THE JOURNAL OF
**DIGITAL FORENSICS,
SECURITY AND LAW**

**Journal of Digital Forensics,
Security and Law**

Volume 10 | Number 4

Article 2

2015

Tracking and Taxonomy of Cyberlocker Link Sharers Based on Behavior Analysis

Xiao-Xi Fan

The University of Hong Kong

Kam-Pui Chow

The University of Hong Kong

Follow this and additional works at: <https://commons.erau.edu/jdfsl>

Part of the [Computer Engineering Commons](#), [Computer Law Commons](#), [Electrical and Computer Engineering Commons](#), [Forensic Science and Technology Commons](#), and the [Information Security Commons](#)

Recommended Citation

Fan, Xiao-Xi and Chow, Kam-Pui (2015) "Tracking and Taxonomy of Cyberlocker Link Sharers Based on Behavior Analysis," *Journal of Digital Forensics, Security and Law*: Vol. 10 : No. 4 , Article 2.

DOI: <https://doi.org/10.15394/jdfsl.2015.1208>

Available at: <https://commons.erau.edu/jdfsl/vol10/iss4/2>

This Article is brought to you for free and open access by the Journals at Scholarly Commons. It has been accepted for inclusion in Journal of Digital Forensics, Security and Law by an authorized administrator of Scholarly Commons. For more information, please contact commons@erau.edu.

(c)ADFSL

TRACKING AND TAXONOMY OF CYBERLOCKER LINK SHARERS BASED ON BEHAVIOR ANALYSIS

Xiao-Xi Fan and Kam-Pui Chow
The University of Hong Kong
Department of Computer Science
Hong Kong, 999077
{xxfan, chow}@cs.hku.hk

ABSTRACT

The growing popularity of cyberlocker service has led to significant impact on the Internet that it is considered as one of the biggest contributors to the global Internet traffic estimated to be illegally traded content. Due to the anonymity property of cyberlocker, it is difficult for investigators to track user identity directly on cyberlocker site. In order to find the potential relationships between cyberlocker users, we propose a framework to collect cyberlocker related data from public forums where cyberlocker users usually distribute cyberlocker links for others to download and identity information can be gathered easily. Different kinds of sharing behaviors of forum user are extracted to build the profile, which is then analyzed with statistical techniques. The experiment results demonstrate that the framework can effectively detect profiles with similar behaviors for identity tracking and produce a taxonomy of forum users to provide insights for investigating cyberlocker-based piracy.

Keywords: identity tracking, taxonomy, user profiling, behavior analysis, cyberlocker, piracy

1. INTRODUCTION

Cyberlocker Service is an Internet hosting service which provides an easy way to host user files. Once a user has uploaded a file successfully, the cyberlocker will then generate a unique download URL (called cyberlocker link) for uploader to copy and share to others. The cyberlocker link can be posted and distributed through third-party sites such as public forums, blogs, etc., and then others can download the file from the cyberlocker with the cyberlocker link (Liu, 2013).

However, cyberlocker piracy become even more rampant, e.g. three-quarters contents on RapidShare is estimated to be illegally traded

(Envisional Ltd, 2011). The reason is that some cyberlocker services encourage user to upload and share files by providing rewards programs to pay uploader money. The more download of a file, the more money the uploader will be paid. It prompts uploader to distribute cyberlocker links on public forums as many as possible to attract downloading. Also, the larger of a file, the more money the uploader will be paid. Therefore, Zhao et al. (Zhao, 2013) analyzed the file size and file number distribution on three cyberlockers, and found that the video files which are generally large took more than 60% of hosted bytes on all the three cyberlockers. Besides, the illegal trade is also driven by the willingness of

downloader who would like to avoid paying for contents.

Many cyberlocker services allow user with no connection to a file's uploader to download the file independently. And the uploader's information cannot be found on the downloading page. Since the IP addresses of users are kept anonymous from one another, cyberlocker can hide the identity of the uploader. In this case, the investigation of cyberlocker piracy is difficult, so an effective model is desired to track user identity on cyberlocker.

