Chapter 6

VALIDATION RULES FOR ENHANCED FOXY P2P NETWORK INVESTIGATIONS

Ricci Ieong and Kam-Pui Chow

Abstract Experiments with the Foxy P2P network have demonstrated that the first uploader of a file can be identified when search queries are submitted to all the network nodes during initial file sharing. However, in real Foxy networks, file search queries are not transmitted to the entire Foxy network and this process may not identify the first uploader. This paper presents a set of validation rules that validate the observed first uploader. The validation rules define the seeder curve that consistently describes the number of uploaders over time. Analysis of four scenarios shows improved accuracy at detecting the first uploader and that, in situations with insufficient competition for file content, the first uploader may not be identified precisely.

Keywords: Peer-to-peer networks, Foxy, first seeder, network simulation

1. Introduction

The Foxy peer-to-peer (P2P) network, which employs the Gnutella 2 protocol, has for several years been a popular file-sharing P2P application in Chinese communities. When the Government of Taiwan shut down the Foxy client publisher in 2009, the use of the Foxy network dramatically decreased and leakage of files through the network dropped temporarily. However, recent incidents have brought considerable media attention. In August and September 2013, sensitive Hong Kong government documents were leaked using the Foxy network [1, 6]. In such situations, if the appropriate legal entities can identify the first uploader of the sensitive files, then they can reduce or even eliminate the leaks at an early stage.

Ieong, *et al.* [5] have shown that the seeder curve – a plot of the number of uploaders (seeders) over time – provides information about

G. Peterson and S. Shenoi (Eds.): Advances in Digital Forensics X, IFIP AICT 433, pp. 73–83, 2014.

the number of seeders as well as the stage of P2P file sharing. However, monitoring rules derived for detecting the first uploader using the seeder curve have high error rates as the number of hub nodes increases [4]. A second monitoring strategy used three rules to monitor search queries in the Foxy P2P network to identify the first uploader [3]. The rules leverage the fact that, in search-based P2P networks, a user locates files by executing a search query in the network. Hub nodes relay and filter search queries from the leaf nodes in the Foxy network. Thus, it is possible to identify the first uploader by monitoring queries at hub nodes.

This paper examines the effects of large numbers of hub nodes on monitoring accuracy. Increasing the number of hubs from one to 1,000 indicates that the number of hubs affects the overall accuracy of the monitoring rules by introducing delays in the appearance of the first uploader without affecting the shape of the seeder curve. To counteract this behavior, supplementary rules are proposed to enhance the accuracy of validation if the first uploader is identified during the slow-rising or rapid-rising periods of the seeder curve. If the first uploader is not identified before the rapid-rising period, then the identified individual is unlikely to be the first uploader.

2. Foxy P2P Query Monitoring

The rules utilized to monitor the Foxy network were presented in a previous study [3]. When file-sharing activities are observed, the monitoring rules initiate file content requests to the nodes that possess the targeted file and then determine if the observed file contains the entire file content. The first uploader returned in the uploader lists that have been verified to possess the entire file content is considered to be first seeder. This rule is referred to as the file existence verification function.

Figure 1 shows a simplified diagram of a Foxy P2P network with four hubs. The solid lines denote the connections between the hubs and leaf nodes. The black and white leaf nodes represent the uploader and downloaders, respectively. The dashed line indicates the direction of file download.

