Insights of File-Sharing System Forums

Guillaume Jourjon, Olivier Mehani and Thierry Rakotoarivelo
NICTA
13 Garden Street
Eveleigh, Sydney, NSW, Australia
Email: first.last@nicta.com.au

***Abstract*—One-click file hosting systems (1-CFHS) have become a prominent means to exchange files across the Internet. Studies have previously identified that a lot of the hosted content is infringing on its owner's copyright, and some of the most well know 1-CFHSs have been taken offline as a result of this. In this paper, we present a pilot study of how links to, and copies of, such content are exchanged via online forums. We have crawled and parsed pages from four of the most prominent sites over a period of a few months in order to extract URLs to these items. These URLs have then been periodically tested until they became unavailable in order to derive the lifespan of these copies on various 1-CFHS. We find that URLs are mostly posted once, presumably by their creators, and that unauthorised content on 1-CFHSs has an availability expectancy of about 40 days before being taken down. We propose an initial simple life-and-death model for such content in the form of a Markov chain. We also show that the 1-CFHS market is still unstable, with most of the past leader services having disappeared from the current charts.**

I. Introduction

File exchange has been one of the main uses of internets since early deployments. Starting with dial-up Bulletin Board Systems (BBSs) and FTP servers, communities of users sharing content have emerged using communications systems such as newsgroups or IRC to share information about the whereabouts of desired file copies. Distributed systems allowing large scale file exchange with little user intervention have appeared in the late 90s, with Napster being the precursor in peer-to-peer music search and exchange. BitTorrent [1] has now become widely used for content acquisition, supported by torrent indexing sites such as The Pirate Bay. More recently, there has been an increase in the use of one-click file hosting systems (1-CFHS) to store content, and user-run forums to share pointers to these files [2]. Nevertheless, what was once considered more of an underground community-based activity, with special chat rooms or newsgroups, has found a new breath and 1-CFHSs now represent around 22 % of current Internet HTTP traffic [2].

1-CFHS systems differ from previous file-sharing systems in their planned distribution and replication of content across several data centres and CDNs around the world. This architecture allows a better availability to the end user but raises some concerns for Internet services providers, as some internal routing might be affected by application-level load balancing implemented by 1-CFHS end-systems [3].

Another concern lies in the legality of the content being exchanged. There is strong indication that a large portion of the files hosted on these systems are unauthorised copies of copyrighted material [4]. The distribution of content was also found to be similar to that available on other systems such as BitTorrent [5]. Since the 2011 FBI injunction to pull the plug on the MegaUpload 1-CFHS and the following arrest of its founder in New Zealand [6], these services have received increased attention from copyright owners.

An important aspect of file sharing, beyond actual hosting, is the discoverability of content. While Napster and other early systems integrated search engines within their user front-end, more recent systems rely on external sources. Dedicated forums and chat rooms have always been one of the main channels to find any type of content for black-market information and data [7]. Nevertheless, the lack of access to their internal databases or logs makes their analysis difficult. Five different 1-CFHSs have been investigated from edge networks [5]; the authors passively analysed incoming and outgoing traffic to and from a university campus and analysed how content consumers were able to locate the desired copies. In [8], the nature of the created social network was also characterised for six different underground forums which main purpose was to exchange stolen personal information such as credit card number. The authors however had complete access to database dumps from these forums.

In this paper, we focus on the observed lifetime of unauthorised copies of copyrighted material uploaded to 1-CFHSs and shared on dedicated forums. Our work is based on two fundamental hypotheses. First, we conjecture information about such content hosted on 1-CFHS is primarily available from these dedicated forums, as supported by [3]. A second hypothesis is that such content eventually disappears mainly because the owners request it to be taken down for copyright reasons.

We therefore aim at developing a birth-and-death model for copyright-infringing content (and their re-spawned copies) hosted in 1-CFHS and publicly advertised through dedicated forums. For this purpose we have selected four well-known such forums based on their Alexa rank and

observed the dynamics of posted links [9]. We further use these links to assess the period during which each item is available on the 1-CFHS host before being taken down. We collected this data for about three months from September to November 2012, and accumulated more than 22,000 posts containing about 230,000 URLs.[1] Based on this dataset, we corroborate findings from previous studies [3], [5] but find that there has been dramatic changes in the landscape of 1-CFHS since their publication. We validate our hypothesis of finite content life expectancy, with more than 90 % of our filtered URL data eventually found dead, and propose an initial Markov-chain model of the lifecycle of 1-CFHS content.

