controversial issues, like those related to racism, it is not uncommon to find a blog entry with thousands of comments where people dispute back and forth on the matter.

Cyber communities have also emerged in blogs. Communities in blogs can be categorized as explicit communities or implicit communities, like some other cyber communities on the Web (Kumar et al., 1999). Explicit communities in blogs are the groups, or blogrings, that bloggers have explicitly formed and joined. Most blog hosting sites allow bloggers to form a new group or join any existing groups. On the other hand, implicit communities can only be defined by the interactions among bloggers, such as subscription, linking, or commenting. For example, a blogger may subscribe to another blog, meaning that the subscriber can get updates when the subscribed blog has been updated. A blogger can also post a link or add a comment to another blog, which are perhaps the most traditional activities among bloggers. These interactions signify some kind of connection between two bloggers. Because of such interactions among bloggers, these communities are less similar to the *cyber communities* as discussed in Kumar et al. (1999) but more resembling to the *virtual communities* which involve the social interaction

between members characterized by memberships, sense of belonging, relationships, shared values and practices, and self-regulation (Erickson, 1997; Roberts, 1998; Rheingold, 2000). Similar to the analysis of hyperlinks among Web pages to identify communities (Chau et al., 2005a,b), analysis of the connections between bloggers could also identify these virtual communities, their characteristics, and their relationships.

2.2. Web mining and social network analysis

Techniques based on both Web mining and social network techniques have been used in intelligence- and security-related applications and achieved considerable success. Web mining techniques are important because the Web has provided a vast amount of publicly accessible information that could be useful in security applications. It has been reported that many terrorists and extremist groups have been using the Web for various purposes (Zhou et al., 2005; Qin et al., 2006). Analysis and mining of such content can provide useful insights that are important to national and international security. Similarly, social network analysis (SNA) techniques also have been increasingly used in security applications in recent years. As SNA analyzes the interactions between individuals, it can often identify communities within a group of individuals and reveal some other interesting findings. This is especially useful for analyzing criminal organizations and terrorist groups and promising results have been reported (Xu and Chen, 2005). In the rest of this subsection, we give a brief review of Web mining and SNA techniques that are relevant to this research.

Web mining techniques have been widely applied to various Web applications in recent years. In these applications, Web spiders usually are an indispensable component. Spiders, also known as crawlers, wanderers, or Webbots, are defined as "software programs that traverse the World Wide Web information space by following hypertext links and retrieving Web documents by standard HTTP protocol" (Cheong, 1996). Since the early days of the Web, spiders have been widely used to build the underlying databases of search engines (e.g., Pinkerton, 1994), to perform personal search (e.g., Chau et al., (2001)), to archive particular Web sites or even the whole Web (e.g., Kahle, 1997), or to collect Web statistics (e.g., Broder et al., 2000).

Web mining techniques can be categorized into three types: content mining, structure mining, and usage mining (Kosala and Blockeel, 2000). Web content mining refers to the discovery of useful information from Web contents, including text, images, audio, video, etc. Web content mining research includes resource discovery from the Web (e.g., Cho et al., 1998; Chakrabarti et al., 1999), document categorization and clustering (e.g., Zamir and Etzioni, 1999; Chen et al., 2003), and information extraction from Web pages. Web structure mining studies the model underlying the hyperlink structures of the Web. It usually involves the analysis of in-links and out-links information of a Web page, and has been used for search engine result ranking and other Web applications. Google's PageRank (Brin and Page, 1998) and HITS (Kleinberg, 1998) are the two most widely used algorithms. Web usage mining employs data mining techniques to analyze search logs or other activity logs to find interesting patterns. One of the main applications of Web usage mining is its use to learn user profiles (e.g., Armstrong et al., 1995).

Specifically, the identification of Web communities has to a large extent relied on Web structure mining (Gibson et al., 1998; Kumar et al., 1999). Many of existing community identification methods are rooted in the HITS algorithm (Kleinberg, 1998). Kumar et al. (1999) propose a trawling approach to find a set of core pages containing both authoritative and hub pages for a specific topic. The core is a directed bipartite subgraph whose node set is divided into two sets with all hub pages in one set and authoritative pages in the other. The core and the other related pages constitute a Web community (Gibson et al., 1998). Treating the Web as a large graph, the problem of community identification can also be formulated as a minimum-cut problem, which finds clusters of roughly equal sizes while minimizing the number of links between clusters (Flake et al., 2000, 2002). Realizing that the minimum-cut problem is equivalent to the maximum-flow problem (Ford and Fulkerson, 1956), Flake et al. (2000) formulate the Web community identification problem as an $s-t$ maximum flow problem, which can be solved using efficient polynomial time methods.

Recently, a number of hierarchical clustering methods have also been proposed for identifying community structure in networks. These methods are especially suitable for unweighted networks such as the Web, in which hyperlinks do not have associated weights. The G–N algorithm (Girvan and Newman, 2002), for example, is a divisive clustering algorithm that gradually removes links to break a connected graph into clusters. However, the algorithm is rather slow. A few alternative methods such as the modularity-based algorithm (Newman, 2004b) and the edge clustering coefficient based algorithm (Radicchi et al., 2004) have improved the efficiency of the G–N algorithm.

Web structure mining, on the other hand, also has a big overlap with social network analysis. SNA is a sociological methodology for analyzing patterns of relationships and interactions between social actors in order to discover the underlying social structure (Wasserman and Faust, 1994). Not only the attributes of social actors, such as their age, gender, socioeconomic status, and education, but also the properties of relationships between social actors, such as the nature, intensity, and frequency of the relationships, are believed to have important implications to the social structure. SNA methods have been employed to study organizational behavior (Borgatti and Foster, 2003), inter-organizational relations (Stuart, 1998), citation patterns (Baldi, 1998), virutal communities (Garton et al., 1999), and many other domains. Recently, SNA has also been

used in the intelligence and security domain to analyze criminal and terrorist networks (Dombroski and Carley, 2002; Krebs, 2001; Xu and Chen, 2004, 2005).

When used to mine a network, SNA can help reveal the structural patterns such as the central nodes which act as hubs, leaders, or gatekeepers, the densely-knit communities or groups, and the patterns of interactions between the communities and groups. These patterns often have important implications to the functioning of the network. For example, the central nodes often play a key role by issuing commands or bridging different communities. The removal of central nodes can effectively disrupt a network than peripheral nodes (Albert et al., 2000).

Moreover, a recent movement in statistical analysis of network topology (Albert and Barabási, 2002) has brought new insights and research methodology to the study of network structure. Networks, regardless of their contents, are classified into three categories: *random network* (Bollobás, 1985), *small-world network* (Watts and Strogatz, 1998), and *scale-free network* (Barabási et al., 1999). In a random network the probability that two randomly selected nodes are connected is a constant p. As a result, each node has roughly the same number of links and nodes are rather homogenous. In addition, communities are not likely to exist in random networks. Small-world networks, in contrast, have a significantly high tendency to form groups and communities. Most empirical networks ranging from social networks (Newman, 2004a), biological networks (Jeong et al., 2001), to the Web (Albert et al., 1999) have been found to be nonrandom networks. In addition, many of these networks are also scale-free networks (Barabási et al., 1999), in which a large percentage of nodes have just a few links, while a small percentage of the nodes have a large number of links. Thus, nodes in scale-free networks are not homogenous in terms of their links. Some nodes become hubs or leaders that play important roles in the operation of the network. The Web has been found to have both small-world and scale-free properties (Albert and Barabási, 2002).

