

On the Topology of the Dark Web of Terrorist Groups

Jennifer Xu¹, Hsinchun Chen², Yilu Zhou², and Jialun Qin²

¹ Computer Information Systems Department, Bentley College, 175 Forest Street,
Waltham, MA 02452, USA
jxu@bentley.edu

² Department of Management Information Systems, Eller College of Management,
The University of Arizona, Tucson, AZ 85721, USA
{hchen, yiluz, qin}@eller.arizona.edu

Abstract. In recent years, terrorist groups have used the WWW to spread their ideologies, disseminate propaganda, and recruit members. Studying the terrorist websites may help us understand the characteristics of these websites and predict terrorist activities. In this paper, we propose to apply network topological analysis methods on systematically collected the terrorist website data and to study the structural characteristics at the Web page level. We conducted a case study using the methods on three collections of Middle-Eastern, US domestic, and Latin-American terrorist websites. We found that these three networks have the small-world and scale-free characteristics. We also found that smaller size websites which share same interests tend to make stronger inter-website linkage relationships.

1 Introduction

Terrorism and terrorist activities substantially threaten national security. Although authorities have taken extensive counter-terrorism measures, terrorist groups remain active, using all kinds of media to disseminate propaganda, seek support, and recruit new members. The WWW, an effective information presentation and dissemination tool, has been widely used by terrorist groups as a communication medium. The Web presence of these terrorist groups reflects their different characteristics and may provide information about planned terrorist activities. Thus, monitoring and studying the content and structural characteristics of terrorist websites may help us analyze and even predict the activities of terrorist groups.

Recently, research on terrorism and terrorist groups on the Web has become an important topic in intelligence and security informatics. Researchers have employed content analysis and Web structure mining to reveal the characteristics of terrorist websites at site level. In this research, we analyze the structural characteristics of terrorist websites at a lower granularity—page level. Based on a systematically collected “Dark Web” data set, we conducted topological analysis to compare the hyper-link structures of terrorist websites from three geographical regions: Middle-East, the United States, and Latin-America.

The remainder of the paper is organized as follows. Section 2 reviews previous work on the structure of terrorist websites and the topological analysis methods. In

section 3, we present our data collection methods and the “Dark Web” dataset. In section 4, we report and discuss our findings from the analysis. Section 5 concludes the paper with implications and future research directions.

2 Literature Review

In this section, we review prior research on the structure of terrorist websites, and the network topology analysis methodology, which has been widely employed in other domains.

2.1 Web Mining Studies on Terrorist Websites

The World Wide Web has been increasingly used by terrorists to spread their ideologies [1]. According to the Southern Poverty Law Center (SPLC) [2], there were 708 active extremist and hate groups in the US in 2002. These groups had 443 websites in 2002 and this number increased to 497 in 2003.

Researchers and watchdog organizations such as SPLC have started monitoring, tracing and studying terrorist websites. The traditional approach is to study the contents and structure of these websites [3]. This approach is limited in the size of dataset and cannot be used to keep track of the dynamic characteristics of terrorist websites. Like other websites, terrorist websites usually suddenly emerge; the content and hyperlinks are frequently modified, and they may also swiftly disappear [4]. In recent years, Web mining techniques have been used in cyber crime and terrorism research [5].

Web mining combines data mining, text mining, and information retrieval techniques in the Web context to discover knowledge from the content of Web pages (content analysis), the structure of hyperlinks (structure analysis), and the visit and usage patterns of websites (usage analysis). In the terrorism domain, content analysis and structure analysis are the most frequently used techniques.

Studies on the content analysis of terrorist websites have shown that terrorist websites present different characteristics from other ordinary websites. For example, Gerstenfeld et al. found that many terrorist websites contain multimedia contents and racist symbols [6]. Gustavson and Sherkat found that terrorist groups used the Internet mainly for ideological resource sharing [7, 8]. This finding was also supported by a few other studies such as Zhou et al. [9], which analyzed the contents of terrorist websites in the United States and found that sharing ideology was one of the most important purposes for building these websites.

Structure analysis based on hyperlinks has also been seen in several previous studies. It is reported that website interconnection relations provide reasonably accurate representation of terrorist groups’ interorganizational structure [9, 10]. However, most of these studies focus on the hyperlink structures at the site level. There are few studies that analyze the hyperlink structure of websites among different terrorist groups at the page level, which may provide insight into the structure of terrorist groups. The topological characteristics of these websites at a lower granularity (page level) remain unknown.