Previous experiments have tested the effect of a small number of hubs on the accuracy of first seeder identification [4]. When a small number of hubs is introduced to the Foxy network, the probability of identifying the first seeder remains the same. To evaluate the functionality of the monitoring rule, it is assumed that the hub node broadcasts all the search queries to all the leaf nodes in the network in a simulated environment. Therefore, all the hubs share the same list of seeders, such that all the leaf

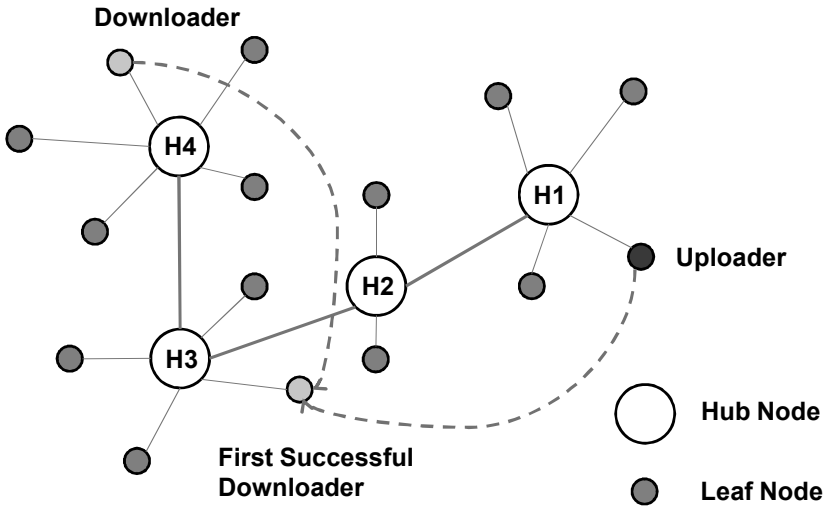

Figure 1. Foxy P2P network with four hubs.

nodes can identify the uploaders. If the search query reaches the entire Foxy network, the identified first uploader should be the first uploader of the file.

However, in a real Foxy network, a hub node does not broadcast a search query message to all the nodes; only a limited set of nodes receive the search query [3]. Thus, a file search request may only be observed by leaf nodes that are connected to a hub where the search query can be accessed. Additionally, since the hub nodes serve as the repository for the list of available uploaders close to the downloader, when a downloader is connected to the hub with the uploader (H3 in Figure 1), its download requests are accepted first. In contrast, although the initial search queries of downloaders located at hubs not connected to the uploader were submitted earlier (e.g., H4 in Figure 1), these downloaders can initiate downloads only after the nodes connected to the hub because uploaders have initiated and completed their file downloads (e.g., leaf node H3 in Figure 1).

The hub node connectivity in a search-based P2P file-sharing network means that, as the number of hub nodes increases, the probability of identifying the first seeder through the investigation algorithm is only affected by the probability of the first uploader and the monitoring node connected to the same hub.

With this implicit nature of the partial uploader list, the investigation algorithm can only confirm if the identified uploader is the first observable full content file holder among the uploaders in the localized

lists of connected hubs, even if the file existence verification function can identify a first seeder. However, it cannot be ascertained if the observed uploader is the actual first uploader. When two full content uploaders are simultaneously present in the network, no mechanism exists to distinguish the first seeder from another seeder unless the target machines are analyzed forensically.

The file existence verification function cannot validate if the collected uploader lists fully cover all the uploaders. Thus, a validation scheme should be designed to utilize the results from the file existence verification function.

2.1 First Uploader Validation Rules

Three validation rules were derived to confirm if the first uploader can be correctly identified in a given situation. The rules focus on the fact that identifying the first uploader depends on the monitoring node that receives the initiation time of the search query. This can be affected by the distance from the query and the size of the file being shared. The first two rules are concerned with the issue of the shared file being small and, therefore, being replicated quickly. The two rules are:

- **Rule 1:** If the time required to download the file is short, then the probability of confirming the observed first uploader is very small. Therefore, the observed uploader should be rejected.

- **Rule 2:** If the entire file existence verification test is completed by the queried uploader in a short time, then the observed first uploader is less likely to be the actual first uploader. Therefore, the observed uploader should be rejected.

