A Framework for Exploring Gray Web Forums: Analysis of Forum-Based Communities in Taiwan

Jau-Hwang Wang[1], Tianjun Fu[2], Hong-Ming Lin[1], and Hsinchun Chen[2]

[1] Department of Information Management, Central Police University,
56 Shu-Ren Road, Ta-Kang, Kwei-Shan, Tao-Yuan, Taiwan, ROC 333
jwang@mail.cpu.edu.tw, im933090@sun4.cpu.edu.tw
[2] Department of Management Information Systems,
The University of Arizona, Tucson, AZ 85721, USA
futj@email.arizona.edu, hchen@eller.arizona.edu

Abstract. This paper examines the "Gray Web Forums" in Taiwan. We study their characteristics and develop an analysis framework for assisting investigations on forum communities. Based on the statistical data collected from online forums, we found that the relationship between a posting and its responses is highly correlated to the forum nature. In addition, hot threads extracted based on the proposed metric can be used to assist analysts in identifying illegal or inappropriate contents. Furthermore, members' roles and activities in a virtual community can be identified by member level analysis.

1 Introduction

Nowadays, computers and computer networks are not only used as tools for processing information, but also have become a new medium to share and access information online. For example, bulletin board systems, internet relay chat systems, and I-phone systems, are all integrated with the WWW and provide various communication channels for individuals to exchange information beyond the limits of time and space. Consequently, our society is in a state of transformation toward a "virtual society," where people's daily activities, such as shopping, getting services, and sharing information, can be accomplished without face-to-face contact with others.

Although the internet has enabled global businesses to flourish, it also allows criminals to make acquaintance of victims, acquiring them and eventually committing crimes. For example, just a few years ago, a National Taipei University student was found dead and wrapped in a luggage box dumped on a street corner by his "net friend," whom he met on a homosexual online forum. Today, many teenagers continue making friends through online activities, such as exchanging e-mails and playing internet video games, without having any physical interaction. The lack of physical interactions leaves few observable trails for parents and makes them less informed on their children's social activities. Just recently, two teenagers in Taiwan who got acquainted through an internet chat room committed suicide together. The breaking news astonished both the two families as well as the Taiwan society.

The advance of computer forensics has shown that online activities often create electronic trails [1]. For examples, bulletin board messages are stored in system

S. Mehrotra et al. (Eds.): ISI 2006, LNCS 3975, pp. 498 – 503, 2006.

archives, and e-mail messages are stored in mail servers or clients. Although archives in private storage can be acquired for analysis only under proper authorization, the information stored in public archives can be retrieved for analysis when necessary. After the 911 disaster, the FBI has shifted the reactive or post crime investigation paradigm to proactive investigation [2]. Thus, precrime investigation and data analysis are of critical importance to mitigate the negative effects of online activities. Although the collection and analysis of "dark web" sites have been under intensive investigations ([3], [4], [5], [6], [7]), only few researches addressed the issue of deriving crime leads or detecting symptoms of crimes from online archives. Wang, et al, conducted a study to gather illegal web sites using special search mechanisms [8]. Dringus, et al., used data mining to discover and build alternative representations for asynchronous education forums [9]. However, no research has been done on gathering crime leads from forum archives. This paper examines the "Gray Web Forums" in Taiwan and develops an analysis framework for assisting investigation on "gray" forums. The organization of this paper is as follows: Section 2 introduces the Gray Web Forum and its implication and challenges on crime investigation. Section 3 describes the framework for exploring Gray Web Forums. Section 4 gives conclusions and the future work.

2 Gray Web Forum-Based Communities

A forum is defined as *the computer-mediated medium for the open discussion of subjects of public interest* [10]. Forum members basically have three types of operations:

1. View existing postings.
2. Reply to an existing posting.
3. Start a new topic (also called a thread) of discussion.

"Postings" are messages that are sent to a forum for public viewing. Forum members may respond to existing postings or create a new topic. A group of postings related to the same topic are called a "thread." Detailed forum descriptions can be found in [10].

Communities are characterized by common interests, frequent interaction, and identification [11]. Internet forum members discuss on a certain topic, seek support from members, exchange information by postings, and are identified by e-mail addresses. Thus internet forums are perfect platforms for the formation of virtual communities. However, due to the anonymity and lack of observability, unscrupulous individuals may exploit internet forums for illegal activities. The *Gray Web Forum-based Virtual Community is* defined as: *the community formed through internet forums, which focused on topics that might potentially encourage biased, offensive, or disruptive behaviors and may disturb the society or threaten the public safety.* Such forums may cover topics such as pirated CDs, gambling, spiritualism, and so on. For examples, members of pirated CD forums may share music CDs without proper licensing; gambling forum members may provide hyperlinks to online gambling web sites; spiritualism forums may misguide teenagers and encourage disruptive behaviors.

Investigations on Gray Web Forums are difficult. Firstly, most forum postings are not indexed and thus can not be detected by search engines. Secondly, internet forums are highly distributed and their huge quantity prohibits effective coverage by manual investigations. Thirdly, a thread of discussion may cause numerous postings or sub-threads and the duration of each thread can be very long, which make manual investigations time consuming and highly inefficient. Finally, access control mechanisms adopted by forums may introduce obstacles for effective/efficient investigation.

