PROFILING AND TRACKING A CYBERLOCKER LINK SHARER IN A PUBLIC WEB FORUM

Xiao-Xi Fan, Kam-Pui Chow, Fei Xu

Chapter 6

PROFILING AND TRACKING A CYBERLOCKER LINK SHARER IN A PUBLIC WEB FORUM

Xiao-Xi Fan, Kam-Pui Chow and Fei Xu

Abstract The expanding utilization of cyberlocker services is driven by the illegal exchange of copyrighted materials. In fact, the illegal exchange of copyrighted materials is the largest contributor to global Internet traffic. However, due to the anonymity provided by cyberlockers, it is difficult to track user identities directly from cyberlocker sites. Since cyberlocker users upload and share links via third-party sites, it is possible to harvest cyberlocker-related data from these sites and connect the data to specific users. This chapter describes a framework for collecting cyberlocker data from web forums and using cyberlocker link sharing behavior to identify users. Multidimensional scaling analysis and agglomerative hierarchical clustering analysis are performed on user profiles to yield clusters of forum users with similar sharing characteristics. The experimental results demonstrate that the framework provides valuable insights in investigations of cyberlocker-based piracy.

Keywords: Cyberlockers, piracy, user profiling, identity tracking

1. Introduction

Cyberlocker services such as RapidShare, MediaFire and BitShare have transformed how Internet users disseminate and share information and media [12]. Cyberlockers provide an easy way for users to upload files to servers. After a file has been uploaded successfully, the cyberlocker generates a unique download URL (cyberlocker link) for the uploader to copy and share with other users. The cyberlocker link can be posted and distributed via third-party sites such as web forums and blogs, enabling other users to easily download the file from the cyberlocker by clicking on the link [11].

Cyberlockers facilitate the sharing of large files, including movies and music. However, they also enable users to exchange illegal or copyrighted media on the Internet – full three-quarters of RapidShare content is estimated to be illegal [6].

Some cyberlockers encourage users to upload and share files by paying them money – the greater the number of downloads of a file, the more the file uploader is paid. This encourages file uploaders to disseminate their cyberlocker links on web forums and other venues to attract downloaders. Although many cyberlockers disaggregate search functionality, copyright owners argue that this has little effect on illegal file sharing [5]. Individuals can search for the desired content along with a specific cyberlocker name using a search engine to obtain links for the desired content as well as third-party sites that aggregate cyberlocker links.

Internet piracy is the distribution and/or copying of copyrighted materials over the Internet. Since Internet piracy is rampant in cyberlockers, an effective model is required to track the identities of cyberlocker users in digital forensic investigations. Many cyberlockers allow users without connections to file uploaders to download files independently [17]. This means that, when downloading a file from a cyberlocker link, information about the uploader is not available on the web page. Unlike a P2P network, a cyberlocker can hide user identities because the IP addresses of users are kept anonymous from each another and are known only to the cyberlocker operator. Without the cooperation of cyberlocker operators, law enforcement agencies are unable to identify users involved in illegal sharing or to download transactions involving copyrighted media [13].

The monitoring and investigation of cyberlocker users are arduous and time consuming, and it is exceedingly difficult to track user identities directly from cyberlocker sites. However, information about cyberlocker link sharing is easily harvested from third-party sites such as web forums; user information can also be gleaned from many web forums. This makes it possible to analyze the behavior of cyberlocker link sharers and provide investigative leads for identity tracking. Moreover, since cyberlocker users tend to have multiple accounts on one or more forums in order to distribute cyberlocker links, a model that could identify accounts that belong to a single user would be very useful in piracy investigations. The research described in this chapter attempts to address these issues. It focuses on building an effective framework for profiling web forum users based on their cyberlocker link sharing characteristics and identifying the relationships between different profiles based on the assumption that the same user has consistent preferences, thereby providing useful identity leads to investigators.

2. Related Work

Despite their popularity, little research has focused specifically on cyberlockers and cyberlocker-based piracy. Zhao et al. [20] have analyzed the numbers and sizes of files of different formats and contents on three cyberlockers. Mahanti [12] has proposed a measurement infrastructure to gather cyberlocker workloads and has compared the characteristics of content popularity, content dissemination and performance related to five cyberlockers. Envisional [6] has estimated the percentage of copyrighted materials exchanged via cyberlockers. Its report revealed that 7% of Internet traffic was related to cyberlocker services and that 73.2% of non-pornographic cyberlocker traffic (5.1% of all Internet traffic) involved copyrighted content being downloaded illegally. Most of the work focused on cyberlockers has involved the analysis of traffic and file characteristics, not user characteristics.