Although numerous downloaders are present in a P2P network, only one uploader has the full file content when the monitoring node initiates the search during the slow-rising period. Many detected uploaders from the list should be confirmed as partial file-content uploaders only, especially during the slow-rising period. A third rule then confirms that there is only one individual with the full version of the file. This rule is given by:

- **Rule 3:** In the list of returned uploaders, if all the detected uploaders are full file-content uploaders with no partial file-content uploaders detected by the file existence verification test during the initial stage, then the observed uploader should be rejected.

3. Experimental Design

Simulations were performed in a Foxy ns-3 simulation environment [4] using the monitoring and validation rules. All the parameters other than the number of hub nodes and the size of the files being shared were the same as specified in [3]. The request broadcast cascade level (defined by the time-to-live (TTL) value in the query) was set to three and the maximum number of connected hubs for a leaf node was set to three according to the GnucDNA implementation [2]. The number of leaf nodes was set to 200, number of downloaders to 100, search time to approximately 500 seconds and transfer bandwidth to 1,024 KB.

Four experiments were performed to determine the effect of hub nodes and file size on identifying the first uploader. In each experiment (set), an average was obtained from five simulations using the same parameters.

- **Set 1:** This set of simulation experiments focused on the random search time with ten hub nodes. Two hundred of the 500 leaf nodes were downloaders. The transferred file size was 13 MB. The mean data rate was approximately 1 Mbps.

- **Set 2:** This set of simulation experiments was performed to analyze the effect of the number of hub nodes on the accuracy of the validation rules. The parameters were the same as those used in Set 1, except that the number of hubs was increased to 1,000 nodes.

- **Set 3:** This set of simulation experiments was derived from the monitoring record of a popular television episode. The file size was set to 100 MB. One hundred and thirty-five of the 200 leaf nodes were downloaders; the leaf nodes were connected to 1,000 hub nodes. The mean data transfer rate was approximately 10 Mbps.

- **Set 4:** This set of simulation experiments was derived from the monitoring record in which a small, popular file was shared over the network. Files of size 512 KB were used in the simulation. One hundred thirty-five of the 200 leaf nodes were downloaders; the leaf nodes were connected to one hub node. The hub node was reduced to one to focus on the validation rules in the simulation experiments. The mean data transfer rate was approximately 1 Mbps.

The simulation results for the four experiment sets were evaluated using four criteria:

Table 1. Simulation experiment results.

Experiment	Correct Identification	Correct Reject	False Reject (Type I)	False Accept (Type II)
Set 1	60%	40%	0%	0%
Set 2	0%	100%	0%	0%
Set 3	0%	20%	80%	0%
Set 4	0%	80%	20%	0%

- **Criterion 1:** A correct identification occurs when the first seeder is identified correctly and is validated correctly by the validation rules.
- **Criterion 2:** A correct rejection occurs when a first seeder is identified incorrectly and is rejected correctly by a validation rule.
- **Criterion 3:** A false rejection or Type I error occurs when the first seeder is identified correctly but is rejected incorrectly by a validation rule.
- **Criterion 4:** A false acceptance or Type II error occurs when a first seeder is identified incorrectly but is validated correctly by the validation rules.

4. Experimental Results

Table 1 shows the results of the simulation experiments. When the P2P investigation algorithm and validation rules were applied to the four experiment sets, more than half of the experiments correctly rejected or confirmed the uploader. However, in two experiment sets, most of the actual seeders were rejected by the validation rules.

4.1 Set 1 Results

Figure 2 shows the seeder curve for Set 1. The two indicators show the times of the initial search query responses of the monitoring node for the two correct rejections. The results show that 40% of the simulations did not locate an uploader in the network. The rejections are correct because the monitoring node submitted the first search query after the rapid-rising period. When the number of available uploaders exceeded the response limit, uploader responses were filtered from the monitoring node list, and the validation rules correctly rejected these first uploaders. The result also demonstrates that, in the case of search-based P2P file-