3 A Framework for Gray Web Forum Analysis

3.1 Research Design

The framework for Gray Web forum analysis mainly consists of four components, as shown in Figure 1 below:

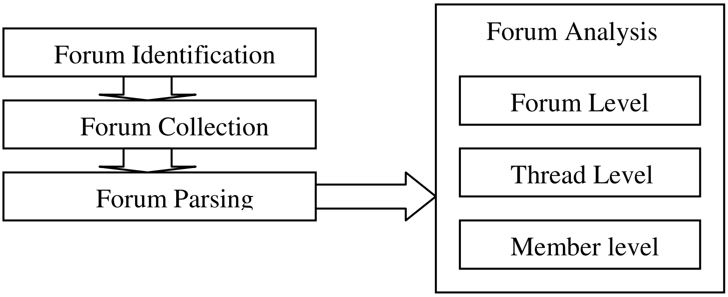

Fig. 1. The Gray Web Forum Analysis Framework

Forum identification is to identify Gray Web Forums based on the knowledge of domain experts. Forum collection is to spider the web pages in forums or boards that have been identified. Forum parsing is to extract useful content, such as the number of threads, the number of members, the duration and distribution of threads, and the frequency as well as the volume of postings. Forum analysis can be divided into three levels: forum, thread, and member levels. Forum level analysis is to provide a broad overview for the forums; thread level analysis is to identify *hot or influential threads*; and member level analysis is to segment members into different categories and identify their roles. Furthermore, forum members can be classified as *initiators*, *active members*, and *followers*. Initiators are those who create threads of discussion. Active members have both high frequency of participation and a large volume of postings. Finally, followers have high frequency of participation but a small volume of postings.

3.2 Forum Level Analysis

Several representatives of Taiwan Gray Web forums are shown in Table 1.

Table 1. Selected Gray Web Forums in Taiwan

FID*	URLs	Type
1	http://bbs.a35.info/thread.php?fid=486	Pirated CD
2	http://oie.idv.tw/cgi-bin/bbs/forums.cgi?forum=1	Gambling
3	http://www.525.idv.tw/bbs/cg-ibin/forums.cgi?forum=9	Sentimentalism
4	http://www.helzone.com/vbb/forumdisplay.php?f=38	Spiritualism

*Forum Identification Number.

Forum 1 provides its members with illegal software and pirated CDs. Forum 2 releases gambling information. Forum 3 allows its members to share and discuss sentimental essays. Forum 4 consists of threads discussing superstitious subjects. The results of forum level analysis are shown in Table 2.

Table 2. Overview of the Gray Web Forum Collection

FID	Type	Threads	Postings	Size(Mb)	Members
1	Pirated CDs	252	7749	32.2	**3013**
2	Gambling	515	**31128**	**292**	539
3	Sentimentalism	**1396**	4452	62.8	463
4	Spiritualism	434	2415	41.4	228

Both Forums 1 and 2 adopt similar access control mechanisms, which only allow members to view the contents of a posting after they have replied a message to the posting. The first postings of many threads in these two forums often contain URL links to pirated CDs or important gambling information, which cannot be viewed unless a viewer replies to them. Therefore, members have to post short/simple message, such as "I want to see it," in order to view the posting. Both Forums 1 and 2 have a high post-per-thread value. However, the average number of postings per member in Forum 1 is 2.6, which is much less than 56.5 in Forum 2. It is because gamblers often reply to as many threads as possible, while people who search for pirated CDs only reply to threads of interest. Forum 3 has 1396 threads but each thread only contains 3.2 postings on average. However, the average number of thread postings in Forum 2 is 60.4. Again, this is because gamblers tend to reply to more threads to gather gambling information.

3.3 Thread Level Analysis

Hot or *influential threads* are threads which statistically have longer duration, more members, more active members, and more postings in numbers and volumes. The following metric is used to calculate scores for every thread in each forum.

$$\textbf{Thread Score} = \mathbf{F_{norm}(N_p) \times F_{norm}(V_t) \times F_{norm}(D_t) \times F_{norm}(N_{am}) \times F_{norm}(N_m)}$$

Where $\mathbf{N_p}$ is the number of postings, $\mathbf{V_t}$ is the volume of postings, $\mathbf{D_t}$ is the duration, $\mathbf{N_{am}}$ is number of active members, and $\mathbf{N_m}$ is the number of members in a thread. The function $\mathbf{F_{norm}}()$ is used to normalize each variable to a range of [0,1]. Note that we classify members who have more than one posting as active members. The hottest threads in each forum and their statistics and topics are shown in Table 3.

Among these threads, Thread 4 has the potential to attract depressed teenagers, who are contemplating suicide. Some robbery cases have also been reported among people who were involved in the hottest thread of Forum 3. We believe such analysis can be used to assist analysts in identifying biased activities.