The remainder of this article is organised as follows. Section II presents an overview of the related work concerning file-sharing systems and forum investigations. Section III introduces the method we have used in this study while Section IV presents our results. We then discuss the potential implications of these results in Section V. Finally, we conclude this study in Section VI.

II. Related Work

In this section, we first describe the general architecture of 1-CFHS and review some recent related studies. We then discuss the methodology of previous studies of online forum based social networks, which informed our data collection process.

A. 1-Click File Hosting Systems

The 1-CFHS ecosystem comprises six main actors: content uploaders ([5] calls them "publishers"), file-hosting services, source sites, index sites, content consumers, and content owners. This is illustrated in Fig. 1, which we extended from [5]. Content uploaders send content copies to file-hosting services, and retrieve an URL for each file. These URLs are then published on source sites, such as public forums or personal blogs, and are eventually crawled by third-party index sites (*i.e.*, search engines). The two last actors, the content consumers and owners can then use source and index sites to search for a particular resource. Once the resource is found, these actors behave differently. Consumers will start downloading the item while owners will request the 1-CFHS to take it offline.

Based on data captured in edge networks, [5] studied the dynamics of interactions around 1-CFHS systems in the context of a university campus. The authors focused on five popular hosting services, namely, RapidShare, Megaupload, zShare, MediaFire and Hotfile.[2] The dataset consists of a trace of HTTP transactions collected over one year. This data was used to study how content consumers find and select items to download. The authors also hypothesised that links in forums were created by content uploaders themselves.

[1]This dataset is available at http://nicta.info/fsstudy#Datasets.

[2]https://rapidshare.com/, http://www.megaupload.com/, http://www2.zshares.net/, http://www.mediafire.com/ and http://hotfile.com/, respectively.

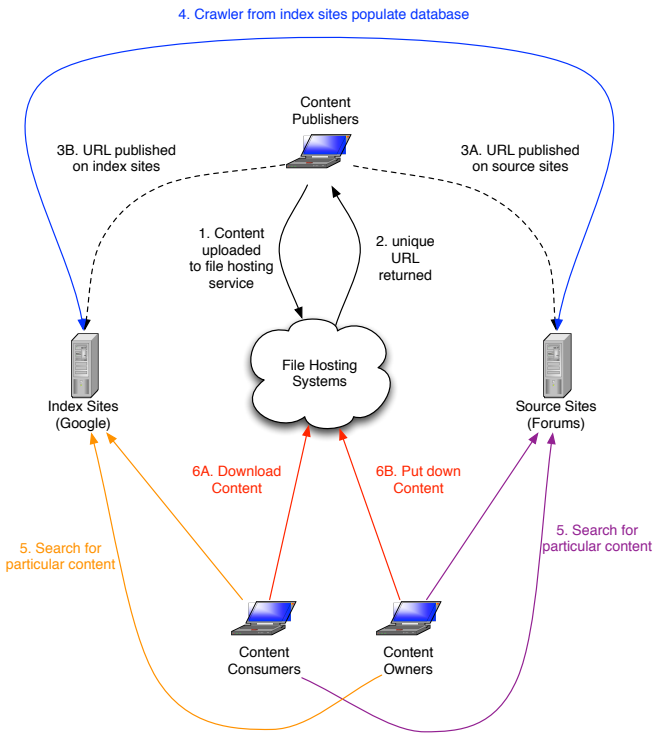

Fig. 1. Dynamic interactions around 1-CFHSs (adapted from [5]).

Following this work, the authors further performed a complete longitudinal study of the content available in 1-CFHS [3]. They found the items shared on these systems were replicated across several services. They also noted that, as compared to P2P services, content would appear earlier on 1-CFHS.

Finally, [4] studied the legal nature of content hosted on 1-CFHSs. The authors found that these systems were largely used to disseminate illegal content, with up to 79 % of the content identified as unauthorised copies of entertainment or software products.

Our goal is to complement this body of work by modelling the birth and death rates of content shared through forums responsible for a large proportion of the traffic to 1-CFHS, as noted in [3]. This model is based on the hypothesis that content on 1-CFHS mainly disappears because copyright owners also scrutinise these forums in order to identify which content to ask the deletion of.