We know that the data related to shared cyberlocker links can be easily collected from public forums, on which the user information can be also collected and associated with cyberlocker links. Besides, cyberlocker user tends to have multiple accounts on a forum or different forums in order to distribute cyberlocker links. It is helpful for law enforcement agencies if a model is able to find out potential accounts which attribute to the same person. Moreover, it is crucial that law enforcement agents and other officials who combat the crimes have better understanding of the underlying criminal behaviors; therefore, the taxonomy of cyberlocker users is helpful to summarize the behaviors and provide investigative leads. Our previous research (Fan, 2015) proposed a preliminary model to construct profile for forum user with shared cyberlocker links as features. In this research, the objective is to build a framework which extracts more comprehensive sharing behaviors to profile forum users, finds relationships between different profiles, and builds the taxonomy of users to generate general profiles for different categories of users.

2. A FRAMEWORK FOR TRACKING AND TAXONOMY OF

CYBERLOCKER LINK SHARERS

Generally, cyberlocker links are distributed within posts on online public forums. A sharer is defined as the forum user who uploads and shares a cyberlocker link on the forum for others to download the file, who is also the potential uploader of this file on cyberlocker site (i.e. a cyberlocker user who has uploaded and distributed this file). This research proposes a model to build profiles for sharers and analyze the relationships between these profiles.

Figure 1. A Framework for Tracking and Taxonomy of Cyberlocker Link Sharers

The architecture of our proposed model is shown in Figure 1. First, public forum and cyberlocker data is collected by Web Forum Crawler which targets on several popular public forums in Hong Kong. The details of data collection are illustrated in Section 2.1.

Then, the observable online traces are detected which may infer the file sharing characteristics of sharer. Sharer profiles are built in vector space model based on extracted characteristics in Section 2.2. In Section 2.3, Multidimensional Scaling (MDS) is applied to analyze the similarity between sharers and map their relationships to a lower dimensional representation. Then, in Section 2.4, agglomerative hierarchical clustering is performed on the coordinates of the resulting points of MDS analysis to identify clusters containing sharers with similar behavior. These similar sharers may be the same person or have a close relationship in real world. Besides, the taxonomy of sharers is identified and the general profile of each cluster is constructed to outline the sharers' behaviors from different categories, which may help narrow down the scope of sharers which have particular characteristics.

2.1 Data Collection

Since there are a vast number of cyberlocker links posted on public forums, our system is designed to collect data on these public forums. The forum is hierarchical structure which may contain several sub forums. Topics come under the lowest level of sub forums. We follow the definition of topic and post in previous research (Bamrah, 2014) as below.

- **Topic:** a collection of posts, which is displayed from oldest to latest or latest to oldest. A topic contains information such as topic title, original post which is the first post starting the topic, and creation date and time.
- **Post:** a message submitted by a forum user, which contains information such as forum user details, content, and creation date and time.

Since online public forum consists of many web pages, our system is designed to collect

public forum and cyberlocker data using Web Forum Crawler, which traverses World Wide Web in a systematic manner with the intention of gathering data. There are two kinds of crawlers in our system, URL Crawlers and Topic Crawlers. URL Crawlers are responsible for extracting URLs of topics from catalog pages of the forum. After URL Crawler Fetch obtains a URL of catalog page, it will call DNS module to resolve the host address of the corresponding web server. Then it connects to the server to download the page and URL Crawler Parse will extract all URLs of topics from this page and store them in URL Frontier. Then, Topic Crawlers are responsible for extracting useful information from topic pages stored in URL Frontier. The system parses topic information, post information, user information, cyberlocker information and cyberlocker link information into structured data and stores in database to build user profile for further analysis.

Our system crawls data on nine online public forums, Hong Kong Discuss Forum (DISCUSS), UWANTS, HK-PUB, KYO, EYNY, TVBOXNOW, CK101, LALULALU and 8CYBER, which rank high in the most frequently visited websites in Hong Kong according to the statistics published by a web technology company (Alexa, 2015).