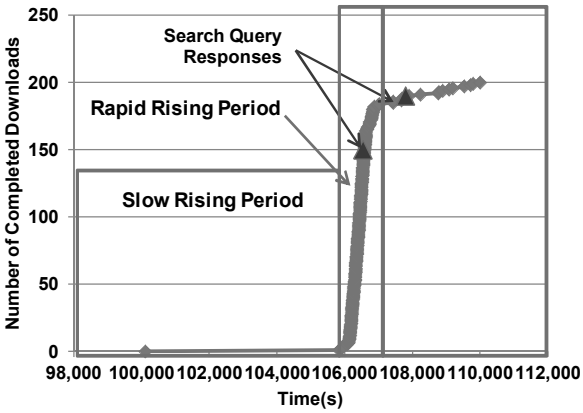

Figure 2. Seeder curve for Set 1.

sharing networks, accurate results can only be obtained if the search is initiated during the slow-rising period.

4.2 Set 2 Results

Set 2 modified Set 1 by increasing the number of hubs and reducing the file size. Two changes are expected in the output. First, with more hubs involved, the probability of correctly identifying the first uploader should be reduced. This is because detection depends on the probability that the monitoring node and the uploader are connected to the same hub node. Second, with a small file size, the download duration decreases and the file should propagate rapidly across the network.

Figure 3 shows the results from a Set 2 simulation. The upper graph shows the seeder curve for the network while the lower graph shows the seeder curve created by the monitoring node. However, the two validation tests failed. The file existence verification was extremely short (rejected by Rule 2). Also, only one partial uploader was identified throughout the verification period (rejected by Rule 3). Since the monitoring node did not see the uploaded file until there were 20 uploaders, the validation rule correctly rejected the incorrectly-identified first uploader.

4.3 Set 3 Results

Figure 4 shows the results from a Set 3 simulation. The upper graph shows the seeder curve created by the monitoring node while the lower graph shows the seeder curve for the first 60 uploaders.

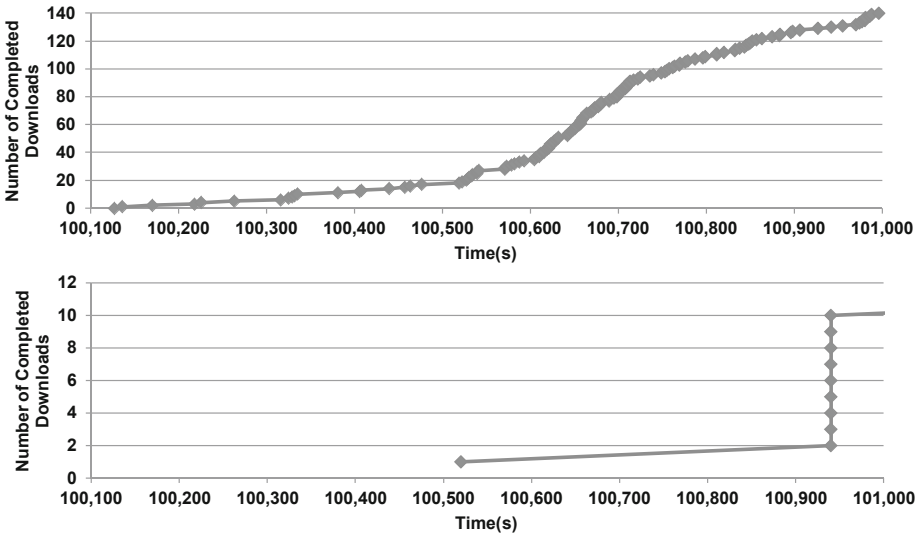

Figure 3. Summary of results from a Set 2 simulation.

Four simulations in Set 3 had a Type I error because the observed seeders were rejected by the validation rules. In this set, the uploader list collected by the monitoring node included the actual first uploader. The file existence verification function also returned the first uploader as the detected first uploader. However, the returned results did not pass the validation rules.