Since Internet piracy can be considered to be a type of serial crime, it is appealing and productive to apply criminal profiling to the problem. However, in the field of criminology, profiling has mainly focused on violent criminal activity and analyses of crime scenes and victims, which are not directly applicable to Internet piracy. Research on criminal profiling in the area of cyber crime is limited. Bhukya and Banothu [3] have captured GUI-based user behavior to construct user profiles and have applied support vector machines to classify users. McKinney and Reeves [14] have proposed a model for developing user profiles by analyzing processes running on a computer and detecting masquerades.

In the field of information retrieval, profiling techniques have been used in the context of recommender systems and e-commerce systems. Schiaffino and Amandi [18] have studied the information content of user profiles, methods for creating user profiles and techniques for analyzing user profiles. Godoy and Amandi [8] have designed a document clustering algorithm to categorize web documents and learn user interest profiles. Fathy et al. [7] have proposed a personalized search approach based on click history data to model user search preferences in ontological user profiles. All these approaches are related to digital profiling and, as such, can be used to develop criminal profiling methodologies for Internet piracy. However, there is little, if any, work on constructing user profiles based on cyberlocker link sharing behavior in web forums.

3. Methodology

Cyberlocker links are typically distributed in posts on web forums. A sharer is a forum user who uploads and shares a cyberlocker link in the forum for others to download a file; the sharer is also the potential

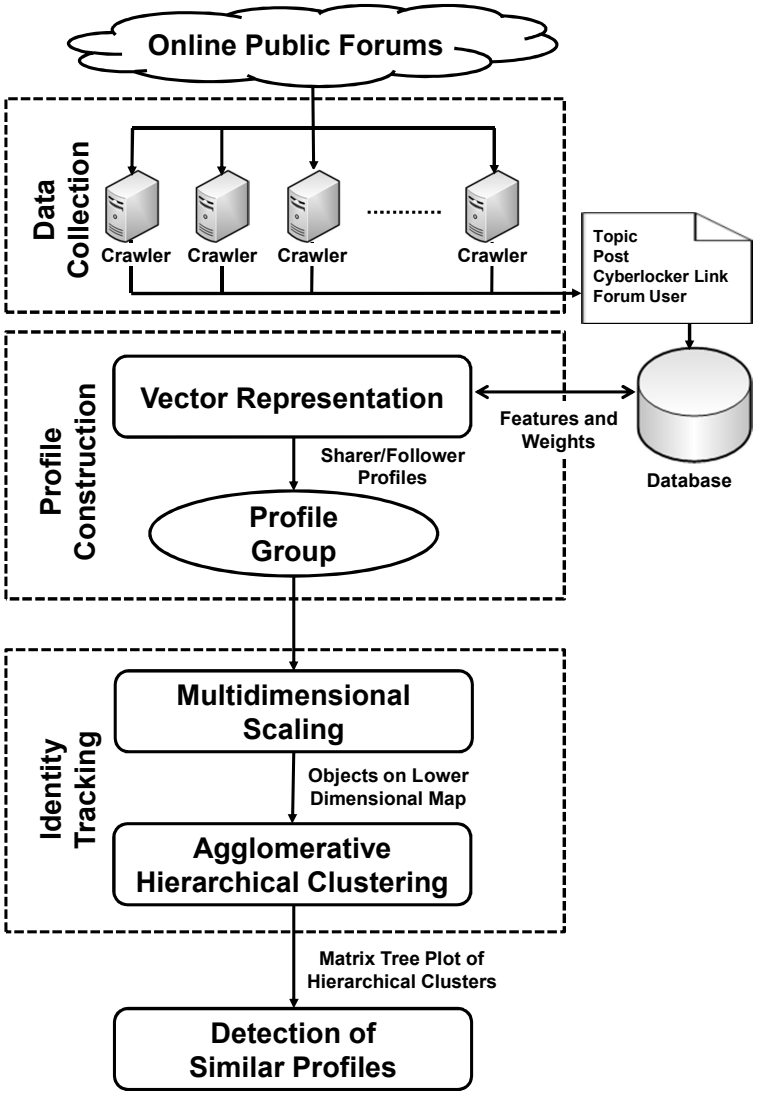

Figure 1. Digital profiling and identity tracking model.

uploader of the file on the cyberlocker site. In contrast, a follower is a forum user who only replies to a post containing a cyberlocker link without uploading or sharing any cyberlocker links.

Figure 1 presents the model used to build profiles for the two types of forum users and analyze the profiles. Web forum and cyberlocker data are collected by web forum crawlers that target several popular Hong Kong web forums. The profiles of sharers and followers are then created

in a vector space model based on the extracted features. Multidimensional scaling (MDS) is applied to analyze the similarity between sharers/followers and to map their relationships to a lower-dimensional representation. Finally, agglomerative hierarchical clustering is performed on the multidimensional scaling results to identify clusters containing sharers/followers with similar behavior.