2.2 Profile Construction

Crime scene evidence is the evidence collected at the crime scene, which is used to infer the behavior of an offender in traditional crime investigation. However, unlike traditional crime, it is not easy to define the scope of crime scene on Internet since it is too broad. It would be more appropriate to use the term "observable online traces" to describe any traces observed on the Internet. For example, in this research, when a cyberlocker link sharer publishes a post containing a cyberlocker link on a forum, the observable online traces are

the created time of the post, the content of the post, the cyberlocker service used by sharer, the content of the cyberlocker link, and etc.

File sharing characteristics are one's sharing behaviors in the cyber world. Especially, in this research, it refers to one's characteristics when participating in the file sharing activities via cyberlocker services. The relationship of observable online traces and file sharing characteristics is based on direct deduction. The observable online traces infer the sharer's file sharing characteristics. In other words, the sharer's sharing behaviors are reflected by the fact of online traces. For instance, if we find a post containing a cyberlocker link which was published at 2:00 pm on Wed, 6 March 2015, it infers that the sharer was online around 2:00 pm on that day. The observable online trace is the created time of the post, while the file sharing characteristic is the sharer was online around 2:00 pm on 6 March 2015 (during work hours on workday).

In this work, file sharing characteristics we focused are divided into five domains. The most representative file sharing characteristics in each domain are extracted based on the sharer's observable online traces. The sharer's profile is comprised of his/her file sharing characteristics which describes his/her regular sharing behaviors.

Sharer profiles are built based on collected data in the same period (observation period). Vector space model usually used in information retrieval is applied to build sharer profile. A sharer is represented as a vector in the following form (Eq. 1), which contains N features. Each feature corresponds to a file sharing characteristic of the sharer in observation period.

$$s_j = (S_1, S_2, \dots, S_N) \quad (1)$$

To yield a more interpretive analysis, all features are normalized to Boolean variables and coded as zero (0 - false) or one (1 - true).

This results in a set of 37 features for sharers, which is detailed in five domains.

1) Domain 1: time period.

a. The time period the sharer usually shares or replies cyberlocker links (workday/weekend).

The corresponding observable online trace is the time when the sharer shares/replies a cyberlocker link. The file sharing characteristics are defined as below. If the proportion of "sharing or replying on workday" is higher than "sharing or replying on weekend", and the number of cyberlocker links "sharing or replying on workday" is not less than 2, then $SV_1 = 1$, and vice versa.

S_1 : usually share or reply cyberlocker links on workday.

S_2 : usually share or reply cyberlocker links on weekend.

b. The time period the sharer usually shares or replies cyberlocker links on workday (during work hours/outside work hours).

Similar to Feature a, if the proportion of "sharing or replying during work hours on workday" is higher than "sharing or replying outside work hours on workday", and the number of cyberlocker links "sharing or replying during work hours on workday" is not less than 2, then $S_3 = 1$, and vice versa.

S_3 : usually share or reply cyberlocker links during work hours on workday.

S_4 : usually share or reply cyberlocker links outside work hours on workday.

c. The time period the sharer usually shares or replies cyberlocker links on weekend (during work hours/outside work hours).

Similar to Feature a, if the proportion of "sharing or replying during work hours on

weekend” is higher than “sharing or replying outside work hours on weekend”, and the number of cyberlocker links “sharing or replying during work hours on weekend” is not less than 2, then $S_5 = 1$, and vice versa.

S_5 : usually share or reply cyberlocker links during work hours on weekend.

S_6 : usually share or reply cyberlocker links outside work hours on weekend.

2) Domain 2: cyberlocker service.

d. The cyberlocker services the sharer usually uses to share cyberlocker links.

The corresponding observable online trace is the cyberlocker service identified from a shared cyberlocker link of the sharer. The file sharing characteristics are defined as below. If the proportion of “the usage of a cyberlocker service” is higher than 30%, and the number of cyberlocker links “using a cyberlocker service” is not less than 2, then $S_i = 1$, and vice versa.

S_7 : usually share cyberlocker links using Bitshare.

S_8 : usually share cyberlocker links using Depositfiles.

S_9 : usually share cyberlocker links using FileFactory.

S_{10} : usually share cyberlocker links using Fileim.

S_{11} : usually share cyberlocker links using FilePost.