B. Social Networks on Online Forums

As mentioned in the previous section, online forums are used to disseminate links to 1-CFHS-hosted content, as source sites. Common approaches study these communities as online social networks (OSNs), and attempt to characterise the nature and dynamics of people social associations. For example, based on the user interaction on Facebook, researchers have proposed to use interaction graph to represent their relationships [10].

Although researchers can evaluate the properties of a social graph, it is difficult to transpose OSN metrics to underground forums. One of the main reasons is that the data of underground forums is often not available as a whole. For example some forums might be accessible by invitation only. Therefore, datasets can only be obtained through inside jobs or external intervention [8].

Another reason why social graph metrics may not be as relevant is that the relationships between users are not fixed (as opposed to a friendship link). Users of underground forums may not have consistent relationship with each other. As a result, many conversations may occur during a single transaction, and less or no communication log can be found after this trade.

Based on a full data dump, [8] characterised the nature of the OSNs formed in six underground forums. In this study, the authors found that commerce forums used mainly private messages to deal with the transactions while a reputation system was generally used to identify a user's elevation in the group. A higher reputation usually resulted from more public activity. In this study, the graph of social degree is also used to indicate how members link and interact with each other. The authors found that this social degree was correlated to the type of underground forums (*e.g.*, carder, black or white hat).

These social graphs have recently gained more attention with the publication of several of them in relation to Facebook or Google and are under discussion for a better overall integration [11]–[13].

Our analysis in this pilot study does not focus on the social network aspect of 1-CFHS forums. However, in our future work we will analyse the social information contained in our collected dataset and investigate the related social networks.

III. Method

As described in Section II and [5], the main vessels through which URL pointers reach content consumers is through both source and index sites (*i.e.*, forums and search engines). In this study we focus solely on a selection of the former which we consider to be representative of the publicly accessible forums. This selection was based on their respective Alexa rank [9] at the beginning of the study (October 2012), and is summarised in TABLE I.

TABLE I
Selected Forums

Name	URL	Alexa Rank
RLSLOG	http://www.rlslog.net/	1,809
Avaxhome	http://avaxhome.ws/	2,122
Sceper	http://sceper.ws/	10,937
SceneSource	http://www.scnsrc.me/	15,708

All these forums follow the same model where a poster starts a thread on a specific category and topic (*e.g.*, a movie or an episode from a series) with some initial 1-CFHS URLs, and followup posts to this thread contain additional URLs pointing to copies or other releases of the same content. All forum users can create threads or post replies.

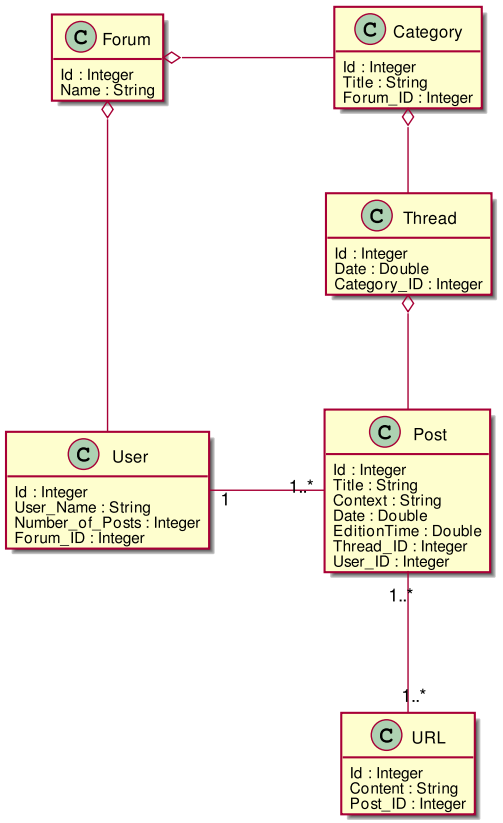

Fig. 2. Relational representation interactions in 1-CFHS-sharing forums.

As we are interested in the birth and death process of files hosted on 1-CFHS, we have developed a crawling and URL-validation system populating a database with the schema shown in Fig. 2. The remainder of this section describes our data collection system and campaign, and discusses the variables used in our analysis.

A. Forum-crawling and URL-validation System

We developed a crawler which operates as illustrated in Fig. 3. It first checks the forums under study for new threads, and creates new entries for these in the measurement database.It also updates the record of already visited threads with new added posts, if any. This crawler is called once per hour per forum. Based on this collection of raw page captures, the database is further populated by parsing the HTML content and extracting URLs from each thread, along with metadata from their containing post.