3.1 Data Collection

Since a vast number of cyberlocker links are posted on web forums, the system described in this chapter is designed to collect data from web forums. A forum is hierarchical in structure and may contain several sub-forums. Topics come under the lowest level of sub-forums. The following definitions of topic and post [2] are used in this research:

- **Topic:** A topic may contain any number of posts from multiple forum users. The information includes the topic title, original post (i.e., first post that started the topic) and creation date and time (which is also the creation date and time of the original post).
- **Post:** A post is a message submitted by a forum user. A post contains information such as the forum user details, content and creation date and time. The original post is the one that started the topic and the posts that follow continue the discussion about the topic.

The system uses web crawlers to collect web forum and cyberlocker data. As shown in Figure 2, two types of crawlers are used: (i) URL crawlers; and (ii) topic crawlers. URL crawlers are responsible for extracting the URLs of topics from the catalog pages in web forums. After a URL crawler uses a fetch command to obtain the URL of a catalog page, it invokes a DNS module to resolve the host address of the web server. Following this, the URL crawler connects to the server to download the page, uses a parse command to extract all the URLs of topics from the page and stores them in the URL frontier.

Topic crawlers are responsible for extracting useful information from topic pages. After a topic crawler uses a fetch to obtain the URL of a topic from the URL frontier, it executes a parse command on the topic page to extract topic information and post information. The system thus parses and extracts cyberlocker links from all the posts encountered by the web crawlers. Several types of information are parsed into structured data and stored in a database.

Compared with the crawlers implemented by cyberlocker indexing sites, the web forum crawlers used in this work specifically target Hong

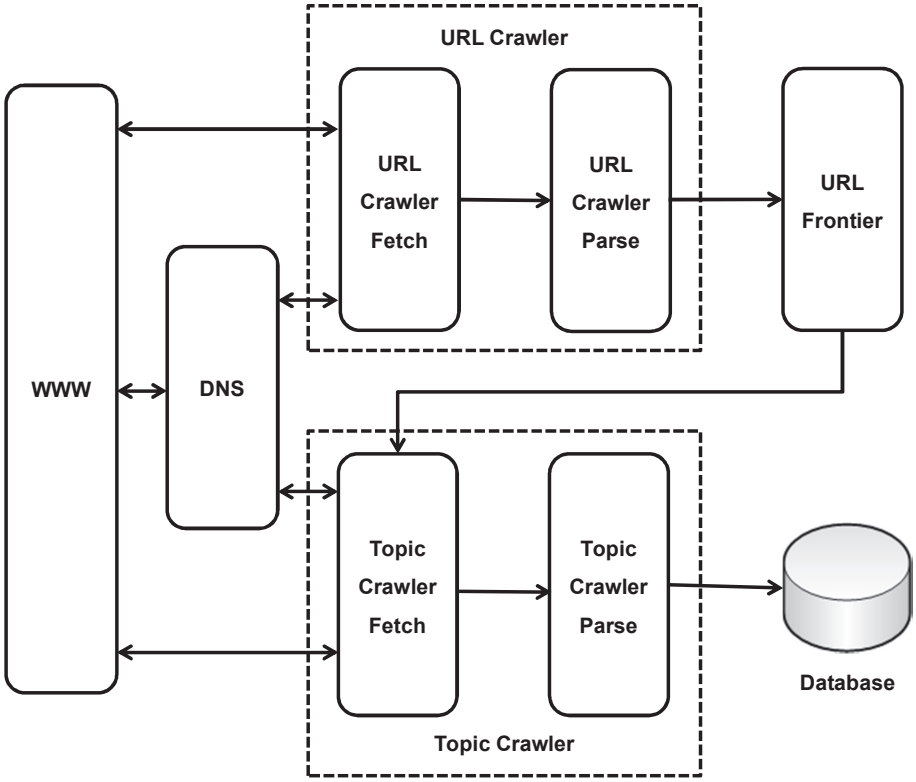

Figure 2. Web forum crawler.

Kong web forums and assign a higher priority to newly created posts to facilitate real-time analysis. The crawlers also download topic content to preserve evidence of cyberlocker-based piracy.

During the data collection process, web crawlers extracted data from the nine web forums presented in Table 1. According to Alexa Internet [1], the nine forums are among the most frequently visited websites in Hong Kong.

Table 2 lists the cyberlocker services considered in this research. They are among the top-ranked cyberlocker services used by Hong Kong web forum users.

3.2 User Profile Construction

This section describes the construction of profiles for sharers and followers, and the detection of relationships between sharers and followers. Sharer and follower profiles are created using data collected during an observation period. A vector space model commonly used in infor-

Table 1. Web forums considered in the study.