2.3. Hate on the Internet

Hate crimes have been one of the long-standing problems in the United States because of various historical, cultural, and political reasons. Race, gender, religion, and disability often become the reason of hate. Over time, hate groups have been formed to unite individuals with similar beliefs as well as to spread such ideology. For example, White supremacist groups such as Ku Klux Klan (KKK), Neo-Nazis, and Racist Skinheads have been active in the United States for a long time (Burris et al., 2000).

Hate groups have been increasingly using the Internet to express their ideas, spread their beliefs, and recruit new members (Lee and Leets, 2002). It has been reported that 60% of hate criminals are youths (Levin and McDevitt, 1993), who are, perhaps unfortunately, also one of the largest groups of Internet users. Glaser et al. (2002) suggest that racists often express their views more freely on the Internet. The Hate Directory (Franklin, 2005) compiles a list of hundreds of Web sites, files archives, newsgroups, and other Internet resources related to hate and racism. Several studies have investigated Web sites that are related to racism or White supremacy. Douglas et al. (2005) studied 43 Web sites that were related to White supremacy. It was found that while these groups showed lower level of advocated violence due to legal constraints, they exhibited high levels of social conflict and social creativity. Lee and Leets (2002) found that storytelling-style, implicit messages often used by hate groups on the Internet were more persuasive to adolescents, who have become the target of new member recruitment of many hate groups. These adolescences might be easily influenced to conduct hate crimes. Gerstenfeld et al. (2003) conducted a manual analysis of 157 extremist Web sites. They found that some hate Web sites were associated with hate groups while others were maintained by individuals. Many of these sites had links to other extremist sites or hate group sites, showing that some of these groups are linked to each other. Burris et al. (2000) systematically analyzed the networks of Web sites maintained by white supremacist groups and found that this network had a decentralized structure with several centers of influence. In addition, communities were present in this network in which groups sharing similar interests and ideologies tended to be closely connected. Zhou et al. (2005) used software to automate the analysis of the content of hate group Web sites and the linkage among them. They found that one of the major objectives of these Web sites was to share ideology. Cyber communities such as White Supremacists and Neo-Nazis were identified among these sites. Recent years have seen the emergence of hate groups in blogs, where high-narrative messages are the norm. This has made blogs an ideal medium for spreading hatred. Blogs have also made it possible for individuals to find others with similar belief and ideology much more easily. As a result, hate groups have emerged in blogs.

To study these online hate groups in blogs, it is important to analyze the content of these blogs as well as the relationships among the bloggers. However, because of the large volume of data involved, it is often a mentally exhausting, if not infeasible, process to perform such kind of analysis manually. There is great potential value to apply Web mining and social network analysis techniques, which have been used successfully in other security-related applications, to analyze hate-related blogs automatically in order to identify patterns and facilitate further analysis. However, we have not been able to identify any prior research in this aspect in the literature.

3. Research questions

As discussed earlier, it is an important and timely issue to identify the hate groups in blogs and analyze their relationships. Web mining techniques have been used to

analyze Web content such as Web pages and hyperlinks; however, few of these techniques have been applied to blog analysis. Based on our review, we pose the following two sets of research questions: (1) Can we use semi-automatic techniques, such as automated text collection, text analysis, and network analysis, to identify hate groups in blogs? (2) What are the structural properties of the social networks of bloggers in the hate groups? Are there bloggers who stand out as leaders of influence in these groups? What is the community structure in these groups? What do the structural properties suggest about the organization of the hate groups? What are the social and political implications of these properties?

4. Proposed approach

In this section, we propose a semi-automated approach for identifying groups and analyzing their relationships in blogs. The approach is diagrammed in Fig. 1. Our approach consists of four main modules, namely Blog Spider, Information Extraction, Network Analysis, and Visualization. The Blog Spider module downloads blog pages from the Web. These pages are then processed by the Information Extraction module. Data about these blogs and their relationships are extracted and passed to the Network Analysis module for further analysis. Finally the Visualization module presents the analysis results to users in a graphical display. In the following, we describe each module in more detail.

4.1. Blog spider

A blog spider program is first needed to download the relevant pages from the blogs of interest. Similar to general Web fetching, the spider can connect to blog hosting sites using standard HTTP protocol. After a blogring description page or a blog page is fetched, URLs are extracted and stored into a queue. However, instead of following all extracted links, the blog spider should only follow links that are of interest, e.g., links to a group's members, other bloggers, comment links, and so on. Links to other external resources are often less useful in blog analysis. Multi-threading also can be used, as in standard spiders, such that multiple Web pages can be downloaded in parallel (Chau et al., 2005a,b). This can avoid bottleneck in the process if any particular Web server is sending malicious response or not responding at all. Alternatively, asynchronous I/O can be used for parallel fetching (Brin and Page, 1998). In either case, after a page is downloaded it can be stored into a relational database or as a flat file. In addition, the spider can use RSS (Really Simple Syndication) and get notification when the blog is updated. However, this is only necessary when monitoring or incremental analysis is desired.

4.2. Information extraction

After a blog page has been downloaded, it is necessary to extract useful information from the page. This includes information related to the blog or the blogger, such as user profiles and date of creation. This can also include linkage information between two bloggers, such as linkage, commenting, or subscription. Because different blogs may have different formats, it is not a trivial task to extract this information from blogs. Even blogs hosted on the same hosting site could have considerably different formats as they can be easily customized by each blogger. Pattern matching or entity extraction techniques can be applied. For example, rule-based algorithms that rely on hand-crafted rules can be used to extract useful information such

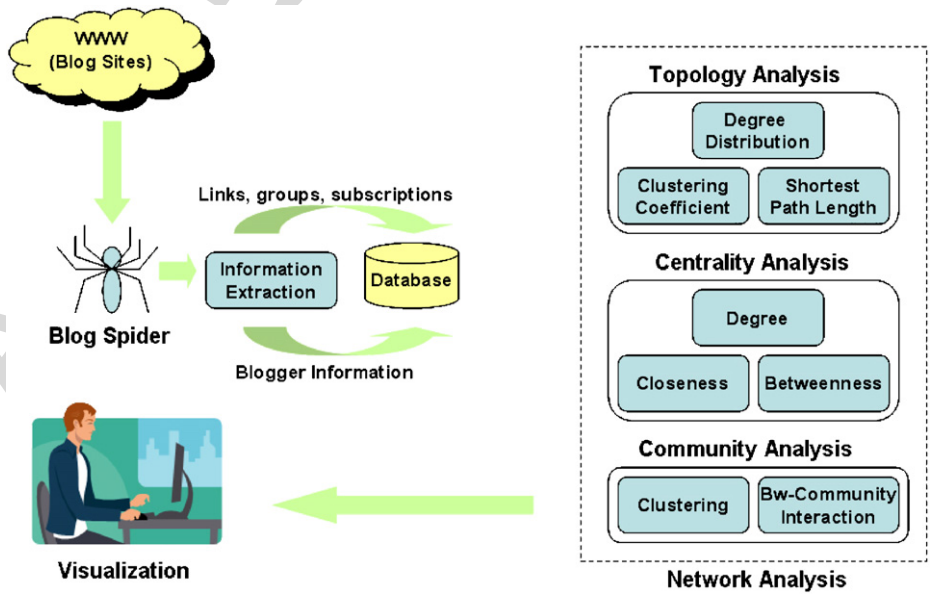

Fig. 1. The proposed semi-automated approach for blog link analysis.

as named entities. The rules may be structural, contextual, or lexical (Krupka and Hausman, 1998). Although such human created rules are usually of high quality, this approach is not scalable to large number of entities and not robust to less structured documents such as blogs. Fortunately, some standard information such as name and location are oftentimes put into specific format (e.g., as a sidebar) in large blog hosting sites, and simple rules should suffice. Nonetheless, different rules may be needed for different blog hosts. Other techniques, such as statistical approach or machine learning techniques, also can be applied if more detailed information needs to be extracted. The information extracted can be stored into database for further analysis. Links extracted can be passed back to the blog spider for further fetching.