2.2 Network Topological Analysis

Statistical analysis of network topology [11] is a recent development in network structure analysis research. In network topological analysis, entities, regardless of their contents, are treated as nodes and their relations are treated as links. The result is a graph representation of the network. Three topologies have been widely studied recently, namely, random network [12], small-world network [13], and scale-free network [14]. Different topologies have different structural characteristics and implications.

It has been found that most empirical networks such as social networks, biological networks, and the Web are nonrandom networks [11]. Actually, many of these networks are found to be small-world networks, which are characterized by small average path length and relatively large clustering coefficient compared to a random network. Moreover, many of these networks are also scale-free, meaning that a large percentage of nodes have just a few links, while a small percentage of nodes have a large number of links. Studies have shown that the WWW, in general, is both a small-world and a scale-free network [11].

A number of measures and approaches, many of which are borrowed from Social Network Analysis research [15], can be employed to reveal the structural patterns of a network. For example, nodes with high degrees may act as hubs or leaders, where the degree of a node is the number of links it has. The patterns found often have important implications to the functioning of the network.

The topological analysis has been used in previous studies on terrorist networks [16-18] and terrorist website structural studies at the site level [9]. In this research, we analyze the topological characteristics of “Dark Web” to reveal the structural properties of terrorist websites at the page level.

3 The Dark Web Dataset

We call the special section of the Web that is used by terrorists, extremist groups, and their supporters the “Dark Web.” As a long-term research project, we have kept collecting and tracing the content and hyperlinks of several terrorist groups’ websites and created a Dark Web test bed. [19]

3.1 Collection Building

In our research, we focused on terrorist groups from three geographical areas: the United States, Latin-America, and Middle-Eastern countries. By November 2004, we had collected three batches of Dark Web data by spidering these websites. To identify the correct terrorist groups to spider, we used the reports from authoritative sources suggested by a domain expert with 13 years of experience. The sources include: Anti-Defamation League, FBI, Southern Poverty Law Center, Militia Watchdog, United States Committee for a Free Lebanon, Counter-Terrorism Committee of the UN Security Council, and etc. From these resources, a total of 224 US domestic terrorist groups and 440 international terrorist groups were identified.

We then identified an initial set of terrorist group URLs and expanded them. In the initial set of URLs, all US domestic group URLs and some international group URLs were directly obtained from US State Department reports and FBI reports. Additional

international group URLs were identified through online searches. We constructed three terrorism keyword lists in terrorist groups' native languages, which contain terrorist organization name(s), leader(s)' and key members' names, slogans, special words used by terrorists, etc. From the search results, those websites that were explicitly purported to be official sites of terrorist organizations and that contained praise of or adopt ideologies espoused by a terrorist group were added to the initial URL set. The initial URL set is expanded by adding the URLs' in-link and out-link websites. Manual filtering was performed on the expanded links to ensure their quality.

After the URL of a terrorist group website was identified, we used a digital library building toolkit to collect the contents and hyperlinks from the sites.

We identified 108 US domestic terrorist websites, 68 Latin-American terrorist websites, and 135 Middle-Eastern terrorist websites.

3.2 Network Creation

After we collected the Web pages, the static HTML files and dynamic files were parsed and the hyperlinks in the files were extracted. We created a Dark Web page network, whose nodes were Web pages in the Dark Web collection and links were hyperlinks between these pages. For the US domestic collection, the network contained 97,391 nodes. The Latin-American collections contained 152,414 nodes. The Middle-Eastern collection contained 298,761 nodes.

4 Dark Web Page Level Topological Analysis

4.1 Network Topological Analysis

In this research, we ignored the directions of hyperlinks, and studied topological characteristics of the resulting non-directional networks. Table 1 shows the basic statistics of the three networks. We defined clustering coefficient as [20]:

$$C = \frac{3 \times \text{number of triangles on the graph}}{\text{number of connected triples of nodes}} \quad (1)$$

Because these networks were rather large, we employed an approximation algorithm called ANF to calculate the average path length [21]. To estimate the diameter of the three networks, we randomly selected 400 nodes in each network and calculated the shortest path length between them.