Forum	**Forum URL**
HK Discuss Forum (DISCUSS)	`http://www.discuss.com.hk`
UWANTS	`http://www.uwants.com`
HK-PUB	`http://hk-pub.com`
KYO	`http://www.kyohk.net`
EYNY	`http://www.eyny.com`
TVBOXNOW	`http://www.tvboxnow.com`
CK101	`http://ck101.com`
LALULALU	`http://www.lalulalu.com`
8CYBER	`http://8cyber.net`

Table 2. Cyberlocker services considered in the study.

Cyberlocker Service	**Cyberlocker URL**
Bitshare	`http://bitshare.com`
FilePost	`http://filepost.com`
FileFactory	`http://www.filefactory.com`
FileIM	`http://www.fileim.com`
Sendspace	`http://www.sendspace.com`
MediaFire	`http://www.mediafire.com`
Ziddu	`http://www.ziddu.com`
RapidShare	`https://rapidshare.com`
DepositFiles	`http://depositfiles.com`
HulkShare	`http://www.hulkshare.com`

mation retrieval is employed to specify the profiles. In particular, a sharer/follower is represented as a vector $s_j = (w_1, w_2, ..., w_N)$ where N is the number of features. Each feature corresponds to a cyberlocker link distributed by a sharer during the observation period.

The weighting scheme depends on whether or not a follower profile is included in a profile group for identity tracking analysis. If a follower profile is not included in a profile group, then the following equation is used to assign weights to each feature to describe the behavior of a sharer:

$$w_i = \begin{cases} 1 \text{ if sharer has shared link } i \\ 0 \text{ if sharer has not shared link } i \end{cases} \quad (1)$$

If the sharer and follower profiles are considered, the following equation based on an ordinal measurement is used to assign weights to each feature to describe the behavior of a sharer/follower:

$$w_i = \begin{cases} 2 \text{ if sharer has shared link } i \\ 1 \text{ if sharer/follower has replied to but not shared link } i \\ 0 \text{ if sharer/follower has not replied to or shared link } i \end{cases} \quad (2)$$

After weights are assigned to each feature for all sharers/followers, each sharer/follower is represented as a vector with components corresponding to the cyberlocker links.

3.3 Multidimensional Scaling Analysis

Multidimensional scaling (MDS) is a data analysis technique for identifying the underlying pattern or structure of a set of objects. Multidimensional scaling has been widely used in criminal profiling to analyze crime behavior patterns [9, 16]. It expresses the similarities and dissimilarities of objects in a geometric representation using a number of distance models.

The primary outcome of multidimensional scaling analysis is a spatial configuration in which the objects are represented as points. The points corresponding to similar objects are located close together, while those corresponding to dissimilar objects are located far apart. If sharers/followers are considered to be objects, then multidimensional scaling can be used to arrange the objects in a lower-dimensional map where the distance between two objects represents the observed correlation of the corresponding sharers/followers. If an object A is in close proximity to an object B but far away from an object C, then object A and object B have a strong relationship while a weak or no relationship exists with the remote object C. A forensic investigator can then attempt to make sense of the derived object configuration by identifying meaningful directions of similar sharers/followers in the space.

Interested readers are referred to Borg and Groenen [4] for additional details about multidimensional scaling. This work employed the SPSS statistical tool to perform multidimensional scaling analysis on the profiles constructed for sharers/followers.

3.4 Cluster Analysis

After multidimensional scaling analysis is performed, the profiles are represented as points in a lower-dimensional map without losing the relationships between each other. Agglomerative hierarchical clustering is then performed on the coordinates of the resulting points to identify potential clusters in which sharers/followers have similar behavior.

Agglomerative hierarchical clustering builds a hierarchy based on each individual object in a cluster. Next, it merges the closest pair of clusters

Table 3. Sharer/follower statistics for four long-term observation periods.

Observation Period	2014.01.01 to 2014.06.30		2013.07.01 to 2013.12.31		2013.01.01 to 2013.06.30		2012.07.01 to 2012.12.31	
Subjects	**Sharers/ Followers**	**Sharers**	**Sharers/ Followers**	**Sharers**	**Sharers/ Followers**	**Sharers**	**Sharers/ Followers**	**Sharers**
Number	1,649	64	814	72	684	91	5,580	215
DISCUSS	0.18%	4.69%	0.98%	11.11%	0.73%	3.30%	0.11%	2.79%
UWANTS	1.09%	10.94%	3.32%	8.33%	2.34%	7.69%	2.06%	8.37%
HK-PUB	1.46%	12.50%	5.77%	9.72%	1.02%	7.69%	0.72%	3.72%
KYO	0.12%	1.56%	2.21%	8.33%	1.90%	7.69%	1.81%	8.37%
EYMY	1.21%	1.56%	2.95%	2.78%	0.73%	1.10%	0.22%	1.86%
TVBOXNOW	56.58%	29.69%	26.78%	22.22%	67.54%	35.16%	66.83%	34.88%
CK101	7.22%	25.00%	6.76%	16.67%	10.23%	15.38%	12.42%	26.05%
LALULALU	31.53%	12.50%	50.98%	18.06%	14.91%	19.78%	15.82%	13.49%
8CYBER	0.61%	1.56%	0.25%	2.78%	0.58%	2.20%	0.02%	0.47%