S_{12} : usually share cyberlocker links using HulkShare.

S_{13} : usually share cyberlocker links using MediaFire.

S_{14} : usually share cyberlocker links using RapidShare.

S_{15} : usually share cyberlocker links using Sendspace.

S_{16} : usually share cyberlocker links using Ziddu.

3) Domain 3: No. of cyberlockers.

e. No. of cyberlocker services the sharer uses to share cyberlocker links (different / single).

The corresponding observable online trace is the cyberlocker service identified from a shared cyberlocker link of the sharer. This feature is to count how many cyberlocker services the sharer uses. The file sharing characteristics are defined as below.

S_{17} : use more than 4 cyberlocker services.

S_{18} : use 2 to 4 cyberlocker services.

S_{19} : use single cyberlocker service.

4) Domain 4: No. of cyberlocker links shared and replied.

f. The activity that has higher proportion (sharing cyberlocker links/replying cyberlocker links).

The corresponding observable online trace is the fact that the sharer shares/replies a cyberlocker link. The file sharing characteristics are defined as below. If the sharer has shared or replied to a cyberlocker link multiple times, then just one time will be counted. If the proportion of “sharing cyberlocker links” is higher than “replying cyberlocker links”, then $S_{20} = 1$, and vice versa.

S_{20} : share more cyberlocker links than replying.

S_{21} : reply more cyberlocker links than sharing.

g. No. of cyberlocker links the sharer shares.

The corresponding observable online trace is the fact that the sharer shares a cyberlocker

link. The file sharing characteristics are defined as below.

S_2 : share no more than 10 cyberlocker links.

S_2 : share more than 10 but no more than 50 cyberlocker links.

S_2 : share more than 50 cyberlocker links.

h. No. of cyberlocker links the sharer replies.

The corresponding observable online trace is the fact that the sharer replies a cyberlocker link. The file sharing characteristics are defined as below.

S_2 : reply no more than 10 cyberlocker links.

S_2 : reply more than 10 but no more than 50 cyberlocker links.

S_2 : reply more than 50 cyberlocker links.

5) Domain 5: category of cyberlocker links shared and replied.

i. The categories of cyberlocker links that the sharer usually shares.

The corresponding observable online trace is the category to which a shared cyberlocker link belongs. The file sharing characteristics are defined as below. If the proportion of “the sharing of a category” is higher than 30%, and the number of shared cyberlocker links of this category is not less than 2, then $S_i = 1$, and vice versa.

S_2 : usually share cyberlocker links of music.

S_2 : usually share cyberlocker links of movies.

S_3 : usually share cyberlocker links of anime.

S_3 : usually share cyberlocker links of TV dramas.

S_3 : usually share cyberlocker links of variety shows.

j. The categories of cyberlocker links that the sharer usually replies.

The corresponding observable online trace is the category to which a replied cyberlocker link belongs. The file sharing characteristics are defined as below. If the proportion of “the replying of a category” is higher than 30%, and the number of replied cyberlocker links of this category is not less than 2, then $S_i = 1$, and vice versa.

S_3 : usually reply cyberlocker links of music.

S_3 : usually reply cyberlocker links of movies.

S_3 : usually reply cyberlocker links of anime.

S_3 : usually reply cyberlocker links of TV dramas.

S_3 : usually reply cyberlocker links of variety shows.

2.3 Multidimensional Scaling

Multidimensional Scaling (MDS) has been widely used as an important tool for analyzing the patterns of crime behavior (Hickey, 2012), which is a data analysis technique for identifying underlying patterns or structure of a set of objects. In particular, it is suitable for analyzing similarity and dissimilarity data that can display similarity and dissimilarity between objects in a geometrical representation by a number of distance models. The Euclidean distance model is the most popular distance model in MDS and it is also applied in this research.

The outcome of MDS analysis is a spatial configuration, in which the objects are represented as points. The points in this spatial representation are arranged in such a

way that the points corresponding to similar objects are located close together, while those corresponding to dissimilar objects are located far apart. If we consider sharers as some objects, the idea of MDS is to arrange these objects onto a lower dimensional map where the distance between two objects represents the observed correlation of the corresponding sharers.