Comparing the topological characteristics of the three networks we found that the Middle-Eastern network is much larger than the US domestic network and the Latin American Network. The number of nodes in the Middle-Eastern network (298,761) is almost three times of that in the US domestic network (97,391) and two times of that in the Latin-American network (152,414). Among the three networks, the Middle-Eastern network has the highest average degree (12.66), indicating that their Web pages tend to point to each other more often than those in the US domestic network and in the Latin-American network. The size of the Middle-Eastern network and the high average degree may indicate their relatively active status and strong intention to cooperate with each other.

Table 1. Basic statistics of the three collections of Dark Web page level networks

<i>Collections</i>	<i>US domestic</i>	<i>Latin-American</i>	<i>Middle-Eastern</i>
<i>No. of Nodes</i>	97,391	152,414	298,761
<i>No. of Links</i>	296,330	586,115	1,914,099
<i>No. of N-Links</i>	239,572	475,748	1,890,728
$\langle k \rangle$	4.92	6.24	12.66
<i>C</i>	0.32	0.31	0.06
C_{rand}	5.05E-05	4.1E-05	4.24E-05
l (by ANF)	3.33	4.70	3.52
l_{rand}	7.21	6.52	4.97
<i>D</i> (by simulation)	≥ 39	≥ 46	≥ 38
<i>NC</i>	4,134	1,110	674
$Node_C$	81,803	22,175	255,699
$Link_C$	239,982	95,346	1,718,626

N-Links: Non-directional links; $\langle k \rangle$: average degree; l : average path length; l_{rand} : average path length of a random network; *C*: clustering coefficient; C_{rand} : clustering coefficient of a random network; *D*: network diameter; *NC*: number of components; $Node_C$: number of nodes in the largest component; $Link_C$: number of links in the largest component

All the three networks have rather high clustering coefficients and small average path length comparing with their random network counterparts. The high clustering coefficient indicates that the networks contain dense and strongly connected local clusters. In this case, it is obvious that the Web pages are more likely to point to pages within the same site, resulting in local clusters. Note that the clustering coefficient of the Middle-Eastern network is smaller than those of the other two networks. This may be caused by two reasons. First, the US domestic and Latin-American networks have a much larger number of components than the Middle-Eastern network. These components serve as local clusters, causing the overall clustering coefficients of the whole networks to be higher. Second, it may be because that the pages in the giant component of the Middle-Eastern network are more decentralized.

Figure 1 shows the in-degree distribution and out-degree distribution of the three networks in log-log plots. All the six degree distributions have a long tail which is often observed in large empirical networks. The in-degree distributions of the Middle-Eastern network and the US domestic network follow a power law degree distribution. The out-degree distribution of the three networks and the in-degree distribution of the Latin American network show two power law parts with a tail.

The power-law distribution takes the form of $P(k) \sim k^{-\gamma}$, where $P(k)$ is the probability that a node has exactly k links. The values of the exponents of the six distributions are shown in Table 2. The special shape of the out-degree distributions of the three networks may be because a Web page normally does not contain so many hyperlinks to the other pages. Thus, the likelihood of such high out-degree pages will quickly drop as degree increases. For the in-degree distribution, as the Latin-American network contains several small components and very few large components, it is difficult for the high in-degree nodes to emerge. As the in-degree increases, the number of nodes with high in-degrees decreases quickly.