Table 3. The Hottest Threads from Each of the Four Forums

TID *	**Type**	**Postings**	**Volume (char)**	**Duration (day)**	**Members**	**Active Members**
1	Pirated CDs	469	7219	69	462	7
2	Gambling	211	4132	4**	186	25
3	Sentimentalism	91	9628	118	16	6
4	Spiritualism	88	5295	962	67	8
TID	**Topics**					
1	Pornography websites recommendation (reply needed to view the content)					
2	Lottery recommendation (made by an experienced member)					
3	A true story about "network love"					
4	What will you do if you know you will die tomorrow?					

* Thread Identification Number.

**4 days is in fact a long time for a typical thread in the gambling forum.

3.4 Member Level Analysis

Table 4 provides the percentage of members who have initiated new threads. The percentages of members who have never created a thread in Forum 1 (99.3%) and Forum 2 (88.5%) are higher than those of the other two forums. This is because that only a small portion of members in Forum1 and Forum 2 have the valuable information, such as pirated CDs or lottery analysis results. Therefore, most members in gambling forums and pirated CD forums are followers. However, members in sentimentalism and spiritualism forums tend to publish their own thoughts. Consequently the percentages of posting initiators are higher. The table also shows that most members in Forum 3 are willing to share their feelings online. Besides the fact that sentimentalism is a more common topic, internet forums are also becoming popular venues for people to share their feelings with strangers. Forum 4 is similar to Forums 1 and 2; however, the percentages of members who have created 1 thread (11.4%) and who have created 2-9 threads in Forum 4 (7.9%) are much higher.

Table 4. Percentage of Members Who Have Created New Threads

FID	Type	Number of threads members created						
		0	1	2-9	10-29	30-49	50-99	>=100
1	Pirated CDs	**99.3%**	0.3%	0.2%	0.1%	0.1%	0%	0%
2	Gambling	**88.5%**	4.5%	3.9%	2.6%	0.4%	0.2%	0%
3	Sentimentalism	**10.6%**	**40.6%**	**44.1%**	3.2%	0.9%	0.6%	0%
4	Spiritualism	77.6%	11.4%	7.9%	2.2%	0.4%	0%	0.4%

4 Conclusions and Future Work

This paper introduced the concept of Gray Web Forums. We developed a framework and analyzed four Gray Web Forums in Taiwan. Gambling forum members reply to as many threads as possible to gather gambling information; while people searching for pirated CDs or illegal software reply only to related threads of interest. Hot thread analysis can be used to assist in manual analysis to identify inappropriate contents. In addition, member level analysis can be used to identify members' roles in virtual communities. Although we were able to find some interesting results, we are far from pinpointing to specific threads or postings with offensive contents. Furthermore, barriers introduced by forum access control also need to be addressed in the future.

Acknowledgements

This research was supported in part by grants: NSC 94-2213-E-015-002, Taiwan, and the Fulbright Visiting Scholar Grant: 68429239, USA. The authors would also like to thank the Artificial Intelligence Laboratory, Department of Management Information Systems, University of Arizona, for providing space and facilities for the research visit.

References

1. Wang, J. H.: *Cyber Forensics – Issues and Approaches*, book chapter in book: Managing Cyber Threats: Issues, Approaches and Challenge, edited by Kumar, et al, Kluwer Academic Publishers, 2005.
2. Mena, J.: *Investigative Data Mining for Security and Criminal Detection*, Butterworth Heinemann, 2003.
3. Zhou, Y., Reid, E., Qin, J., Lai, G., Chen, H.: *U.S. Domestic Extremist Groups on the Web: Link and Content Analysis,* IEEE Intelligent Systems, Special Issue on Artificial Intelligence for National and Homeland Security, Vol. 20, Issue 5, 2005, pp. 44-51.
4. Chen, H.: *The Terrorism Knowledge Portal: Advanced Methodologies for Collecting and Analyzing Information from the Dark Web and Terrorism Research Resources*, presented at the Sandia National Laboratories, August 14 2003.
5. Elison, W.: *Netwar: Studying Rebels on the Internet*, The Social Studies 91, pp 127-131, 2000.
6. Tsfati, Y. and Weimann, G.: *www.terrorism.com: Terror on the Internet, Studies in Conflict & Terrorism 25*, pp 317-332, 2002.
7. Weimann, G.: *www.terrorism.net: How Modern Terrorism Uses the Internet*, Special Report 116, U.S. Institute of Peace, http://usip.org/pubs, 2004.
8. Wang, J. H., et al.: *A study of Automatic Search System on Cyber Crimes*, Research Report to Telecommunication Bureau, Ministry of Traffic and Communication, Taiwan, 1999.
9. Dringus, L. P. and Ellis, T.: *Using Data Mining as a Strategy for Assessing Asynchronous Discussion Forums*, Computer & Education 45, pp141-160, 2004.
10. Spaceman: http://www.vbulletin.c*om/forum/showthread.php?t=32329, 2001.*
11. *Bock, W.: Christmas, Communities, and Cyberspace,*
12. http://www.bockinfo.com/docs/community.htm , 2001