4.3. Network analysis

Network analysis is a major component in our approach. In this module we propose three types of analysis: *topological analysis centrality analysis* and *community analysis*.

The goal of topological analysis is to ensure that the network extracted based on links between bloggers is not random and it is meaningful to perform the centrality and community analysis. We use three statistics that are widely used in topological studies to categorize the extracted network (Albert and Barabási, 2002): *average shortest path length*, *clustering coefficient* and *degree distribution*. Average path length is the mean of the all-pair shortest paths in a network. It measures the efficiency of communication between nodes in a network. Clustering coefficient indicates how likely nodes in a network form groups or communities. The degree distribution, $P(k)$, is the probability that a node has exactly k links. Another measure related to average path length is the network's global efficiency, which is defined as the average of the inverses of shortest path lengths over all pairs of nodes in a network (Crucitti et al., 2003). It is shown that a random network usually has a small average path length and is more efficient because an arbitrary node can reach any other node in a few steps. Small-world networks usually have significantly high clustering coefficients than their random network counterparts. Scale-free networks are categorized by a power-law degree distribution, which is different from the Poisson degree distribution presented in random networks.

Centrality analysis follows the topological analysis if the extracted network is shown to be a nonrandom network in which node degrees may vary greatly. The goal of centrality analysis is to identify the key nodes in a network. Three traditional centrality measures can be used: *degree*, *betweenness*, and *closeness* (Freeman, 1979). Degree measures how active a particular node is. It is defined as the number of direct links a node has. "Popular" nodes with high degree scores are the leaders, experts, or hubs in a network. It has been shown that these popular nodes can be a network's "Archilles' Heel," whose failure or removal will cause the network to quickly fall apart (Albert et al., 2000; Holme et al., 2002). In the intelligence and security context, the removal of key offenders in a criminal or terrorist network is often an effective disruptive strategy (McAndrew, 1999; Sparrow, 1991). Betweenness measures the extent to which a particular node lies between other nodes in a network. The betweenness of a node is defined as the number of geodesics (shortest paths between two nodes) passing through it. Nodes with high betweenness scores often serve as gatekeepers and brokers between different communities. They are important communication channels through which information, goods, and other resources are transmitted or exchanged (Wasserman and Faust, 1994). The removal of nodes with high betweenness scores can be even more devastating than the removal of nodes with high degrees (Holme et al., 2002). Closeness is the sum of the length of geodesics between a particular node and all the other nodes in a network. A node with low closeness may find it very difficult to communicate with other nodes in the network. Such nodes are thus more "peripheral" and can become outliers in the network (Sparrow, 1991; Xu and Chen, 2005).

Community analysis is to identify social groups in a network. In SNA a subset of nodes is considered a community or a social group if nodes in this group have stronger or denser links with nodes within the group than with nodes outside of the group (Wasserman and Faust, 1994). A weighted network in which each link has an associated weight can be partitioned into groups by maximizing the within-group link weights while minimizing between-group link weights. Because the link weight represents node similarity or link strength and intensity, nodes in the same group are more similar to each other or more strongly connected. An unweighted network can be partitioned into groups by maximizing within-group link density while minimizing between-group link density. In this case, groups are densely-knit subsets of the network. Note that community and groups here do not refer to the explicit groups (blogrings). They refer to a subset of nodes which form implicit clusters through various relationships. In these communities, members subscribe to or post comments to each other's blogs frequently even though they may not belong to the same blogrings.

After a network is partitioned into groups, the between-group relationships become composites of links between individual nodes. In SNA, a method called *blockmodeling* is often used to reveal the patterns of interactions between groups (White et al., 1976). Given groups in a network, blockmodel analysis determines the presence or absence of a relationship between two groups based on the link density (Wasserman and Faust, 1994). When the density of the links between the two groups is greater than a predefined threshold value, a between-group relationship is present, indicating that the two groups interact with each other constantly and thus have a strong relationship. By this means, blockmodeling summarizes individual relational

details into relationships between groups so that the overall structure of the network becomes more prominent.

4.4. Visualization

The extracted network and analysis results can be visualized using various types of network layout methods. Two examples are *multidimensional scaling* (MDS) and *graph layout* approaches. MDS is the most commonly used method for social network visualization (Freeman, 2000). It is a statistical method that projects higher-dimensional data onto a lower-dimensional display. It seeks to provide a visual representation of proximities (dissimilarities) among nodes so that nodes that are more similar to each other are closer on the display, while nodes that are less similar to each other are further apart (Kruskal and Wish, 1978). Graph layout algorithms have been developed particularly for drawing aesthetically pleasing network presentations (Fruchterman and Reingold, 1991). A type of graph layout algorithm called spring embedder, also known as force-directed method (Davidson and Harel, 1996; Eades, 1984; Fruchterman and Reingold, 1991; Kamada and Kawai, 1989) has been widely used to visualize networks. This algorithm treats a network as an energy system in which steel rings (nodes) are connected by springs (links). Nodes attract and repulse each other and finally settle down when the total energy carried by the springs is minimized.

5. Case study

5.1. Focus and Methods

We applied our approach to conduct a case study of hate groups in blogs. As pointed out in our review, hate crimes have been a long-standing problem in the United States and it is important to study the phenomenon of online hate groups on the Internet, particularly in new media such as blogs. We chose to study the hate groups against Blacks. There are two reasons for the focus. First, the nature of hate groups and hate crimes is often dependent on the target "hated" group. By focusing on a type of hate groups it is possible to identify relationships that are more prominent. Second, among different hate crimes, anti-Black hate crimes have been one of the most widely studied (e.g., Burris et al., 2000; Glaser et al., 2002). This allows us to compare our results with previous research in the literature.

To keep the study at a manageable size, we limit our study to the blogs on Xanga (www.xanga.com). According to statistics provided by Alexa (2005), Xanga is the second most popular Web site that is primarily devoted to blogs, only after Blogger (www.blogger.com). It is also ranked 17th in traffic (visit popularity) among all Web sites in English. We chose Xanga over Blogger because Xanga has more prominent features to support subscriptions and groups (as blogrings). These features are useful for the identification of hate groups in the blogs and the relationships between bloggers.

After choosing our focus, we had to identify a set of hate groups on Xanga. We used the search feature in Xanga to semi-automate the task. First, a set of terms related to Black-hatred, such as "KKK", "niggers", "white pride", were identified. We used these terms to search for groups (blogrings) on Xanga that have any of these words in their group name or description. We then checked these groups and filtered out those not related to anti-Black. Groups with only one single member, which were likely to have been formed by one blogger with no one else joining afterwards, were also removed from our list. This resulted in a set of 40 groups. While most of these groups showed some beliefs of racism or White supremacies, we tried to further narrow these down to groups that demonstrated explicit hatred, so as to make sure that our analysis focused on "hate groups". Therefore, we manually checked these groups and only included those that explicitly mentioned hatred (e.g., "I hate black people", "hate the black race") or used offensive languages (e.g., "nigger beaters", four-letter words) towards the Blacks in their group name or description. Finally we had a list of 28 groups. These groups are listed in Table 1.