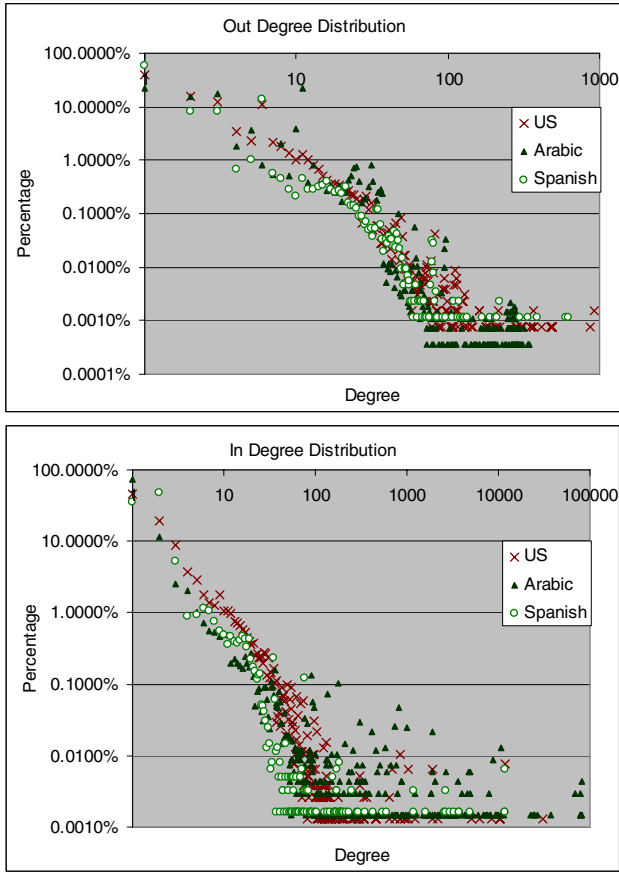


Fig. 1. Degree distributions of the three networks

Table 2. Exponents of the three networks' degree distributions

<i>Collections</i>	<i>US domestic</i>	<i>Latin-American</i>	<i>Middle-Eastern</i>
<i>In degree exponent</i>	1.94	2.16/2.53	1.60
<i>Out degree exponent</i>	1.95/2.30	2.24/2.26	1.88/ 2.44

4.2 Giant Component Analysis

Although the Middle-Eastern network is the largest network, it has fewer components than the other two networks. Table 3 shows the three networks' top five components' node percentage in their networks. The three networks all have a giant component (The Latin-American network's largest component has 22.77% of the nodes. But it is still very big compared with the other components in the network.). We also observed that these giant components usually are composed of several terrorist websites.

Table 3. The node percentages of the top 5 components in the three networks

<i>Component Size Rank</i>	<i>US domestic</i>	<i>Latin-American</i>	<i>Middle-Eastern</i>
1	53.67% pages 54 websites	22.77% pages 9 websites	85.62% pages 68 websites
2	2.31% pages 1 website	6.58% pages 1 website	2.73% pages 1 website
3	0.68% pages 1 website	5.84% pages 10 websites	1.66% pages 1 website
4	0.56% pages 1 website	4.61% pages 11 websites	1.35% pages 2 websites
5	0.43% pages 1 website	2.79% pages 1 website	1.13% pages 10 websites
<i>Other Components</i>	42.35% pages	57.41% pages	7.51% pages

The giant component of the Latin-American network contains fewer websites than the giant components in the other two networks. This may be because that these Latin-American terrorist groups have diverse ideologies and beliefs. As a result, it is less likely for them to refer to each other on their websites or to seek cooperation.

The three giant components compose the bulk of the three networks. We thus focused only on the giant components of the three networks.

In general, there is a positive correlation between a website’s size (number of pages) and number of internal links. This is also observed in the websites and pages included in the giant components of the three networks (Figure 2). However, these terrorist networks show special characteristics on inter-site links. Figure 3 presents the relationship between the website sizes and the number of inter-site links of the three networks. In this figure, the vertical axis represents the number of hyperlinks between a pair of websites in the giant component and the other two axes represent the number