2.4 Cluster Analysis

MDS arranges the profiles as points on a lower dimensional map without losing the relationships between each other. We then perform agglomerative hierarchical clustering on the coordinates of these resulting points to identify potential clusters in which sharers have similar file sharing characteristics.

In agglomerative hierarchical clustering, since clusters generated in early stages are nested in those generated in later stages, it can produce an ordering of the objects which is informative for hierarchically display. We apply matrix tree plot to show the clustering outcome which can visually highlight clustering relationships. Ward's minimum variance method (Murtagh, 2011) is adopted as the criterion for choosing the pair of clusters to merge at each step since it is more aligned with our requirements for clustering in practice, because it tends to generate smaller

distance between two small clusters so that they have high priority to be merged.

After the clustering process, the sharer profiles falling in the same cluster have more possibility belonging to the same person in real world, which may help investigator collect more evidences about the target's identity. Besides, the general profile of each cluster can be drawn to detect the characteristics of these clusters, which may help investigator find a specific group, e.g. "professional sharers" who share lots of cyberlocker links using cyberlocker services that have rewards programs.

3. EXPERIMENTS

We collected data in four observation periods and the basic statistics of sharers is concluded in Table 1. We also collected the traces of repliers who have replied the posts containing a cyberlocker link but have not uploaded any cyberlocker links. Among all sharers and repliers, we can find that the proportion of sharers is very small, which means thousands of cyberlocker links are shared by only a small number of forum users. Besides, we find that these three forums, TVBOXNOW, and LALULALU and CK101, have the largest number of sharers and repliers, which means these three active forums are forum users' favorite sites to distribute cyberlocker links.

Table 1
Basic Statistics of Sharers in Four Observation Periods

Observation Period	No. of Sharers/Repliers		Percentage of Sharers/Repliers on Different Forums								
			DISCUSS	UWANTS	HK-PUB	KYO	EYMY	TVBOXNOW	CK101	LALULALU	8CYBER
2014.01.01 – 2014.06.30	Sharers & Repliers	1649	0.18%	1.09%	1.46%	0.12%	1.21%	56.58%	7.22%	31.53%	0.61%
	Sharers	64	4.69%	10.94%	12.5%	1.56%	1.56%	29.69%	25%	12.5%	1.56%
2013.07.01 – 2013.12.31	Sharers & Repliers	854	0.98%	3.32%	5.77%	2.21%	2.95%	26.78%	6.76%	50.98%	0.25%
	Sharers	74	11.11%	8.33%	9.72%	8.33%	2.78%	22.22%	16.67%	18.06%	2.78%
2013.01.01 – 2013.06.30	Sharers & Repliers	684	0.73%	2.34%	1.02%	1.90%	0.73%	67.54%	10.23%	14.91%	0.58%
	Sharers	91	3.30%	7.69%	7.69%	7.69%	1.10%	35.16%	15.38%	19.78%	2.20%
2012.07.01 – 2012.12.31	Sharers & Repliers	5580	0.11%	2.06%	0.72%	1.81%	0.22%	66.83%	12.42%	15.82%	0.02%
	Sharers	215	2.79%	8.37%	3.72%	8.37%	1.86%	34.88%	26.05%	13.49%	0.47%

3.1 Evaluation Metric

We apply STRESS measure to judge the goodness of fit, which is the most commonly-used measure to evaluate MDS (Wickelmaier, 2003). If the input matrix to MDS is not a proximity matrix, which represents the distance (similarity/dissimilarity) between rows or columns, proximities will be created. Then a monotonic transformation of the proximities is conducted to yield scaled proximities. The objective of MDS is to find a configuration of points that minimizes the squared differences between the optimally scaled proximities and the distances between points. In other words, MDS is to minimize STRESS, which is defined as

$$\text{STRESS} = \sqrt{\frac{\sum (f(p) - d)^2}{\sum d^2}} \quad (2)$$

where $f(p)$ is a monotonic transformation of the proximities p , and d is the distances between points.

Table 2
STRESS and Goodness of Fit of MDS

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

In order to provide a standard, Kruskal (1964) gives the guidelines to interpret the goodness of fit of MDS based on the STRESS value (Table 2).