that satisfy a similarity criterion in each successive iteration until all the objects are in one cluster. Agglomerative hierarchical clustering does not need the number of clusters as input and investigators can choose appropriate clusters according to their needs. This work adopted Ward's minimum variance method [15] as the criterion for choosing the pair of clusters to merge at each step. Compared with other criteria, Ward's minimum variance method is more aligned with the clustering requirements for the cyberlocker problem because it tends to generate a larger distance between two large clusters so that they are less likely to be merged, and vice versa.

Agglomerative hierarchical clustering can produce an ordering of objects that is very informative as a hierarchical display. This work uses a matrix tree plot to show the clustering outcome in a visually appealing manner that highlights clustering relationships.

4. Experiments

Experiments were conducted over a long-term observation period (six months) and a short-term observation period (one month). In each experimental design, experiments were repeated to detect general data trends.

4.1 Datasets

Datasets were extracted from the nine web forums listed previously. Since web forum users have to register and log in to post messages, their user details and sharing information were gathered quite easily.

Data was first collected over four long-term observation periods. Table 3 presents the basic statistics for sharers/followers over the four long-term observation periods. For all the sharers/followers collected over an

Table 4. Sharer/follower statistics for six short-term observation periods.

Observation Period	2014.01.01 to 2014.01.31	2014.02.01 to 2014.02.28	2014.03.01 to 2014.03.31	2014.04.01 to 2014.04.30	2014.05.01 to 2014.05.31	2014.06.01 to 2014.06.30
No. Sharers and Followers	117 117	209 209	511 511	320 320	431 431	295 295
No. Sharers and Followers rel. > 3 links	41	104	141	152	167	196
No. Sharers	26	26	34	21	28	26

observation period, the proportion of sharers is very small, which implies that thousands of cyberlocker links are shared by only a small number of web forum users. In addition, three web forums, TVBOXNOW, LALU-LALU and CK101, have the largest numbers of sharers and followers, which means that these three forums are the most popular venues for distributing cyberlocker links.

Table 4 presents the basic statistics for sharers/followers over six short-term observation periods. The data has the same trends as for the long-term observation periods: cyberlocker links are distributed by a few sharers, some of whom are active over all six short-term observation periods. Moreover, since a short-term observation period extended over one month, the number of followers decreased significantly when sharers and followers who shared or replied to more than three cyberlocker links were counted.

4.2 Evaluation Metric

The mapping process of multidimensional scaling is a critical step that influences the degree to which multidimensional scaling represents the correlation of the data. As a result, the stress measure, which is the most commonly-used measure to evaluate multidimensional scaling results [19], was selected to assess the goodness of fit.

If the input to multidimensional scaling is not a proximity matrix, then proximity values are created. A monotonic transformation of the proximity values is then performed to yield scaled proximities. The objective of multidimensional scaling is to find a configuration of points that minimizes the squared differences between the optimally-scaled proximity values and the distances between points. In other words, multidimensional scaling seeks to minimize the stress metric:

Table 5. MDS stress values and goodness of fit.

Stress Value	Goodness of Fit
Stress = 0	Perfect
0 < Stress ≤ 2.5%	Excellent
2.5% < Stress ≤ 5%	Good
5% < Stress ≤ 10%	Fair
10% < Stress ≤ 20%	Poor

$$Stress = \sqrt{\frac{\sum (f(p) - d)^2}{\sum d^2}} \tag{3}$$

where $f(p)$ is a monotonic transformation of the proximity p and d is the distance between points. A small stress value implies a good fit of the multidimensional scaling results, and vice versa.

Table 5 presents the guidelines proposed by Kruskal [10] to interpret the goodness of fit of multidimensional scaling.

4.3 Analysis of Sharers

This section reports on the results of multidimensional scaling and cluster analysis performed on sharer profiles weighted according to Equation (1). The sharer profiles were collected over the four-long term observation periods. SPSS was used for multidimensional scaling; it provides two options: ALSCAL and PROXSCAL. Unlike ALSCAL, PROXSCAL can handle similarity and dissimilarity matrices and construct a proximity matrix if the input data does not have a proximity measure. Since these characteristics meet the experimental requirements, PROXSCAL was used in this research.

Data from July 1, 2013 to December 31, 2013 is used to illustrate the analysis that was conducted to discover the underlying relationships between sharer profiles. Since the data has some isolated noise (which implies that the sharers corresponding to the noise points have totally different behavior from others, e.g., sharing less popular cyberlocker links), the noise points were removed to leave 66 sample points. Following this, multidimensional scaling analysis was performed on the 66 sample points. Figure 3 shows the resulting object configuration (top left) where the similarity between sharers is visualized using a two-dimensional representation (stress = 0.0390).