The URL verifier in Fig. 3 is in charge of sequentially resolving and testing the availability of the content pointed by the URLs in the database. Indeed as mentioned previously, 1-CFHSs are expected to remove such content after a while due to copyright infringement.

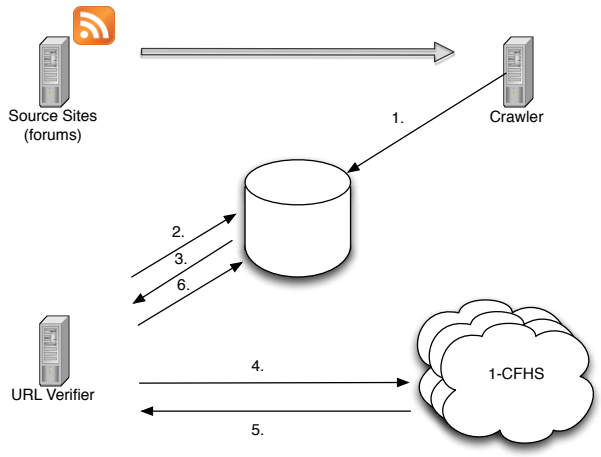

Fig. 3. Method in place during our study. Source sites were periodically crawled in search for new URLs, which were then checked sequentially and repeatedly by a URL verifier.

In order to perform this verification, we have modified an open-source downloader application, JDownloader,[3] to support injection of new URLs through its web interface. The choice of JDownloader as the base for our URL verifier stems from the fact that it can extract metadata such as file name, size and availability for more than 2,000 1-CFHSs. It is also natively able to resolve shortened or indirect URLs, which constitute a large portion of the posted links, into final unique links to the actual content.

For each new URLs, the verifier checks its content availability and extracts its file metadata. This information is then updated in the database. Each URL in the database is periodically checked for availability until its linked content has been removed. We are thus able to approximate the life duration of every file uploaded in the various 1-CFHS. As opposed to the crawler, the URL verifier was run continuously on the collected URLs, looping to the first still-alive entry when the last one had been checked.

We have observed the different forums listed in TABLE I for 3 months from September 2012 to November 2012. We have limited the URLs to the Music and Video categories of the various forums. Overall we have collected more than 22,000 posts from the forums resulting in about 1/4 million URLs for which the birth and death times were observed.

B. Observed Dependent Variables

We defined and measured the following dependent variables to characterise the birth-and-death process of (illegal) content announced on Internet forums and hosted on 1-CFHS.

1) Popularity of 1-CFHS Systems: Given a number of URLs pointing to content hosted on various 1-CFHSs, we can infer the popularity of each of these systems by ranking them based on the number of items hosted and shared.

[3]http://www.jdownloader.org/

2) URL Duplication: Links to a specific piece of content, as identified by its unique, fully-resolved URL, might be posted more than once. We study the duplication of these URLs across all threads, posts and forums.

3) Post Size: More than one URL is often given in each post. It might therefore be interesting to study the distribution of post sizes.

4) Content Lifecycle: To study our main concern, we have collected the times of birth B_i (as given by the date of the earliest post pointing to that URL) and observed death D_i (taken as the first time the verifier found the content to be unavailable) for every URL i. We can then define the lifespan L_i of each content i as

$$L_i = D_i - B_i. \tag{1}$$

C. Independent Variables

To explain variations in our observations we consider two kinds of factors: those related to forums where URLs are published and those related to the 1-CFHS where content is hosted. TABLE II presents a summary of these independent variables.

TABLE II
EXPLANATORY FACTORS

Name	Possible Value
URL Provenance	Forum ID
Type	Category ID
File-sharing system	1-CFHS ID

1) Forum-related Variables: We consider two variables related to forums: the forum's name, and by extension, its Alexa rank, and the thread category, as shown in Fig. 2.

2) 1-CFHS-related Variables: Our modification of JDownloader allowed access to additional information such as the 1-CFHS ID (through resolution of short URLs into final unique URLs), the real name of the file, or its size. In this preliminary study, we decided to only consider the hosting identification (1-CFHS ID).