3.2 Detection of Similar Profiles

We use the data from 1 July 2013 to 31 Dec 2013 to illustrate the analysis conducted to discover underlying relationships between the collected sharers' profiles. After MDS analysis, the resulting object configuration is shown in Figure 2, which visualizes the

similarity between sharers in two-dimensional representation (STRESS=0.0731).

Then, we performed agglomerative hierarchical clustering analysis on the coordinates of the resulting points produced by MDS analysis. The clustering result is shown in matrix tree plot (Figure 3), which is a close reflection of agglomerative hierarchical clustering algorithm. If we put a dashed line on the matrix tree plot, each horizontal line intersecting with the dashed line represents a cluster. 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, it will be divided into sub-clusters according to the agglomerative process, e.g. cluster 5-8 are sub-clusters produced in the second division. The corresponding result on two-dimensional representation is shown in Figure 4. The investigators can then explore the objects within the cluster to which the pirate suspect belongs.

Figure 2. Two-dimensional Object Configuration

Figure 3. Matrix Tree Plot

The same analysis was performed on sharers' profiles collected in other three observation periods and the STRESS is concluded in Table 3. It implies the analysis result can well reflect the relationships between objects. If investigators desire to trace the identity of a pirate suspect, the model can find out the users with similar sharing characteristics and narrow down the scope of investigation, which can help collect more evidences about the suspect's identity and piratic activities.

Figure 4. Two-dimensional Representation of Clusters

Table 3

STRESS of Four Observation Periods

Observation Period	STRESS
2014.01.01 – 2014.06.30	0.0830
2013.07.01 – 2013.12.31	0.0731
2013.01.01 – 2013.06.30	0.0784
2012.07.01 – 2012.12.31	0.0804

3.3 Construction of General Profiles

In Section 3.2, the agglomerative hierarchical clustering analysis yields a taxonomy comprising eight types of sharers, each with different sets of file sharing characteristics. It is significant to construct the general profile of each type, which would aid forensic scientists and practitioners in understanding the behaviors and devising efficient and effective investigation plans. Therefore, we performed statistical analysis on each type of sharers and summarized the representative characteristics of each type, where “most of them” means

more than 50% of sharers have this characteristic, while “several of them” means less than 50% of sharers have this characteristic.

1) Profile 1 (Cluster 1): sharing a lot cyberlocker links using single cyberlocker service.

Time period: They usually share or reply cyberlocker links on workday (SV₁). No matter on workday or weekend, they usually share or reply cyberlocker links outside work hours (SV₄, SV₆).

Cyberlocker service: Most of them usually share cyberlocker links using Bitshare (SV₇), while several of them using FileFactory (SV₉) or FilePost (SV₁).

No. of cyberlockers: *Most of them use single cyberlocker service (S₁).*

No. of cyberlocker links shared and replied: They share more cyberlocker links than replying (SV₂). *They share more than 50 cyberlocker links (S₂).* Most of them reply more than 10 but no more than 50 cyberlocker links (SV₂), while several of them reply more than 50 cyberlocker links (SV₂).

Category of cyberlocker links shared and replied: Several of them usually share and reply cyberlocker links of movies and anime (SV₂, SV₃, SV₃, SV₃).

2) Profile 2 (Cluster 2): sharing a lot cyberlocker links using multiple cyberlocker services.

Time period: They usually share or reply cyberlocker links on workday (SV₁). No matter on workday or weekend, they usually share or reply cyberlocker links during work hours (SV₃, SV₅).

Cyberlocker service: They usually share cyberlocker links using Bitshare (SV₇),

while most of them usually share cyberlocker links using FileFactory (SV₉).

No. of cyberlockers: *They use 2 to 4 cyberlocker services (S₁).*

No. of cyberlocker links shared and replied: They share more cyberlocker links than replying (SV₂). *They share more than 50 cyberlocker links (S₂).* Most of them reply more than 10 but no more than 50 cyberlocker links (SV₂).

Category of cyberlocker links shared and replied: Most of them usually share and reply cyberlocker links of movies (SV₂, SV₃).