Next, agglomerative hierarchical clustering analysis was performed on the coordinates of the points produced by multidimensional scaling analysis. The clustering result, shown in the matrix tree plot (right panel of

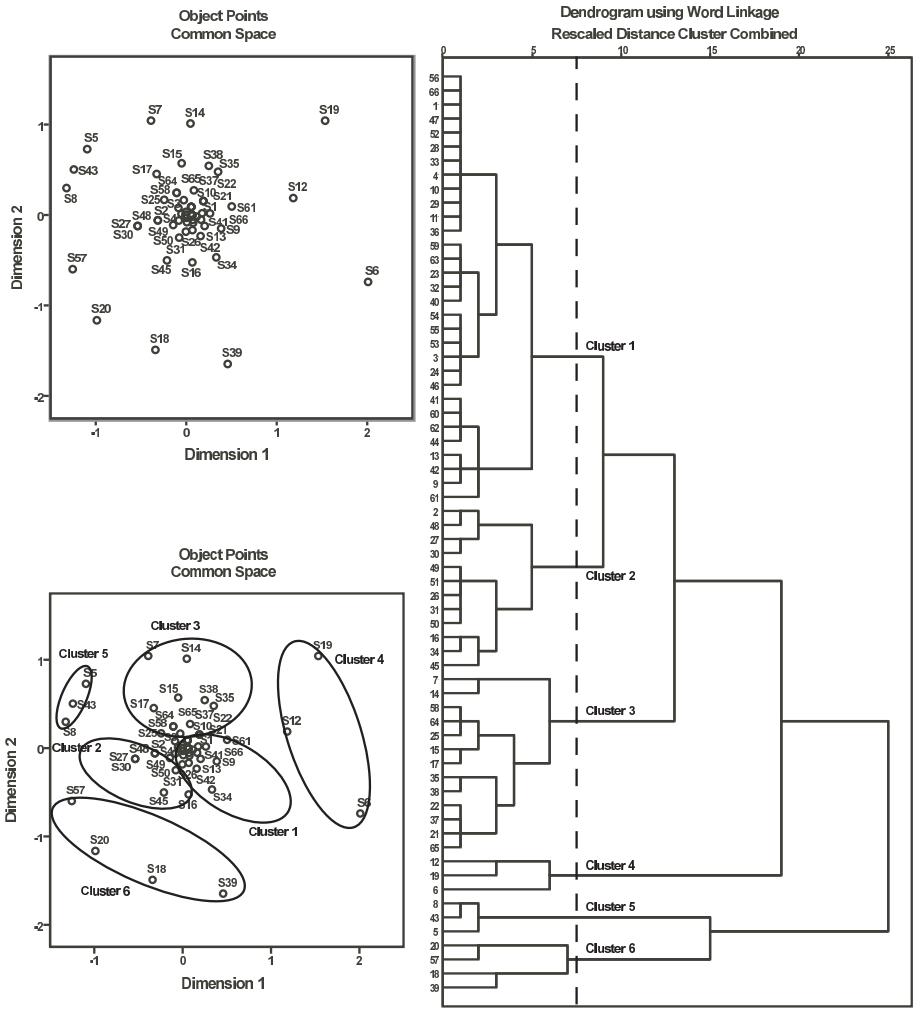

Figure 3. Results using PROXSCAL and agglomerative hierarchical clustering.

Figure 3), is a close reflection of the agglomerative hierarchical clustering algorithm. The left column of nodes represents the objects (sharers) while the remaining nodes represent the clusters to which the objects belong. During the clustering, two very similar objects were first joined at a node and the resulting line was joined to the next closest object or sub-cluster by another line. The length of the line is proportional to the value of the inter-group dissimilarity between its two child nodes. As shown in Figure 3, when a dashed line is drawn in the matrix tree plot, each horizontal line intersecting the dashed line represents a cluster, following which the appropriate clusters can be found. In total, six

Table 6. Hierarchical clusters of sizes three to thirteen.

Cluster	Subjects (Sharers)	Sub-Cluster	Subjects (Sharers)	Cluster Size
1	56, 66, 1, 47, 52, 28, 33, 4, 10, 29, 11, 36, 59, 63, 23, 32, 40, 54, 55, 53, 3, 24, 46, 41, 60, 62, 44, 13, 42, 9, 61	1-1	56, 66, 1, 47, 52, 28, 33, 4, 10, 29, 11, 36	12
		1-2	59, 63, 23, 32, 40, 54, 55, 53, 3, 24, 46	11
		1-3	41, 60, 62, 44, 13, 42, 9, 61	8
2	2, 48, 27, 30, 49, 51, 26, 31, 50, 16, 34, 45			12
3	7, 14, 58, 64, 25, 15, 17, 35, 38, 22, 37, 21, 65			13
4	12, 19, 6			3
5	8, 43, 5			3
6	20, 57, 18, 39			4

clusters were obtained. Figure 3 (bottom left) shows these clusters in a two-dimensional representation.