In order to further justify the validity of the list of the 28 groups identified, we tried to compare this list with the "racist blogs" list in the Hate Directory (Franklin, 2005). However, only 14 bloggers (10 hosted on Xanga) were listed in the Hate Directory, and no blogrings were listed. To make the comparison possible, we compared our list with the blogrings that these bloggers belong to. As two out of the 10 bloggers on Xanga were no longer available, we had 8 bloggers for our comparison. We found that all these 8 bloggers are members of at least one of the groups identified in our list. Five of these bloggers are members of "! White Power !", the biggest blogring identified in our list. While the comparison was not direct, it showed that our list of groups had covered all of the racist bloggers listed in the Hate Directory, who were some of the more prominent bloggers.

Spiders were used to automatically download the description page and member list of each of these groups. A total of 820 bloggers were identified from these 28 groups. The spiders further downloaded the blogs of each of these bloggers. The extraction program was then executed to extract the information of each blogger, including user id, real name, date of creation, date of birth, city, state, and country. One should note that these data were self-reported; they could be fraud or even missing.

The extraction program also analyzed the relationship between these bloggers. In this study, two types of relationships were extracted:

1. *Group co-membership*: two bloggers belong to the same group (blogring). This is an undirected relationship with an integer weight (based on the number of groups shared by the two bloggers).

Table 1
The 28 groups (blogrings) identified in our analysis

Blogring name	URL	No. of members
! White Power !	http://www.xanga.com/groups/group.aspx?id=76863	371
KKK white is right	http://www.xanga.com/groups/group.aspx?id=84971	135
The KKK (Ku Klux Klan)	http://www.xanga.com/groups/group.aspx?id=164821	67
Are u racist? hate queers and niggers? Me too.	http://www.xanga.com/groups/group.aspx?id=191887	67
Angry and White	http://www.xanga.com/groups/group.aspx?id=258062	48
! ! ! !Hatred 4 SoCiEtY! ! ! !	http://www.xanga.com/groups/group.aspx?id=709048	43
ALL NiGgErS sTiNk	http://www.xanga.com/groups/group.aspx?id=107296	40
I HATE BLACK PEOPLE	http://www.xanga.com/groups/group.aspx?id=525845	37
::White::Power::	http://www.xanga.com/groups/group.aspx?id=261285	37
Nigger beaters	http://www.xanga.com/groups/group.aspx?id=58711	29
WHITE f**kin PRIDE	http://www.xanga.com/groups/group.aspx?id=240012	24
KKK WE GONNA KILL THE NIGGERS!!!	http://www.xanga.com/groups/group.aspx?id=250810	21
I HATE THE FREAKIN PORCH MONKEYS	http://www.xanga.com/groups/group.aspx?id=1066584	14
White-power	http://www.xanga.com/groups/group.aspx?id=1244436	9
** WHITE POWER NATION **	http://www.xanga.com/groups/group.aspx?id=1298240	8
N I G G E R S L A Y E R	http://www.xanga.com/groups/group.aspx?id=1184100	7
K.K.K. MEMBERS	http://www.xanga.com/groups/group.aspx?id=1447382	5
The "i like to beat negros and mexicans" blog rin	http://www.xanga.com/groups/group.aspx?id=387326	5
Honor + The + Ku + Klux + Klan	http://www.xanga.com/groups/group.aspx?id=323614	4
THE REAL KU KLUX KLAN	http://www.xanga.com/groups/group.aspx?id=1315712	4
Dj aNDIE	http://www.xanga.com/groups/group.aspx?id=794267	4
I hate G-Unit	http://www.xanga.com/groups/group.aspx?id=916827	4
∼I Hate negros∼	http://www.xanga.com/groups/group.aspx?id=1470409	4
Niggerzsmellfunny	http://www.xanga.com/groups/group.aspx?id=317512	4
Ku Klux Klan_White Knights Of America	http://www.xanga.com/groups/group.aspx?id=1014877	3
'*∼NEGRO HATERS∼*'	http://www.xanga.com/groups/group.aspx?id=325247	2
Black Haters and Negro Hangers	http://www.xanga.com/groups/group.aspx?id=1401428	2
||||| I HATE N1GGERS |||||	http://www.xanga.com/groups/group.aspx?id=1733297	2

2. *Subscription*: blogger A subscribes to blogger B. This is a directed, binary relationship.

The first type of relationship (group co-membership) assumes that two bloggers are related as long as they join the same blogring. Co-membership relationship is important as it often reflects that the two bloggers involved have similar belief or ideology. The resulting network based on such relationships consisted of several fully-connected cliques. Each clique corresponded to a blogring, in which each node was connected with every other node. This made the nodes indistinguishable from each other in terms of degrees. Thus, we included in our network only co-membership links whose weight was greater than one. That is, we considered a co-membership link between two bloggers a valid link only if they shared memberships of at least two common groups. For the second type of relationship, we included all subscription links. Subscription link is important because it means that the subscriber is interested in reading the subscribed blogs. It suggests that there is an information flow between the two individuals.

5.2. Analysis and results

After collecting the blogs and extracting information from them, we performed demographical and network analysis on the data set in order to reveal the characteristics of these groups and ascertain whether any patterns exist. Visualization was then applied to present the results. We discuss the details of our analysis in the following sections.

5.2.1. Demographical analysis

We provide a brief summary of the demographical information of the bloggers of interest and the growth patterns of the blog space of hate groups. As in many other Internet-based media such as forums and chat rooms, the real identities of bloggers are unknown. Thus, the self-reported demographical information of bloggers is subject to the problems of anonymity. However, since blogs are often personal online diaries many bloggers still choose to release partial information about their demographics such as gender and country. Many bloggers even post their personal photos on their blogs making the true gender information accessible, assuming the photos are really those of the bloggers themselves.

Among the 820 bloggers in our data set, 659 explicitly indicate their gender. Sixty three percent of them are male and 37% are female. These bloggers are from various countries. Among the 529 bloggers who explicitly report the country information, 81.9% are from the United States. The two next countries are Germany and Afghanistan, taking on 2.6% and 2.1%, respectively. The remaining 13.4% of bloggers are from 43 other countries. It can be seen that hate groups are dominated by male bloggers from the United States. However, we are aware that this

finding is based on the problematic source of demographic information. The actual distribution of gender and country may be different.

Nevertheless, this finding provides the evidence that the Internet and blogs have become convenient media for hate groups to penetrate into different countries and regions. The anti-Black ideology and Black-hatred groups are no longer unique to specific Western countries but have the potential to reach international audience. This finding is consistent with previous studies on online hate groups (Burris et al., 2000; Gerstenfeld et al., 2003). For example, Burris et al. (2000) found that the Internet had made it possible and rather easy for white supremacist groups to form a transnational cyber-community via hyperlinks between their Web sites.