3) Profile 3 (Cluster 3): sharing little cyberlocker links of movies.

Time period: They usually share or reply cyberlocker links outside work hours (SV₄) on workday (SV₁).

Cyberlocker service: Most of them usually share cyberlocker links using Bitshare (SV₇), while several of them using FileFactory (SV₉).

No. of cyberlockers: They use single cyberlocker service (SV₁).

No. of cyberlocker links shared and replied: They share more cyberlocker links than replying (SV₂). *Most of them share no more than 10 cyberlocker links (S₂).* They reply no more than 10 cyberlocker links (SV₂).

Category of cyberlocker links shared and replied: *Most of them usually share cyberlocker links of movies (S₂).* Several of them usually reply cyberlocker links of movies (SV₃).

4) Profile 4 (Cluster 4): sharing some cyberlocker links of movies.

Time period: They usually share or reply cyberlocker links on weekend (SV_2). No matter on workday or weekend, they usually share or reply cyberlocker links during work hours (SV_3 , SV_5).

Cyberlocker service: Most of them usually share cyberlocker links using Bitshare (SV_7), while several of them using FileFactory (SV_9).

No. of cyberlockers: Most of them use single cyberlocker service (SV_1).

No. of cyberlocker links shared and replied: They share more cyberlocker links than replying (SV_2). *Most of them share more than 10 but no more than 50 cyberlocker links (S_2).* Most of them reply no more than 10 cyberlocker links (SV_2).

Category of cyberlocker links shared and replied: *Most of them usually share cyberlocker links of movies (S_2),* while several of them usually share cyberlocker links of anime (SV_3). Several of them usually reply cyberlocker links of movies and anime (SV_3 , SV_3).

5) Profile 5 (Cluster 5): sharing one cyberlocker link.

They just *share one cyberlocker link without reply* any cyberlocker link in observation period. Under normal conditions, these cyberlocker links are shared by users who come to the site infrequently and never share other cyberlocker links again.

6) Profile 6 (Cluster 6): sharing little cyberlocker links of music outside work hours on weekend.

Time period: They usually share or reply cyberlocker links outside work hours (S_6) on weekend (S_2).

Cyberlocker service: Most of them usually share cyberlocker links using FilePost (SV_1), while several of them using MediaFire (SV_1) or Sendspace (SV_1).

No. of cyberlockers: They use single cyberlocker service (SV_1).

No. of cyberlocker links shared and replied: Most of them share more cyberlocker links than replying (SV_2). They share no more than 10 cyberlocker links (S_2). Also they reply no more than 10 cyberlocker links (SV_2).

Category of cyberlocker links shared and replied: Most of them usually share cyberlocker links of music (S_2). Several of them usually reply cyberlocker links of movies (SV_3).

7) Profile 7 (Cluster 7): sharing some cyberlocker links of music.

Time period: They usually share or reply cyberlocker links on workday (SV_1). No matter on workday or weekend, they usually share or reply cyberlocker links outside work hours (SV_4 , SV_6).

Cyberlocker service: They usually share cyberlocker links using Bitshare (SV_7).

No. of cyberlockers: They use single cyberlocker service (SV_1).

No. of cyberlocker links shared and replied: They share more cyberlocker links than replying (SV_2). *They share more than 10 cyberlocker links (S_2 , S_2).* They reply no more than 10 cyberlocker links (SV_2).

Category of cyberlocker links shared and replied: *Most of them usually share cyberlocker links of music (S_2).* Several of them usually reply cyberlocker links of music (SV_3).

8) Profile 8 (Cluster 8): sharing little cyberlocker links of music outside work hours on workday.

Time period: They usually share or reply cyberlocker links outside work hours (SV4) on workday (S_1).

Cyberlocker service: Most of them usually share cyberlocker links using Bitshare (SV7). Several of them using FilePost, MediaFire, Sendspace and Ziddu (SV1, SV1, SV1, SV1).

No. of cyberlockers: They use single cyberlocker service (SV1).