IV. RESULTS

We have collected 231,679 URLs, in the first step of Fig. 3, which have been later filtered according to several factors. TABLE III shows the number of URLs before and after filtering broken down depending on their status. The filtering process mainly consists in removing URLs of which the associated content could not be verified and content from unpopular 1-CFHS. The following section breaks down the details of this filtering process and its parameters. These filtered URLs will be later used in Section IV-C to study the dynamics of creation and destruction of content on 1-CFHS hosts while the general consideration analysis will take advantage of the full dataset regardless of the content's status or provenance.

TABLE III
Raw URL Measurement Data

Name	Alive	Dead	No status	Total
Original Dataset	11,272	144,943	75,464	231,679
Filtered URLs	11,237	142,490	0	153,727

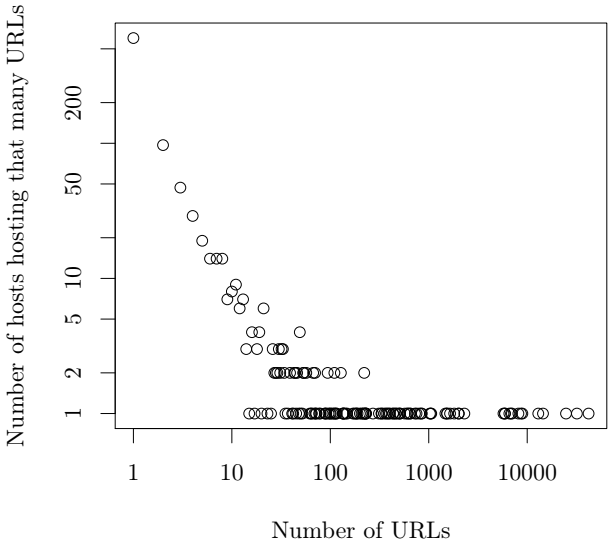

Fig. 4. Distribution of content across 1-CFHS hosts in our dataset.

A. URL Filtering Criteria

The first factor is the URL's status. In our study, three statuses were assigned to these URLs: *none*, *alive* and *dead*. When a URL is initially scraped from a forum post (step 1 in Fig. 3), its status is set to none. The URL verifier then connects to that URL and checks its status. If the content is still available, the URL status is set to alive. This status remains until the content disappears, when it is set to dead, and the observed time of death is recorded. In our filtered dataset, we only retained URLs which were either alive or dead, discarding any other. This reduced the number of URLs from about 231 k to 156 k.

A second factor is the 1-CFHS system hosting each piece of content. As shown in Fig. 4, a few servers host most of the content (around 10,000 items) while most of the others only harbour very little content. We filtered the present dataset to remove hosts pointed to by less than 50 URLs, as they were deemed not representative of the service itself. This reduced the number of considered 1-CFHS hosts from 1,047 to 108. The total number of URLs was however only further reduced to 153,k.

B. General Considerations

In this section, we study the relationship between the crawled forums and the 1-CFHS services. To this end we use the full dataset summarised in TABLE III.

1) Overview of 1-CFHSs: As mentioned in the previous section, after filtering the number of 1-CFHS hosts, about a hundred of them remained. TABLE IV shows the top-20 hosts. They account for more than 86 % of the crawled URLs (including those with no status).

TABLE IV
Top-20 1-CFHS

Name	Number of URLs
Uploaded	42,035
RapidGator	31,971
FilePost	24,796
TurboBit	14,469
RapidShare	12,960
BitShare	8,923
DepositFiles	8,650
NetLoad	8,137
ExtaBit	6,994
UltraMegabit	6,846
SharpFile	6,588
LetitBit	5,953
FreakShare	5,920
ShareFlare	5,741
FileFactory	2,299
Vip-File	2,024
HenchFile	2,002
NewTorrents	1,814
LumFile	1,646
TorrentDownload	1,536

Interestingly, the top-ranked hosts are not the ones which were considered in previous 1-CFHS studies [3]–[5]. The first considered system appearing in our list, RapidShare, is only found to be the 5$^{\text{th}}$ largest host. We discuss the implications of this finding in Section V.

2) URL Duplication: Another factor of interest is the duplication of URLs. While initial content uploaders share their 1-CFHS URLs with their forums of choice, some other content consumers visiting them might repost the links to other threads or forums. Fig. 5 shows how often URLs are found in more than one post. Most URLs (168,081) are posted only once across all forums, which represents 72% of the entire dataset. However, the most duplicated URL had up to 940 reposts but appeared to be spam.[4] This specific URL was pointing to a dead page at the time we manually checked it.