We also analyzed the growth of the hate group blogs over the years. Unlike demographical information, the exact time when a blogger registered on the blog hosting site (Xanga) is recorded by the server and thus is generally not subject to fraud. Fig. 2 presents the growth of the number of bloggers who registered on and joined the anti-Black blogrings since the first quarter of 2002. The number increased steadily between 2003 and the third quarter of 2004 and started to fluctuate since the forth quarter of 2004. This may be because some bloggers who have recently registered have not joined those popular blogrings or have not formed into large communities. As a result, some of them were not included in the data set after we filtered the raw data. Fig. 2 implies that hate groups have been gaining popularity in blog space over years as more and more individual bloggers joined these groups. Such a trend should not be underestimated because the ideas, beliefs, and opinions advocated by racists and extremists may pose potential threats to the society. More importantly, if more youths join these anti-Black blogrings, they can be influenced by the ideology easily and become future members of extremist organizations. As extremist organizations gain more members, popularity, and powers, their advocated illegal activities and crimes would substantially hamper the stability of the society.

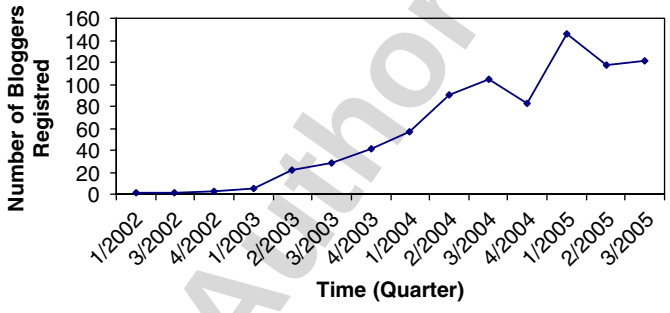

Fig. 2. The number of hate-group bloggers registered over the years.

5.2.2. Topological analysis

When analyzing the topology of the network, we ignored the weight of co-membership relationship and the direction of subscription. We connected two nodes (bloggers) if they belonged to at least two common groups or one subscribed to the other. As a result, there could only be at most one link between a pair of nodes. The resulting network was an unweighted, undirected network consisting of 1193 links. This network was not a connected graph in that it consisted of several disjoint components, between which no link existed. The largest connected component, often called a giant component in graph theory (Bollobás, 1985), contained 273 nodes connected by 1115 links. This giant component was a rather dense graph with an average node degree of 8.2.

We performed topological analysis for the giant component. Table 2 provides the statistics of the average shortest path length, global efficiency, and clustering coefficient. To compare the giant component with its random graph counterpart, we generated 30 random networks consisting of the same number of nodes (273) and links (1115) with the giant component. The resulting statistics are also reported in Table 2. It shows that the giant component is slightly less efficient than its random graph counterpart. On average, a node in the giant component must take 0.75 more steps than in the random graph to reach another arbitrary node. Although it is less efficient than a random network, the giant component is indeed an efficient network. It takes a blogger no more than four steps (three intermediate bloggers) to reach another arbitrary one among the 273 bloggers. This short path length has an important implication to the diffusion and communication of information in the network. In other words, new ideas, opinions, beliefs, and propaganda can be quickly distributed among the bloggers in the network, making it easier for bloggers to influence each other.

Moreover, the giant component has a significantly higher clustering coefficient, which is 31 times more than its random graph counterpart. This implies that the giant component is a small world, in which densely-knit communities are very likely to exist. Communities have also been observed in other online media of hate groups. For example, Burris et al. (2000) found that there was a community of Holocaust Revisionist Web sites, consisting

Table 2
The topological properties of the giant component (number in the parentheses are standard deviations)

	Average shortest path length	Global efficiency	Clustering coefficient
Giant component	3.62	0.33	0.37
Random counterpart	2.89 (0.03)	0.37 (0.00)	0.03 (0.00)

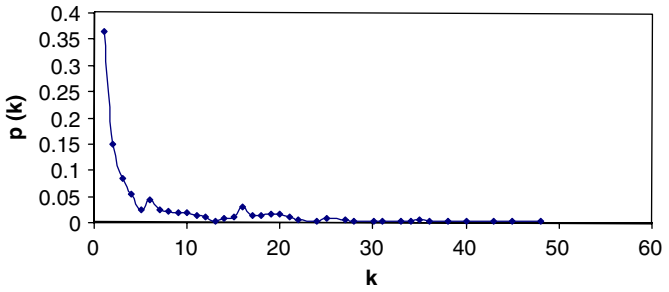

Fig. 3. The degree distribution of the giant component.

of 11 sites connected by dense hyperlinks. A community plays an important role in reinforcing the interests and beliefs of its members, and helps create a "collective identity" (Gerstenfeld et al., 2003). By comparing with other members in the community, one may feel that his/her extremism ideas and beliefs are shared by many other people and thus are not extreme at all.

The degree distribution of the giant component seems to follow a power-law distribution (see Fig. 3). The most distinctive feature of a power-law distribution curve is its long tail for large degree (k), which is significantly different from a bell-shaped Poisson distribution. The long tail indicates that a small number of nodes in the network have a large number of links and they are the "centers" of the network. Actually, the power-law degree distribution manifests the "rich-get-richer" phenomenon (Albert and Barabási, 2002), where nodes with many links are more capable of attracting links. If a blogger has many subscribers, he/she may attract more bloggers to visit, make comments on, and subscribe to his/her blog. Such a blogger become an "authoritative" (Kleinberg, 1998) and influential leader in the network. His/her beliefs and even advocated crimes could influence more people faster. On the other hand, a blogger subscribing to many other bloggers can appear as a "hub" that serves as a pointer and channel to many other bloggers including those central, influential leaders.

5.2.3. SNA and visualization

We used a prototype system we developed (Xu and Chen, 2005) to find the central nodes and to identify the communities. The system has a graphical user interface to facilitate interaction between users and the system (see Fig. 4). The user interface visualizes the network and presents the results of social network analysis. In the visualization, each node represents a blogger. A straight line connecting two nodes indicates that the two corresponding bloggers either co-register in more than one blogring, or one blogger subscribes to the other. Note that the directions of subscription links are not shown.

The layout of the network is determined using the MDS method. In order to position nodes which are likely to belong to the same community close to each other on the display, we assigned each link an "edge clustering coefficient", which measures the likelihood of two incident nodes of the link to form a cluster (Radicchi et al., 2004).

The community analysis can be performed by adjusting the "level of abstraction" slider at the bottom of the panel. At the lowest level of abstraction, each individual node and link are presented. As the abstraction level increases, the system employs hierarchical clustering method to gradually merge nodes, which are connected by links of high edge clustering coefficients. When the highest abstraction level is reached, the whole giant component becomes a big community.

At any level of abstraction, a circle represents a community. The size of the circle is proportional to the number of bloggers in the community. Straight lines connecting circles represent between-group relationships, which are extracted using blockmodel analysis. The thickness of a line is proportional to the density of the links between the two corresponding communities.

Fig. 4(a) presents the giant component at its lowest abstraction level. The bloggers who have the highest degree and betweenness scores are highlighted and labeled with their usernames. These bloggers are those who may participate in multiple blogrings or have many subscription relationships with other bloggers. It is interesting to see that in addition to joining explicit groups (blogrings), bloggers have also formed implicit communities through co-membership and subscription. Three circles of nodes are apparently communities in which bloggers share many common memberships. The communities found by the system (see Fig. 4(c)) also confirm this pattern. The two circles indicate that at least two big communities exist among the bloggers. Moreover, the two communities also interact with each other very actively, as represented by the thick line between the two circles.