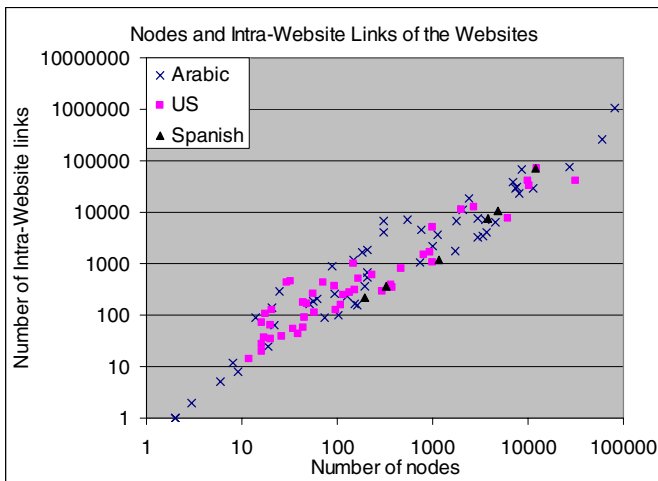


Fig. 2. Website size and the number of internal links

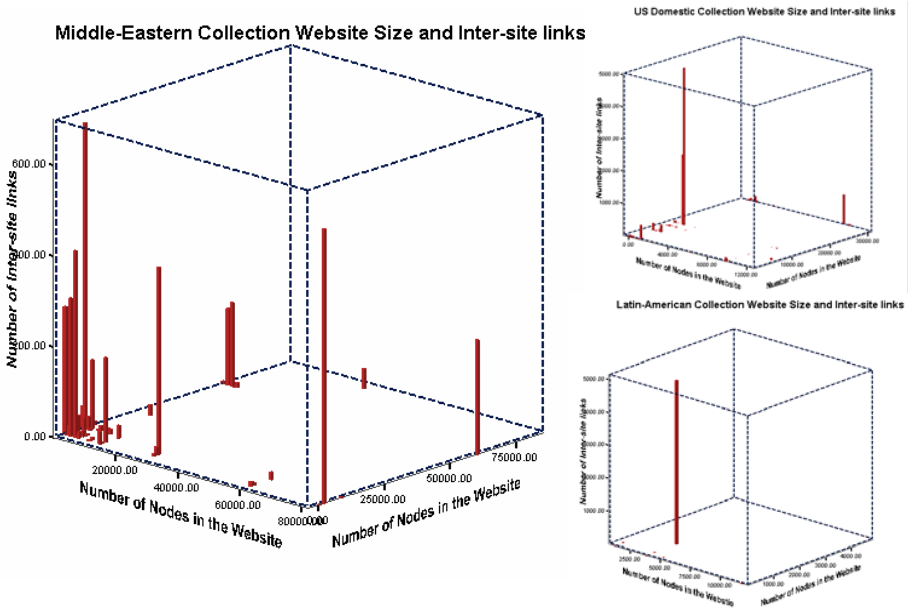


Fig. 3. Website size and the number of inter-site links

of pages in the two websites. For all three networks, we observe that most of the inter-site links are not present between large websites. For example, in the Middle-Eastern network, most of the inter-site links appear between websites that have less than 10,000 pages. It is normal for large websites to share a large number of inter-links. However, if two websites with relatively small number of pages are connected by many inter-links, it means that the two websites must have a close relationship.

To further study the relationships between the small and middle sized websites, which usually are connected by many inter-site links, we selected and examined some of these websites. For example, in the US domestic network, there are 4,875 inter-site links between www.resistance.com (12,188 pages) and www.natall.com (943 pages), 2,173 links between www.resistance.com and www.natvan.com (814 pages), and 414 links between www.natvan.com and www.natall.com. After we examined their websites, we found that the three websites have very close relationship. www.natall.com is the official Website of the National Alliance, a white supremacist group. www.natvan.com is another domain name of www.natall.com. www.resistance.com is an e-commerce website owned by Resistance Records, which is a music production company affiliated with National Alliance. Therefore, the dense inter-site hyperlinks may reflect the close relationship between the organizations.

In the Latin-American Eastern network, clajadep.lahaine.org (3,796 pages) and www.carteleralibertaria.org (1,177 pages) are connected by 4,979 links. The Clajadep group (clajadep.lahaine.org) is focused more on broadcasting affairs in Mexico, while the Cartelera Libertaria group (www.carteleralibertaria.org) more on Spain. These two groups all belong to a terrorist alliance called “La Haine,” which has people from different Spanish-speaking and Latin America countries. The dense inter-site links may result from the fact that members in La Haine share similar ideologies, beliefs and interests.

Similarly, in the Middle-Eastern network, there are some small websites of close relationship. For example, www.daymohk.info (737 pages) and www.chechen.org (7,042 pages) have 676 links, which are both websites for extremists in Chechnya. www.palestine-info.info (3,698 nodes) and www.infopalestina.com (550 nodes) shared 410 interlinks, which are both news websites for Palestinians.

These cases show that the similarities in terrorist groups' ideologies, beliefs, interests, and geographical closeness may cause their websites to frequently point to each other. From Figure 3, we can see that the Middle-Eastern giant component has many more and denser inter-site links than the other two giant components. It implies that the terrorist groups in Middle-Eastern have relatively closer relationships and more interconnections than those in the United States and Latin-American. Such dense inter-site links also enable the emergence of the giant component in the network.

5 Conclusions and Future Directions

In this research, we analyzed the structural properties of the Dark Web at the page level based on systematically connected terrorist websites data. Our goal was to reveal the characteristics of these websites. We conducted a case study based on a "Dark Web" test bed of US domestic, Latin-American, and Middle-Eastern terrorist websites. From the case study we found that:

- The three networks are small worlds.
- The three networks' in-degree and out-degree distributions roughly follow a power-law degree distribution, indicating that they have the scale-free characteristics. In addition as degree increases in the three networks, the probability of having nodes with high degrees decreases more quickly than in scale free networks.
- The giant components of the three networks contain several websites, which are not very large and share the same interests. They also have more inter-site links and closer relationships.