3) URL and Post Relations: Our dataset lists 21,985 posts (*i.e.*, a single message within a thread, including the initial one). As shown in Fig. 6, posts contain a variable number of URLs. 80 % (17,507) of posts contains less than 11 URLs, while a few outliers list up to 1,466 URLs, which is probably an indication that they are spam posts. The highest number of posts contains two URLs, which we hypothesise is due to either file-size limitations on the 1-CFHS requiring files to be split and uploaded separately, or items distributed across a few 1-CFHS by their uploader, as mentioned in [3].

[4]A URL which does not lead to the content stated by the original post but rather to advertisement or other unrelated material.

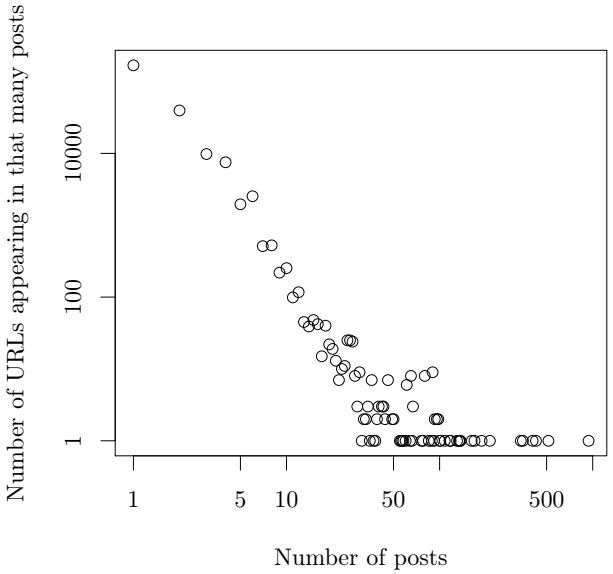

Fig. 5. Distribution of duplicated URLs

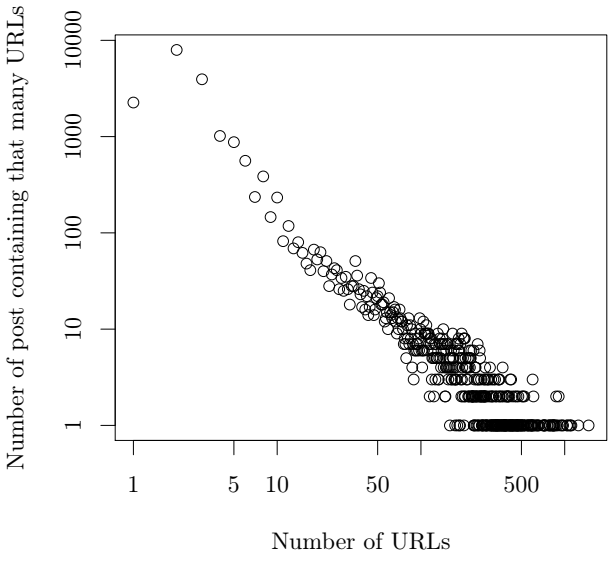

Fig. 6. Distribution of the number of URLs in each post.

4) Data Collection Process: During the data collection process of this pilot study, we identified some issues that we will address in the refined design of our future long term study. For example, we ended up collecting many more URLs than what we expected, which created a bottleneck at the verifier. Indeed, it took about 5 s for it to resolve one URL, and therefore created a backlog of increasing length at this step. This verification duration was considered as the observed worst case scenario for the resolution of short URL referencing content on 1-CFHS. One approach to address this issue would be to design a measurement process which will be less greedy in collecting new posts, and which will rather maintain a manageable sample set from each forum potentially through a tighter coupling with the URL verifier (*e.g.*, collect new posts from a forum, only when enough of its URLs have been found as dead).

C. Lifecycle Analysis

In the lifecycle analysis, we only consider our filtered dataset, *i.e.*, only those URLs our verifier has seen alive, then eventually dead. This class represents most of our dataset, with 142 kURLs, or 61 % of the entire unfiltered set and 93 % of the filtered data. We studied the birth and observed-death times of the URLs, and estimated their lifespan. The characteristics of their distribution is shown in TABLE V, though we note neither followed a normal law ($p < 22 \times 10^{-16}$ in all cases).