Because the communities can be analyzed at different levels, the two big communities may consist of smaller communities. For example, the inner structure of the bigger community shows that two smaller circles of nodes, which may also be communities, exist in the community (see Fig. 4(d)). The Fig. 4(d) also displays at the right-hand side of the screen the rankings of the community members in terms of their centrality scores within the specific community. These three centrality measures are interpreted as the leader role, gatekeeper role, and outlier role. Therefore, a blogger which ranks the highest in degree in the community would likely be the centers of the community. Actually, these central bloggers may be either authoritative leaders or hubs. The leaders with many in-links from subscribers can spread their beliefs and opinions to many of their subscribers; and the hubs with many out-links may facilitate such a spread of influence by directing their visitors to the leaders via subscription links.

To separate the co-membership links and subscription links we also present the network with only subscription links in Fig. 4(b). Comparing Fig. 4(a) and (b) one can see that among the five highlighted nodes, two bloggers (the upper right one and the lower left one) stand out due to

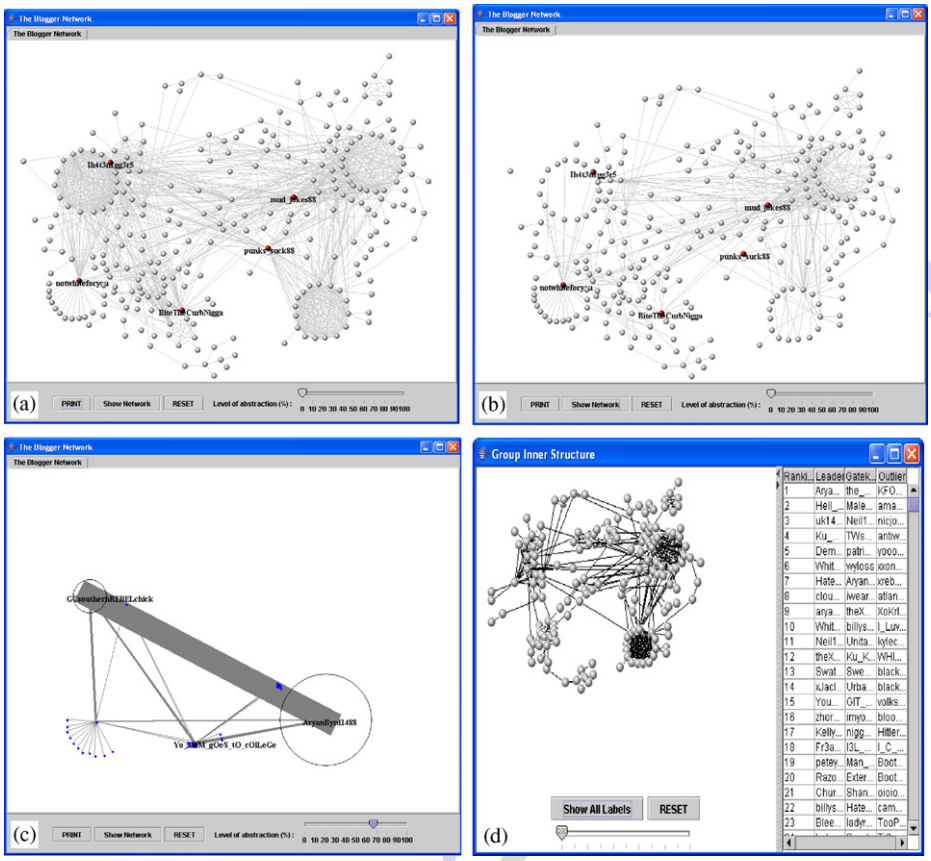

Fig. 4. The prototype system for SNA and visualization. (a) The giant component with both co-membership and subscription links. The highlighted nodes are those who have large degree or betweenness scores. (b) The giant component with only subscription links. Note that the directions of subscription are omitted. (c) The two major communities found in the network. (d) The inner structure of a community which is represented by the large circle in (c).

their large number of subscription links. Because the system interface does not show the directions of the subscription links it is difficult to visually determine whether they are leaders or hubs. By calculating the in-degree and the out-degree of these nodes, we found that they were actually hubs who subscribed to many other blogs.

In general, it is worth a closer look at the bloggers with many subscribers and a high in-degree. A content analysis on their blogs may reveal whether they are actually popular leaders who often express extremism ideas, beliefs, and opinions. Because such ideologies may easily be spread among and influence their subscribers, these bloggers need to be closely monitored. On the other hand, the hubs with a high out-degree should also be paid attention since they may be intermediate channels leading to the extremism ideologies.

In addition, although the network in Fig. 4(b) is not as dense as the one in Fig. 4(a), some part of the network is still rather dense indicating a tendency for forming communities. Interestingly, the upper right highlighted node seems to be a joining point that holds a community around it. This further implies that hubs, although they may not be leaders, can be important communication channels.

5.3. Discussion

In this section we compare our findings with previous studies while providing answers to our research questions regarding online hate groups. Specifically, we choose to compare with the findings in Burris et al. (2000), primarily because among the limited number of studies on online hate groups, this article studies this social movement from a social network perspective. The authors focused on the structural properties of the Web site networks among white supremacist groups, making it appropriate for us to align and compare our findings with theirs.

a. *What are the structural properties of the social networks of bloggers in the hate groups*? Similar to the network of white supremacist Web sites (Burris et al., 2000), the network of bloggers in hate groups is decentralized. The data set contains a number of isolated clusters and a giant component consisting of roughly one third of the bloggers. Centralized structures such as star and hierarchical ones are not observed in the network. This is not out of our expectation because the hate groups under study were mostly formed spontaneously. They have not evolved into organizations or associations with formal organizational structure.

b. *Are there bloggers who stand out as leaders of influence in these groups*? Burris et al. (2000) found that the decentralized white supremacist groups had different centers of influence. Our analysis shows that there are some bloggers who are connected with many others through common membership and subscription. These bloggers may be either leaders of opinions and ideologies or hubs of communication. Both types of center require closer analysis and examination.
c. *What is the community structure in these groups*? Communities, represented by densely-knit clusters of Web sites in Burris et al. (2000), are also present in the blogger network. However, these communities are not composed of Web sites but individual bloggers. Communities provide an environment for its members to exchange their ideas and opinions and reinforce the shared ideology. A community becomes the collective social identity of its members and is potential of facilitating the communication and collaboration among its members to carry out illegal activities and crimes. In addition, we observed that the two communities identified are closely connected. It suggests that members of these communities may share common ideologies and interests, and it is not unlikely that they will merge as a larger community in the future.
d. *What do the structural properties suggest about the organization of the hate groups*? As mentioned in point (a), the structure of the network suggests that the hate groups in blogosphere have not formed into centralized organizations. However, it does not eliminate the possibility that these groups may help prepare future members for extremist organizations such as the Ku Klux Klan. Especially, their influence on youths should not be underestimated.
e. *What are the social and political implications of these properties*? Burris et al. (2000) commented that extremist groups are a type of social movement which has profound social and political implications. At this stage of our study we cannot find any evidence that the bloggers in hate groups are affiliated with any political entities or associations. Indeed, there has not been any group that takes a form of organization. However, we do find that the hate groups are gaining popularity among bloggers as more and more individuals join the groups. In addition, these hate groups also show transnational characteristics, meaning that hatred and hate-related illegal activities are not a problem to a particular country or region but an international phenomenon. Such a movement should not be overlooked and ignored by authorities, researchers, and other watchdog organizations such as the Hate Directory (Franklin, 2005).