A limitation of our study is that focused only on the structural properties of the Dark Web without performing content analysis that might reveal important insights into the ideology, mission, and other information about these terrorists groups. Caution must be made when any interpretation is drawn based solely on the structure of the Dark Web. In the future, we plan to perform in-depth content analysis on these terrorist web sites and combine it with other structural analysis methods such as cluster analysis from the network structural perspective to advance our knowledge of the Dark Web.

References

1. E. Lee and L. Leets, "Persuasive storytelling by hate groups online - Examining its effects on adolescents," *American Behavioral Scientist*, vol. 45, pp. 927-957, 2002.
2. S. P. L. Center, "Hate Groups, Militias on Rise as Extremists Stage Comeback," 2004, pp. www.splcenter.org/center/splcreport/article.jsp?aid=71.

3. M. Whine, "Far Right on the Internet," in *Governance of Cyberspace*, B. Loader, Ed.: Routledge, 1997, pp. 209-227.
4. G. Weimann, "How Modern Terrorism Uses the Internet," United States Institute of Peace, www.terror.net Special Report 116, 2004.
5. H. Chen, W. Chung, J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining: A general framework and some examples," *Computer*, vol. 37, pp. 50-+, 2004.
6. P. B. Gerstenfeld, D. R. Grant, and C.-P. Chiang, "Hate Online: A Content Analysis of Extremist Internet Sites," *Analyses of Social Issues and Public Policy*, vol. 3, pp. 29, 2003.
7. A. T. Gustavson and D. E. Sherkat, "Elucidating the Web of Hate: The Ideological Structuring of Network Ties among White Supremacist Groups on the Internet," presented at presented at Ann. Meeting Am. Sociological Assoc., 2004.
8. C. C. Demchak, C. Friis, and T. M. L. Porte, "Webbing Governance: National Differences in Constructing the Face of Public Organizations," in *Handbook of Public Information Systems*, G. D. Garson, Ed.: Marcel Dekker, 2000.
9. Y. Zhou, J. Qin, G. Lai, E. Reid, and H. Chen, "Building Knowledge Management System for Researching Terrorist Groups on the Web," presented at Proceedings of the Eleventh Americas Conference on Information Systems, Omaha, NE, USA, 2005.
10. V. Burris, E. Smith, and A. Strahm, "White Supremacist Networks on the Internet," *Sociological Focus*, vol. 33, pp. 215-235, 2000.
11. R. Albert and A. L. Barabasi, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, pp. 47-97, 2002.
12. P. Erdos and A. Renyi, "On random graphs," *Publ Math-Debrecen*, vol. 6, pp. 290-297, 1959.
13. D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, pp. 440-2, 1998.
14. A. L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509-12, 1999.
15. S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*. Cambridge: Cambridge University Press, 1994.
16. D. McAndrew, "The structural analysis of criminal networks," in *The Social Psychology of Crime: Groups, Teams, and Networks*, I, Offender Profiling Series, II, D. Canter, Alison, L., Ed. Aldershot, Dartmouth, 1999, pp. 53-94.
17. J. S. McIllwain, "Organized crime: A social network approach," *Crime, Law & Social Change*, vol. 32, pp. 301-323, 1999.
18. M. K. Sparrow, "The application of network analysis to criminal intelligence: An assessment of the prospects," *Social Networks*, vol. 13, pp. 251-274, 1991.
19. J. Qin, Y. Zhou, G. Lai, E. Reid, M. Sageman, and H. Chen, "The Dark Web portal project: Collecting and analyzing the presence of terrorist groups on the web," *Intelligence and Security Informatics, Proceedings*, vol. 3495, pp. 623-624, 2005.
20. M. E. Newman, D. J. Watts, and S. H. Strogatz, "Random graph models of social networks," *Proc Natl Acad Sci U S A*, vol. 99 Suppl 1, pp. 2566-72, 2002.
21. C. Palmer, P. Gibbons, and C. Faloutsos, "ANF: A fast and scalable tool for data mining in massive graphs," presented at In Proc. of the 8th ACM SIGKDD Internal Conference on Knowledge Discovery and Data Mining, 2002.