The advantage of agglomerative hierarchical clustering is that the agglomerative process and the relationships between objects can be viewed clearly. If the cluster size exceeds the maximum cluster size, the cluster is divided into sub-clusters according to the agglomerative process. For example, based on the results in Figure 3 (right panel), Cluster 1 can be further divided into three sub-clusters as shown in Table 6, and nine clusters of sizes three to thirteen are identified. A forensic investigator can then explore the objects within the clusters to determine the cluster to which the piracy suspect belongs.

Table 7. Stress values for the four long-term observation periods.

Observation Period	Stress
2014.01.01 to 2014.06.30	0.0336
2013.07.01 to 2013.12.31	0.0390
2013.01.01 to 2013.06.30	0.0273
2012.07.01 to 2012.12.31	0.0295

Table 7 shows the analysis results for the sharer profiles collected during the four long-term observation periods along with the stress values. The goodness of fit obtained via multidimensional scaling is adequate and the results reflect the relationships between objects. Therefore, if a forensic investigator desires to trace the identity of a piracy suspect, this methodology can narrow the scope of the investigation and identify users with similar sharing behavior. This can help collect more evidence

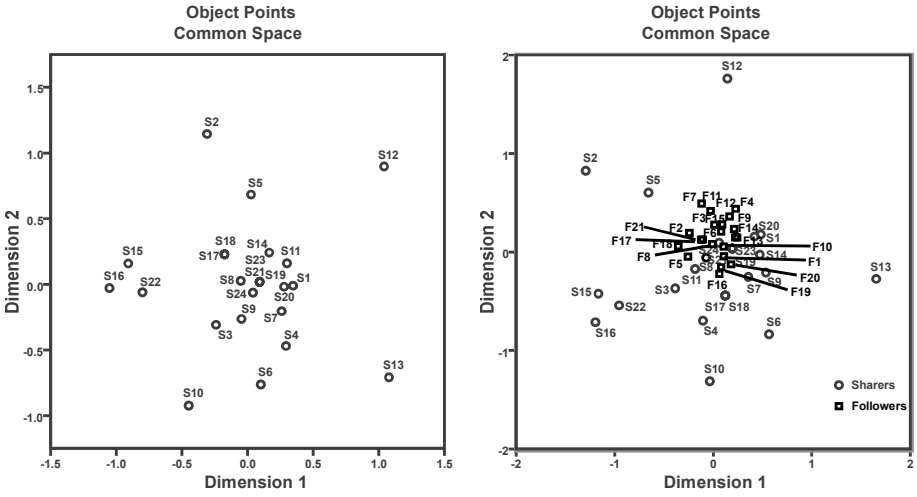

Figure 4. Two-dimensional object configuration of 24 sharers and 21 followers.

about the suspect's cyberlocker-based piracy behavior and also his/her identity.

4.4 Analysis of Sharers and Followers

Multidimensional analysis and cluster analysis were also performed on the sharer and follower profiles collected during the six short-term observation periods. The weighting schemes corresponding to Equations (1) and (2) were applied to create sharer and follower profiles, respectively, which were then compared. Also, the relationships between the sharers and followers were identified and analyzed.

The data from January 1, 2014 to January 31, 2014 is used to illustrate the analysis. During this observation period, data about 26 sharers and 91 followers was collected. The sharer profiles were created using the weighting scheme specified by Equation (1), which ignores the replying behavior of sharers. After removing two isolated noise points, multidimensional scaling analysis was performed on the remaining 24 sharers. The resulting object configuration is shown in Figure 4 (left panel; stress = 0.0516).

Next, the sharer and follower profiles were created using the weighting scheme specified by Equation (2), which considers the sharing and replying behavior of sharers and followers. During the short-term observation period, a number of inactive followers replied to less than three cyberlocker links. These inactive followers were removed and multidimensional scaling analysis was performed on the remaining 21 followers

and 24 sharers (same 24 sharers as before). The resulting object configuration is shown in Figure 4 (right panel; stress = 0.0496).

Note that the object configurations of sharers in the left and right panels of Figure 4 are similar. This implies that the replying behavior of sharers has little discriminating power in determining the relevance of sharers, while the discriminating power of the sharing behavior of sharers is high. Moreover, as seen in Figure 4 (right panel), the object configuration of the followers is agminated, which is readily seen in the distribution. These followers have close relationships with the sharers who are close to them because they replied to the cyberlocker links posted by the sharers. Other sharers are far away from the followers because the cyberlocker links they shared had few, if any, replies from the followers.