6. Conclusion and future directions

In this paper, we have discussed the problems of the emergence of hate groups and racism in blogs. Our contributions are twofold. First, we have proposed a semi-automated approach for blog analysis. Our approach consists of a set of Web mining and network analysis techniques that can be applied to the study of blogosphere. Such techniques as network topology analysis, which has not been employed in blog or hate group related research, are proposed and have been demonstrated useful and relevant in this context. While the approach has been proposed and studied in the context of hate group analysis, we have tried to keep the approach general such that it is not specific to such analysis. We believe that the approach can also be applied to other domains that involve virtual community analysis and mining, which we believe would be an increasingly important field for various applications. These applications include not only other security informatics research such as bioterrorism (Raghu and Vinze, In Press) but also other applications such as marketing analysis and business intelligence analysis (Chau et al., 2005a,b).

Second, we applied this approach to investigate the characteristic and structural relationships among the hate groups in blogs in our case study. Hatred and hate groups are a type of social movement which should not be ignored in today's information age. Extremist groups are using the Internet to disseminate their propaganda, spread their ideologies, and recruit new members. These groups have also been present in Blogs, a new online communication and publication tool, thereby posing threat to our society. However, there has not been much research on this movement. We believe that our research is timely and important to the security of society. By disseminating both explicit and implicit hatred messages through blogs, racists can easily target youths with ubiquitous coverage—basically anyone who has access to the Internet—that was never possible in the past. Youths are often easily influenced by these messages and could eventually become terrorists and pose a threat to our society (Blazak, 2001). Our study not only has provided an approach that could facilitate the analysis of law enforcement and social workers in studying and monitoring such activities, but also has brought insights into the structural properties of online hate groups and helped broaden and deepen our understanding of such a social movement.

In the future, we will extend our study in three major directions to address some limitations of the current study. First, the current study only investigated the hate group activities on one single blog site, Xanga. Although Xanga has been reported to have the most number of blogs associated with hate groups (Franklin, 2005), a further study that includes other popular sites such as Blogger would be more comprehensive. Second, only two types of relationships, namely co-membership and subscription, were considered in the present study. It would be interesting to expand our study to include other types of relationships, such as commenting and hyperlinking, in the network analysis. Inclusion of these relationships could reveal other implicit linkages among the bloggers. Third,

in-depth content analysis can be performed on blogs. Important nodes identified by the current approach can be further analyzed using text mining techniques such as co-occurrence analysis or entity extraction (Chau et al., 2002). Such analysis will help further unveil online hate groups.

Acknowledgement

This project has been supported in part by a grant from the University of Hong Kong Seed Funding for Basic Research, "Using Content and Link Analysis in Developing Domain-specific Web Search Engines: A Machine Learning Approach," 10205294 (PI: M. Chau), February 2004—January 2006.

We would like to thank Porsche Lam and Boby Shiu of the University of Hong Kong for their participation and support in this project.

References

Albert, R., Barabási, A.-L., 2002. Statistical mechanics of complex networks. Reviews of Modern Physics 74 (1), 47–97.

Albert, R., Jeong, H., Barabási, A.-L., 1999. Diameter of the world-wide web. Nature 401, 130–131.

Albert, R., Jeong, H., Barabási, A.-L., 2000. Error and attack tolerance of complex networks. Nature 406, 378–382.

Alexa, 2005. Top English language sites. [Online] Retrieved from http://www.alexa.com/site/ds/top_sites?ts_mode=lang&lang=en on October 7, 2005.

Anti-Defamation League, 2001. Poisoning the Web: Hatred online. [Online] Retrieved from http://www.adl.org/poisoning_web/poisoning_toc.asp on October 7, 2005.

Armstrong, R., Freitag, D., Joachims, T., Mitchell, T., 1995. WebWatcher: A learning apprentice for the World Wide Web. In: Proceedings of the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, March 1995.

Baldi, S., 1998. Normative versus social constructivist processes in the allocation of citations: a network-analytic model. American Sociological Review 63 (6), 829–846.

Barabási, A.-L., Albert, R., Jeong, H., 1999. Mean-field theory for scale-free random networks. Physica A 272, 173–187.

Blazak, R., 2001. White boys to terrorist men: Target recruitment of Nazi skinheads. American Behavioral Scientist 44 (6), 982–1000.

Blood, R., 2004. How blogging software reshapes the online community. Communications of the ACM 47 (12), 53–55.

Bollobás, B., 1985. Random Graphs. London, Academic.

Borgatti, S.P., Foster, P.C., 2003. The network paradigm in organizational research: a review and typology. Journal of Management 29, 991–1013.

Brin, S., Page, L., 1998. The anatomy of a large-scale hypertextual web search engine. In: Proceedings of the 7th WWW Conference, Brisbane, Australia, April 1998.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J., 2000. Graph Structure in the Web. Proceedings of the 9th International World Wide Web Conference, Amsterdam, Netherlands, May 2000.

Burris, V., Smith, E., Strahm, A., 2000. White supremacist networks on the internet. Sociological Focus 33 (2), 215–235.

Chakrabarti, S., van den Berg, M., Dom, B., 1999. Focused crawling: a new approach to topic-specific web resource discovery. In: Proceedings of the 8th International World Wide Web Conference, Toronto, Canada, May 1999.

Chau, M., Zeng, D., Chen, H., 2001. Personalized Spiders for Web Search and Analysis. Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries, Roanoke, Virginia, USA, June 24–28, 2001, pp. 79–87.

Chau, M., Xu, J. J., Chen, H., 2002. Extracting Meaningful Entities from Police Narrative Reports. Proceedings of the National Conference for Digital Government Research (dg.o 2002), Los Angeles, California, USA, May, 2002, pp. 271–275.

Chau, M., Qin, J., Zhou, Y., Tseng, C., Chen, H., 2005a. SpidersRUs: Automated development of vertical search engines in different domains and languages. In: Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries, Denver, Colorado, USA, June, 2005, pp. 110–111.

Chau, M., Shiu, B., Chan, I., Chen, H., 2005b. Automated identification of web communities for business intelligence analysis. In: Proceedings of the Fourth Workshop on E-Business (WEB 2005), Las Vegas, USA, December, 2005.

Chen, H., Fan, H., Chau, M., Zeng, D., 2003. Testing a cancer meta spider. International Journal of Human-Computer Studies 59 (5), 755–776.

Chen, H., Chung, W., Xu, J., Wang, G., Qin, Y., Chau, M., 2004. Crime data mining: a general framework and some examples. IEEE Computer 37 (4), 50–56.

Cheong, F.C., 1996. Internet Agents: Spiders, Wanderers, Brokers, and Bots. New Riders Publishing, Indianapolis, Indiana, USA.

Cho, J., Garcia-Molina, H., Page, L., 1998. Efficient crawling through URL ordering. In: Proceedings of the 7th WWW Conference, Brisbane, Australia, April 1998.

CNN (1999). Hate group web sites on the rise. CNN News [Online] Retrieved from http://edition.cnn.com/US/9902/23/hate.group.report/index.html on October 7, 2005.

Crucitti, P., Latora, V., Marchiori, M., Rapisarda, A., 2003. Efficiency of scale-free networks: error and attack tolerance. Physica A 320, 622–642.

Davidson, R., Harel, D., 1996. Drawing graphs nicely using simulated annealing. ACM Transactions on Graphics 15 (4), 301–331.

Dombroski, M.J., Carley, K.M., 2002. NETEST: estimating a terrorist network's structure. Computational and Mathematical Organization Theory 8, 235–241.