No. of cyberlocker links shared and replied: They share more cyberlocker links than replying (SV2). Most of them share no more than 10 cyberlocker links (S_2). They reply no more than 10 cyberlocker links (SV2).

Category of cyberlocker links shared and replied: Most of them usually share cyberlocker links of music (S_2), while several of them usually share cyberlocker links of TV dramas (SV3). Several of them usually reply cyberlocker links of TV dramas (SV3).

Among these general profiles, we can find Profile 2 is the “professional” sharer who uses different kinds of cyberlocker services, which provide rewards programs such as Bitshare and FileFactory, to distribute many cyberlocker links. This kind of sharers usually share cyberlocker links of movies that may be because movies are large files so the sharers can be paid more money. The sharers are usually active during work hours on workday that infers the sharers may not have regular jobs. Compared to Profile 2, Profile 1 is usually active outside work hours. This kind of sharer also shares a lot cyberlocker links but uses single cyberlocker service, which is not as professional as Profile 2.

Besides, Profile 4 and Profile 7 are “medium-level” sharers who share more than 10 cyberlocker links. Profile 4 usually shares cyberlocker links of movies, while Profile 7 usually shares cyberlocker links of music. The remaining profiles are summarized in four profiles with different sharing time and sharing category. These profiles do not share cyberlocker links regularly.

Through this approach, we can find that general profiles with distinguished characteristics can be detected, which makes it possible for investigators to focus on a particular profile, especially the “professional” sharer, to help monitor and investigate cyberlocker piracy.

4. CONCLUSION

In this research, a framework for profiling cyberlocker link sharer is proposed. We introduce five domains from which the file sharing characteristics of the sharer are extracted, namely time period, cyberlocker service, No. of cyberlockers, No. of cyberlocker links shared and replied, and category of cyberlocker links shared and replied. The sharers’ file sharing characteristics are reflected by their observable online traces collected on nine public forums. The sharer profiles are built based on the file sharing characteristics and analyzed with statistical techniques including multidimensional scaling and agglomerative hierarchical clustering. The results can detect profiles with similar behaviors which may belong to the same person in real world, and lead to a taxonomy of several types of sharers with different sharing characteristics. This framework suggests a kind of tendency or higher probability of relationships between sharers. Currently the internal evaluation has been conducted. For future work, we will carry out empirical tests on the data with known class

labels to adjust the model and provide more accurate prediction for identity tracking.

REFERENCES

- Alexa. (2015). Top sites in Hong Kong. Retrieved from <http://www.alexa.com/topsites/countries/HK>.
- Bamrah, N. H., Satpute, B. S., & Patil, P. (2014). Web Forum Crawling Techniques. *International Journal of Computer Applications*, 85(17).
- Envisional Ltd. (2011). Technical report: an estimate of infringing use of the Internet. Retrieved from http://documents.envisional.com/docs/Envisional-Internet_Usage-Jan2011.pdf.
- Fan, X., Chow, K. P., & Xu, F. (2015). Profiling and tracking a cyberlocker link sharer in a public web forum. In *Advances in Digital Forensics XI* (pp. 97-113). Springer Berlin Heidelberg.
- Hickey, E. W. (2012). *Serial murderers and their victims*. Cengage Learning.
- Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1), 1-27.
- Liu, M., Zhang, Z., Hui, P., Qin, Y., & Kulkarni, S. R. (2013). Measurement and understanding of Cyberlocker URL-sharing sites: Focus on movie files. In *Advances in Social Networks Analysis and Mining (ASONAM)*, 2013 IEEE/ACM International Conference on (pp. 902-909). IEEE.
- Murtagh, F., & Legendre, P. (2011). Ward's hierarchical clustering method: Clustering criterion and agglomerative algorithm. *arXiv preprint arXiv:1111.6285*.
- Wickelmaier, F. (2003). *An introduction to MDS*. Sound Quality Research Unit, Aalborg University, Denmark.
- Zhao, N., Baud, L., & Bellot, P. (2013). Characteristics analysis for the cyberlockers files study on Rapidgator Speedyshare and 1Fichier. In *Internet Technology and Secured Transactions (ICITST)*, 2013 8th International Conference for (pp. 176-181). IEEE.