TABLE V
Lifecycle Analysis Summary

Variable	**Mean**	**Standard Dev.**
Inter-birth time	46.84 s	1,509.84 s
Inter-death time	5.68 s	347.87 s
Lifetime	41.85 d	12.74 d

1) Birth and Death Rates: We first propose to quantify the birth and death phenomenon. These results allow us to propose a comprehensive model later.

The birth of the content was computed and compiled based on the date of the post in every forum, thus assuming that uploaders would almost immediately publish URLs upon successful completion of the file transfer to the 1-CFHS. Its death was taken to be the first time when the content referenced by the URL was found removed by the verifier.

2) Life Expectancy: The graph on Fig. 7 shows the life expectancy of URLs. It is computed as the difference between its birth time and its death time (1).

3) Markov Chain Model: TABLE V shows the observed birth and death rates, as well as life expectancy of the content published on 1-CFHS. Based on these measurements, we propose a birth-and-death process which would allow to predict the number of items in these systems. Such prediction may be used to dimension storage resources in 1-CFHS data centres. This model is presented in Fig. 8 where $\{\lambda_i\}_{i=0,1\ldots\infty}$ and $\{\mu_i\}_{i=0,1\ldots\infty}$ represent the birth

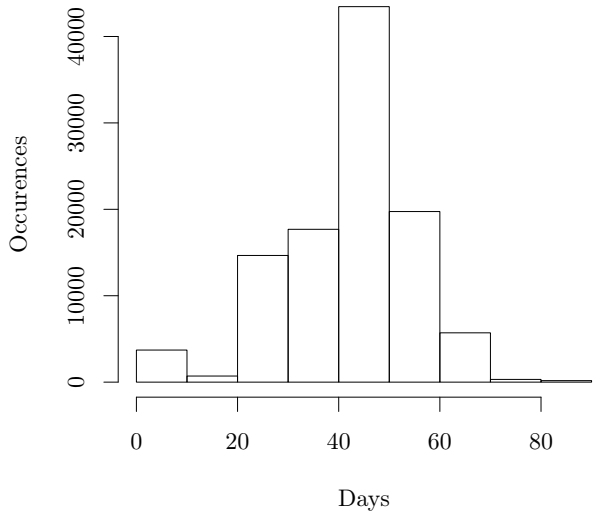

Fig. 7. Distribution of URLs' life expectancy.

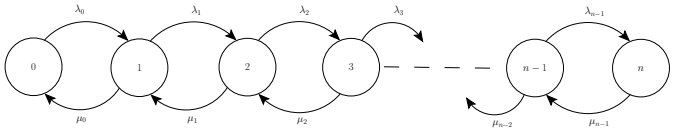

Fig. 8. Markov chain model of the number of items hosted by one 1-CFHS.

and death rates, respectively, according to the state of the system. In our model, we identify the state i as the cardinality of the set of contents.

Based on this model, we can extract the expected sojourn time for every state of the Markov chain following classical analysis methods for such system. Unfortunately, due to the peculiar characteristics of the results presented in previous section, presumably due to the bottleneck formed at the verifier during the data collection, we are unable to propose a full analytical solution to this problem at this stage. We are however currently investigating its resolution, which will be the subject of future work.

V. Discussion

The results presented in Section IV, while confirming previous studies, open some interesting questions for the future and could allow us to better define in the future a comprehensive model using the concept of birth-and-death processes briefly introduced in Section IV-C.

One of the questions raised by [3], [5] was to determine whether or not URLs were posted across different forums and whether these possible multiple posts were done by the content uploader or by a third person. While these studies clearly identified how end-users find content, they did not provide much insights on the content uploaders, apart from the assumption that they may duplicate their posted items across different 1-CFHS. Our results in Fig. 5 show that URLs are rarely duplicated across posts and forums, suggesting that only content uploaders post links after uploading items to a new 1-CFHS. Link posters also appear to limit themselves to only one forum. Nevertheless, some URLs were duplicated more than 50 times (Fig. 5), and some posts were found to contain more than 1,000 URLs (Fig. 6), which raises questions about spam levels in these forums.

In the year and a half since [3]–[5] the landscape of 1-CFHSs providers has drastically changed, with only RapidShare surviving in the top-20 shown in TABLE IV, at the exception of all others (particularly, MegaUpload, FileServe or FileSonic). It can be hypothesised that RapidShare managed to survive by removing their term-of-service clause favouring big uploaders, while, *e.g.*, MegaUpload did not [6]. While this shift does not impact findings in previous works *per se*, it indicates that the 1-CFHS market is still in a transient phase, with no clear leader.