Douglas, K.M., McGarty, C., Bliuc, A., Lala, G., 2005. Understanding cyberhate: social competition and social creativity in online white supremacist groups. Social Science Computer Review 23 (1), 68–76.

Eades, P., 1984. A heuristic for graph drawing. Congressus Numerantium 42, 149–160.

Erickson, T., 1997. Social interaction on the net: Virtual community as participatory genre. Proceedings of the Thirtieth Hawaii International Conference on System Sciences, January 7–10, 1997, Wailea, Hawaii, USA.

Flake, G.W., Lawrence, S., Giles, C.L., 2000. Efficient identification of web communities. In: Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD 2000), Boston, MA.

Flake, G.W., Lawrence, S., Giles, C.L., Coetzee, F.M., 2002. Self-organization and identification of web communities. IEEE Computer 35 (3), 66–71.

Ford Jr., L.R., Fulkerson, D.R., 1956. Maximal flow through a network. Canadian Journal of Mathematics 8, 399–404.

Franklin, R.A., 2005. The hate directory. [Online] Retrieved from http://www.bcpl.net/~rfrankli/hatedir.htm on October 7, 2005.

Freeman, L.C., 1979. Centrality in social networks: conceptual clarification. Social Networks 1, 215–240.

Freeman, L.C., 2000. Visualizing social networks. Journal of Social Structure 1 (1).

Fruchterman, T.M.J., Reingold, E.M., 1991. Graph drawing by force-directed placement. Software-Practice and Experience 21 (11), 1129–1164.

Garton, L., Haythornthwaite, C., Wellman, B., 1999. Studying online social networks. Doing Internet Research. Sage Publications, S. Jones. Thousand Oaks, CA, 75–105.

Gerstenfeld, P.B., Grant, D.R., Chiang, C.P., 2003. Hate online: a content analysis of extremist internet sites. Analyses of Social Issues and Public Policy 3, 29–44.

Gibson, D., Kleinberg J., Raghavan, P., 1998. Inferring web communities from link topology. In: Proceedings of the 9th ACM Conference on Hypertext and Hypermedia, Pittsburgh, PA.

Girvan, M., Newman, M.E.J., 2002. Community structure in social and biological networks. In: Proceedings of the National Academy of Science of the United States of America, vol. 99, pp. 7821–7826.

Glaser, J., Dixit, J., Green, D.P., 2002. Studying hate crime with the internet: what makes racists advocate racial violence? Journal of Social Issues 58 (1), 177–193.

Hof, R., 2005. Blogs on ice: Signs of a business model? Business Week Online—The Tech Beat, June 2, 2005. [Online] Retrieved from http://www.businessweek.com/the_thread/techbeat/archives/2005/06/blogs_on_ice_si.html on October 7, 2005.

Holme, P., Kim, B.J., Yoon, C.N., Han, S.K., 2002. Attack vulnerability of complex networks. Physical Review E 65, 056109.

Jeong, H., Mason, S.P., Barabási, A.-L., Oltvai, Z.N., 2001. Lethality and centrality in protein networks. Nature 411 (6833), 41.

Kahle, B., 1997. Preserving the Internet. Scientific America, March 1997.

Kamada, T., Kawai, S., 1989. An algorithm for drawing general undirected graphs. Information Processing Letters 31 (1), 7–15.

Kleinberg, J., 1998. Authoritative sources in a hyperlinked environment. In: Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms, San Francisco, California, USA, January 1998, pp. 668–677.

Kosala, R., Blockeel, H., 2000. Web mining research: a survey. ACM SIGKDD Explorations 2 (1), 1–15.

Krebs, V.E., 2001. Mapping networks of terrorist cells. Connections 24 (3), 43–52.

Krupka, G.R., Hausman, K., 1998. IsoQuest Inc.: Description of the NetOwlTM extractor system as used for MUC-7. In: Proceedings of the Seventh Message Understanding Conference, April 1998.

Kruskal, J.B., Wish, M., 1978. Multidimensional Scaling. Sage Publications, Beverly Hills, CA.

Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A., 1999. Trawling the web for emerging cyber-communities. Computer Networks 31 (11–16), 1481–1493.

Lee, E., Leets, L., 2002. Persuasive storytelling by hate groups online: Examining its effects on adolescents. American Behavioral Scientist 45, 927–957.

Levin, J., McDevitt, J., 1993. Hate crimes: The Rising Tide of Bigotry and Bloodshed. Plenum, New York.

McAndrew, D., 1999. The structural analysis of criminal networks. In: The Social Psychology of Crime: Groups, Teams, and Networks, Offender Profiling Series, III. D. Canter and L. Alison. Dartmouth, Aldershot, pp. 53–94.

Nardi, B.A., Schiano, D.J., Gumbrecht, M., Swartz, L., 2004. Why we blog. Communications of the ACM 47 (12), 41–46.

Newman, M.E.J., 2004a. Coauthorship networks and patterns of scientific collaboration. Proceedings of the National Academy of Science of the United States of America, vol. 101, pp. 5200–5205.

Newman, M.E.J., 2004b. Fast algorithm for detecting community structure in networks. Physical Review E 69 (6), 066133.

Pinkerton, B., 1994. Finding What People Want: Experiences with the WebCrawler. Proceedings of the 2nd International World Wide Web Conference, Chicago, Illinois, USA, 1994.

Qin, J., Zhou, Y., Reid, E., Lai, G., Chen, H., 2006. "Unraveling International Terrorist Groups' Exploitation of the Web: Technical Sophistication, Media Richness, and Web Interactivity," Proceedings of the Workshop on Intelligence and Security Informatics (WISI-06), Singapore, April 9, 2006.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D., 2004. Defining and identifying communities in networks. Proceedings of the National Academy of Science of the United States of America, vol. 101, pp. 2658–2663.

Raghu, T.S., Vinze, A., In Press. A business process context for knowledge management. *Decision Support Systems*, forthcoming.

Rheingold, H., 2000. The Virtual Community: Homesteading on the Electronic Frontier. MIT Press, Cambridge, Massachusetts, USA.

Roberts, T.L., 1998. Are newsgroups virtual communities? Proceedings of the CHI'98 Conference, April 18–23, 1998, Los Angeles, California, USA.

Sparrow, M.K., 1991. The application of network analysis to criminal intelligence: an assessment of the prospects. Social Networks 13, 251–274.

Stuart, T.E., 1998. Network positions and propensities to investigation of strategic alliance formation in a high-technology industry. Administrative Science Quarterly 43, 668–698.

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

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

White, H.C., Boorman, S.A., Breiger, R.L., 1976. Social structure from multiple networks: I. Blockmodels of roles and positions. American Journal of Sociology 81, 730–780.

Xu, J.J., Chen, H., 2004. Fighting organized crime: using shortest-path algorithms to identify associations in criminal networks. Decision Support Systems 38 (3), 473–487.

Xu, J.J., Chen, H., 2005. Crimenet explorer: a framework for criminal network knowledge discovery. ACM Transactions on Information Systems 23 (2), 201–226.

Zamir, O., Etzioni, O., 1999. Grouper: A dynamic clustering interface to web search results. In: Proceedings of the 8th World Wide Web Conference, Toronto, May 1999.

Zhou, Y., Reid, E., Qin, J., Chen, H., Lai, G., 2005. US domestic extremist groups on the web: Link and content analysis. IEEE Intelligent Systems 20 (5), 44–51.