In most of the Set 3 simulations, the file existence verification function completed in a very short time. Therefore, the observed uploader was rejected by Rule 2. As a result, the first uploader in Set 3 was incorrectly rejected by the validation rules.

A review of the overall seeder curve in Figure 4 shows that the competition among downloaders might not have been sufficiently demanding. Thus, the download activities had a very short slow-rising period and quickly entered the rapid-rising period. This behavior is similar to that observed when sharing a small file.

The file existence verification function also executed at an early stage of the slow-rising period. The result is that the uploader may not be sharing the file requested by the other downloaders. Without competition for file content, a delay in the file request might not be observed.

Further analysis of the overall seeder curve and the corresponding download duration time shows that shortly after the appearance of three uploaders in the P2P file-sharing network, the download activities proceeded to the rapid-rising period. Therefore, the actual slow-rising pe-

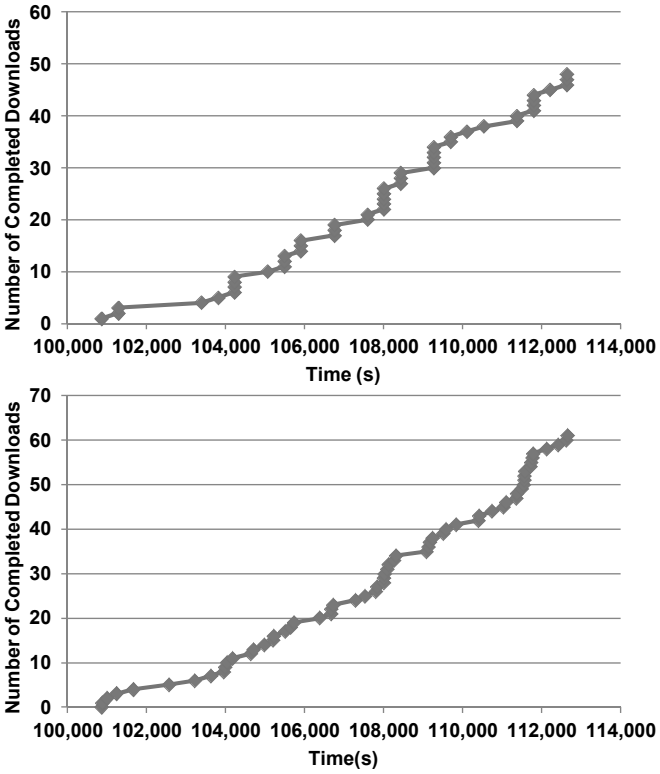

Figure 4. Summary of results from a Set 3 simulation.

riod in Experiment Set 3 was rather short. Even with the complete hub information, the probability of rejecting the first seeder was relatively high.

4.4 Set 4 Results

The Set 4 file size was small, which resulted in short file transfer times. Although the hub node serviced all the uploaders and downloaders, the time available for the monitoring node to submit the initial search query was insufficient. Figure 5 shows the seeder curve for Set 4. The query occurred after the slow-rising period and the search query response occurred at 100,400 seconds during the rapid-rising period. The observed first uploader was rejected by all three validation rules.

When the file size is smaller than 512 KB, the Gnutella2 protocol transmits the file in one fragment. Since the transfer time was so short, confirming that a file was originally from a particular uploader was impossible for the monitoring node even if the file locating function in-