5. Conclusions

The framework proposed for digital profiling and identity tracking of sharers and followers of cyberlocker links is very useful in investigations of cyberlocker-based piracy. Experiments with data collected from nine web forums and targeting ten cyberlockers demonstrate that multidimensional scaling and agglomerative hierarchical clustering adequately capture the relationships between sharers and followers in a lower-dimensional hierarchical representation. The framework provides clear insights into the identities and relationships of cyberlocker link sharers and followers, significantly reducing the time and effort needed to investigate cyberlocker-based piracy.

Note that the analytic results only suggest tendencies related to cyberlocker link sharers and followers. Future research will apply the framework to profile and track real cyberlocker pirates, providing useful insights into the effectiveness of the framework in real-world piracy investigations. Also, the research will consider other characteristics of sharers and followers such as online time frames and time periods, types of shared content and different languages to construct more comprehensive profiles for identity tracking and forensic investigations.

References

[1] Alexa Internet, Top Sites in Hong Kong, San Francisco, California (`www.alexa.com/topsites/countries/HK`), 2014.

[2] N. Bamarah, B. Satpute and P. Patil, Web forum crawling techniques, *International Journal of Computer Applications*, vol. 85(17), pp. 36–41, 2014.

[3] W. Bhukya and S. Banothu, Investigative behavior profiling with one class SVM for computer forensics, in *Multi-Disciplinary Trends in Artificial Intelligence*, C. Sombattheera, A. Agarwal, S. Udgata and K. Lavangnananda (Eds.), Springer-Verlag, Berlin Heidelberg, Germany, pp. 373–383, 2011.

[4] I. Borg and P. Groenen, *Modern Multidimensional Scaling: Theory and Applications*, Springer-Verlag, New York, 2005.

[5] R. Drath, Hotfile, Megaupload and the future of copyright on the Internet: What can cyberlockers tell us about DMCA reform? *John Marshall Review of Intellectual Property Law*, vol. 12(205), pp. 204–241, 2012.

[6] Envisional, Technical Report: An Estimate of Infringing Use of the Internet, Cambridge, United Kingdom, 2011.

[7] N. Fathy, N. Badr, M. Hashem and T. Gharib, Enhancing web search with semantic identification of user preferences, *International Journal of Computer Science Issues*, vol. 8(6), pp. 62–69, 2011.

[8] D. Godoy and A. Amandi, A conceptual clustering approach for user profiling in personal information agents, *AI Communications*, vol. 19(3), pp. 207–227, 2006.

[9] E. Hickey, *Serial Murderers and Their Victims*, Wadsworth, Belmont, California, 2012.

[10] J. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika*, vol. 29(1), pp. 1–27, 1964.

[11] M. Liu, Z. Zhang, P. Hui, Y. Qin and S. Kulkarni, Measurement and understanding of cyberlocker URL-sharing sites: Focus on movie files, *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 902–909, 2013.

[12] A. Mahanti, Measurement and analysis of cyberlocker services, *Proceedings of the Twentieth International Conference Companion on the World Wide Web*, pp. 373–378, 2011.

[13] N. Marx, Storage wars: Clouds, cyberlockers and media piracy in the digital economy, *Journal of E-Media Studies*, vol. 3(1), 2013.

[14] S. McKinney and D. Reeves, User identification via process profiling, *Proceedings of the Fifth Annual Workshop on Cyber Security and Information Intelligence Research: Cyber Security and Information Intelligence Challenges and Strategies*, article no. 51, 2009.

[15] F. Murtagh and P. Legendre, Ward's hierarchical clustering method: Which algorithms implement Ward's criterion? *Journal of Classification*, vol. 31(3), pp. 274–295, 2014.

[16] W. Petherick (Ed.), *Serial Crime: Theoretical and Practical Issues in Behavioral Profiling*, Elsevier Academic Press, Burlington, Massachusetts, 2009.

[17] R. Raysman and P. Brown, Cyberlockers, file-sharing and infringement in the cloud, *New Jersey Law Journal*, September 12, 2012.

[18] S. Schiaffino and A. Amandi, Intelligent user profiling, in *Artificial Intelligence: An International Perspective*, M. Bramer (Ed.), Springer-Verlag, Berlin Heidelberg, Germany, pp. 193–216, 2009.

[19] F. Wickelmaier, An Introduction to MDS, Report No. R000-6003, Institute for Electronics Systems, Aalborg University, Aalborg, Denmark, 2003.

[20] N. Zhao, L. Baud and P. Bellot, Characteristic analysis for the cyberlockers files study on Rapidgator, Speedyshare and 1Fichier, *Proceedings of the Eighth International Conference on Information Science and Transactions*, pp. 176–181, 2013.