The main finding of this work is in confirming our hypothesis that content likely to infringe on some copyright has a finite lifetime. Indeed, as shown in TABLE III, most URLs (more than 90%) were eventually found dead. Fig. 7 and TABLE V further characterise this lifespan, which has an average of 40 days. We keep for future work the investigation of the impact of the independent variables, which were presented in Section III-C (*e.g.*, thread category, forum ID).

While this study supports our initial hypothesis and confirms the claims in previous works, it also raises several questions regarding both the communities created on these forums, and the fine line for 1-CFHS between popularity in hosting illegal content and compliance with copyright laws. Future work will measure these aspects using the experience acquired in this pilot study. More specifically, we want to study the interaction between users within and across forums using methods developed for OSNs as described in Section II. Orthogonally, the rate at which clones are created in 1-CFHS and their URLs posted on these forums needs to be studied more in depth, taking into account the death of previous clones, and the popularity of the content, as evaluated by audience ratings and thus continuing the work presented in [3].

VI. Concluding Remarks

In this paper we have presented our initial work on assessing the on-going trend and interactions between 1-CFHS and their companion source sites (forums). We have presented our methodology based on a periodic crawler and URL verifier, and identified variables of interest, both dependent and independent. Analysis of the dependent variables confirmed the findings of previous studies, and allowed us to build an initial life-and-death Markov model

for 1-CFHS content. Overall, 90 % of URLs were found to eventually have died, on average 40 days after their birth. This confirms the hypothesis that they have a finite lifetime, which is presumably due to the content-publishing industry scrutinising these forums and sending takedown requests.

This work has highlighted the fidelity of content uploaders/posters to only one forum, whereas [5] showed that consumers where much more volatile. We also observed a variability of the 1-CFHS landscape, with favourites from the past years having almost entirely disappeared from the top-20.

Future work will incorporate lessons learned from the challenges encountered by our crawling system. It will also investigate the probability distributions of birth and death rates more thoroughly and incorporate them into a more elaborate iteration of the presented Markov model.

References

[1] B. Cohen, "The BitTorrent protocol specification," BitTorrent, Tech. Rep. 3, Jun. 2009.

[2] J. Sanjuàs-Cuxart, P. Barlet-Ros, and J. Solé-Pareta, "Measurement based analysis of one-click file hosting services," *Journal of Network and Systems Management*, vol. 20, no. 2, pp. 276–301, 2012.

[3] A. Mahanti, N. Carlsson, and C. Williamson, "Characterizing the global file hosting landscape," in *MASCOTS 2012*, Aug. 2012.

[4] T. Lauinger, K. Onarlioglu, A. Chaabane, E. Kirda, W. Robertson, and M. A. Kaafar, "Holiday pictures or blockbuster movies? copyright infringement in user uploads to one-click file hosters," in *RAID 2013*, 2013.

[5] A. Mahanti, C. Williamson, N. Carlsson, M. Arlitt, and A. Mahanti, "Characterizing the file hosting ecosystem: A view from the edge," *Performance Evaluation (IFIP Performance 2011 Special Issue)*, vol. 68, no. 11, pp. 1085–1102, Oct. 2011.

[6] U. S. D. court for Eastern District of Virginia, "United States of America vs. Kim Dotcom," 2012.

[7] J. Franklin, V. Paxson, A. Perrig, and S. Savage, "An inquiry into the nature and causes of the wealth of internet miscreants," in *CCS 2007*, 2007.

[8] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker, "An analysis of underground forums," in *IMC 2011*, 2011.

[9] Alexa — the web information company. [Online]. Available: http://www.alexa.com/

[10] C. Wilson, B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao, "User interactions in social networks and their implications," in *Eurosys 2009*, Apr. 2009.

[11] T. Berners-Lee. (2007, Nov.) Giant global graph. Accessed 2013-09-20. [Online]. Available: http://dig.csail.mit.edu/breadcrumbs/node/215

[12] B. Fitzpatrick. (2007, Aug.) Thoughts on the social graph. Accessed 2013-09-20. [Online]. Available: http://bradfitz.com/social-graph-problem/

[13] A. Iskold. (2007) Social graph: Concepts and issues. Accessed 2013-09-20. [Online]. Available: http://readwrite.com/2007/09/11/social_graph_concepts_and_issues