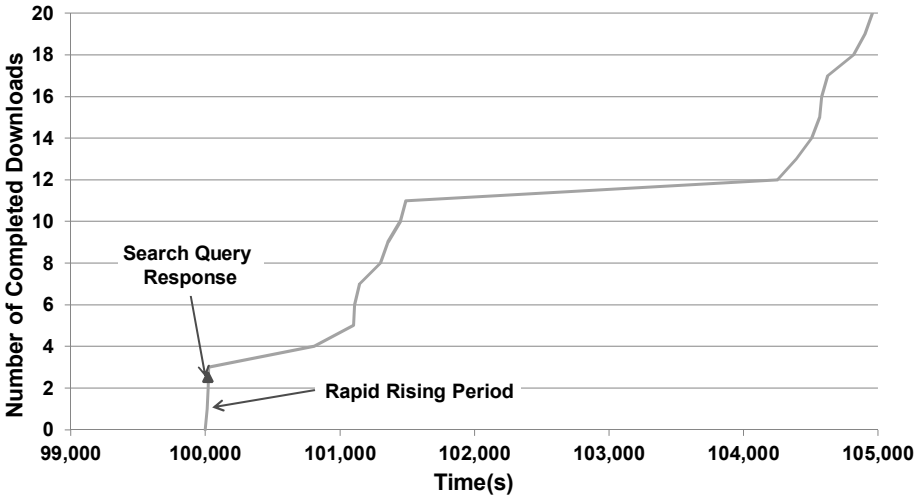

Figure 5. Seeder curve for Set 4.

cluded the first uploader in the search environment. This is due to two reasons. First, a node that successfully downloads one file fragment immediately becomes a full file content holder; the monitoring node then considers the node to be a potential first uploader. Second, if the file is popular, the file search, which can take between ten seconds to a few minutes, may not complete before the end of the slow-rising period. The result is that multiple file uploaders are identified in the network.

Based on these simulation results, it can be concluded that, when a small file is shared, a forensic investigator is unlikely to distinguish the first uploader even if the monitoring node was enabled since the beginning of the file download. However, the investigator can identify the participants, including the potential uploaders and downloaders.

5. Conclusions

The research presented in this paper enhances the first uploader detection strategy in Gnutella networks. The three proposed validation rules validated and correctly rejected all the incorrectly-identified first uploaders in the simulation experiments. However, the validation rules incorrectly rejected 80% of the actual first uploaders in Set 3 and 20% in Set 4, resulting in a Type I error of 25% over all the experiments. The error is due to the short slow-rising periods and the competition among downloaders.

The overall analysis suggests that, if the seeder curve has an extremely short slow-rising period or if the seeder is identified after the slow-rising

period, then the observed first seeder cannot be accurately validated. The validation rules can verify an observed seeder, but they eventually result in an incorrect rejection of the seeder.

Future research will investigate P2P network scenarios with short slow-rising periods to enhance the monitoring rules. Also, the research will analyze the characteristics of file-sharing activities and extract attributes that can accurately determine the slow-rising period.

References

[1] ckseol, More personal data identified by Foxy King (in Chinese), HKHeadline.com, Hong Kong, China (`blogcity.me/blog/reply_blog_express.asp?f=C6T2E3Z19P239366&id=566836`), August 18, 2013.

[2] Gnucleus, GnucDNA – Free P2P Core Component (`www.gnucleus.org/GnucDNA`).

[3] R. Ieong and K. Chow, Enhanced monitoring rule through direct node query for Foxy network investigation, presented at the *First International Conference on Digital Forensics and Investigation*, 2012.

[4] R. Ieong, K. Chow and P. Lai, Validation of rules used in Foxy peer-to-peer network investigations, in *Advances in Digital Forensics VIII*, G. Peterson and S. Shenoi (Eds.), Springer, Heidelberg, Germany, pp. 231–245, 2012.

[5] R. Ieong, P. Lai, K. Chow, M. Kwan, F. Law, H. Tse and K. Tse, Forensic investigation and analysis of peer-to-peer networks, in *Handbook of Research on Computational Forensics, Digital Crime and Investigation: Methods and Solutions*, C. Li (Ed.), IGI Global, Hershey, Pennsylvania, pp. 355–378, 2010.

[6] Sing Tao, Another personal data leak from HK Police Force identified by Foxy King (in Chinese), Toronto, Canada (`news.singtao.ca/toronto/2013-09-25/hongkong1380094959d4714768.html`), September 25, 2013.