# Measurement and Analysis of Child Pornography Trafficking on Gnutella and eMule

Ryan Hurley, Swagatika Prusty, Hamed Soroush, Robert J. Walls
Jeannie Albrecht†, Emmanuel Cecchet, Brian Neil Levine, Marc Liberatore, Brian Lynn
Department of Computer Science, University of Massachusetts, Amherst
†Department of Computer Science, Williams College

## ABSTRACT

Peer-to-peer networks are the most popular mechanism for the criminal acquisition and distribution of child pornography (CP). In this paper, we examine observations of peers sharing known CP on the Gnutella and eMule networks, which were collected by law enforcement using forensic tools that we developed. We characterize a year's worth of network activity and evaluate different strategies for prioritizing investigators' limited resources.

First, we focus on strategies for reducing the number of CP files available on the network by removing a minimal number of peers. We present a metric for peer removal that is more effective than simply selecting peers with the largest libraries or the most days online. We show that any successful strategy must target offenders from all countries. Second, we characterize the aggressiveness of six peer subgroups, including: peers using Tor, peers that bridge multiple p2p networks, and the top 10% of peers contributing to file availability. We find that these subgroups are more aggressive in their trafficking, having more known CP and more uptime, than the average peer. Finally, while in theory Tor presents a challenge to investigators, we observe that in practice offenders use Tor inconsistently. Over 90% of regular Tor users send traffic from a non-Tor IP at least once after first using Tor.

## 1. INTRODUCTION

Peer-to-peer (p2p) networks are the most popular mechanism for the criminal acquisition and distribution of *child sexual exploitation imagery*, commonly known as *child pornography* (CP)[1]. Investigating CP trafficking online is critical to law enforcement because it is the only effective *proactive*[2] method of finding persons, known as *contact offenders*, who directly and physically abuse children. A previous study has shown that 16% of investigations of CP possession ended with discovering contact offenders [31]. These investigations have two primary goals: to stop the distribution of CP; and to catch child molesters and help children that are being victimized, often by family members. Understanding the behavior of users of these networks, and of the network as a whole, is of critical importance to investigators.

We have built several tools for conducting forensically valid investigation of these crimes. These tools are in daily use by over 2,000 U.S. law enforcement officers. As a result, thousands of persons, many of whom had directly abused children, have been arrested for these crimes. Additionally, our tools log an arsenal of data about these online activities, permitting an in-depth analysis of the use of these networks. In this paper, we examine CP trafficking using data gathered mainly from the Gnutella and eMule p2p file sharing networks.

Our study differs significantly from past work in both duration and detail. Most previous studies of p2p networks have taken place over just several days [8, 9], several weeks [12], or a few months [7, 17, 21, 26]; our study is comprised of many thousands of observations per day for a full year. Past measurements focused on online CP trafficking have examined only observed search terms related to CP [8, 9, 25]; our study has information about peers sharing files verified as child pornography and follows strict forensic standards of measurement.

We present a detailed analysis of the one-year period of observations of Gnutella and eMule. We examine methods of proactively reducing file availability and present characterizations of subgroups of users. Our focus is on *files of interest* (FOI), which include CP images, as well as stories, child erotica, and other collections that are strongly associated with this crime. Our high-level findings are as follows.

- CP trafficking is global in scope. We observed over 1.8 million distinct peers, from over 100 countries sharing CP on eMule, and over 700,000 peers on Gnutella.
- We observed that the majority of CP files are shared by a relatively small set of aggressive users. These users are geographically diverse and thus require international cooperation to be apprehended.

[1]These are not "sexting" crimes by late teens: 21% of CP possessors have images depicting sexual violence to children such as bondage, rape, and torture; 28% have images of children younger than 3 years old [31].

[2]This method is proactive in the sense that law enforcement is not waiting for children or others to come forward and report abuse.

- While most FOI are only available for a short amount of time (only about 30% are available for more than 10 days of the year), there are still tens of thousands of unique FOI available for download each day. On Gnutella alone, the daily average is over 9,700 unique FOI.

In addition to the general findings above, our key contributions are as follows.

- We propose and evaluate three strategies for prioritizing law enforcement resources in investigating CP trafficking. We conclude that removing peers with largest *contributions* (a weighted measure of days of uptime and files made available) is most effective, but with Pareto-like diminishing returns.
- We examine subgroups of peers of particular interest, such as peers seen using Tor, peers on multiple p2p networks, and four other subgroups. We find that all appear to be more aggressive in their trafficking, having more FOI and more uptime than the average peer.
- We find that offenders using Tor use it inconsistently. Over 60% of linkable user sessions send traffic from non-Tor IPs at least once after first using Tor, thus removing its protection; over 90% of sessions observed on three or more days fail likewise.

Further, given that peers are not one-to-one with users we examine our dataset for evidence of aliasing. We find little evidence to suggest that users are changing their application-level identifiers but keeping their libraries.

Our findings are based on a partnership with over 2,000 U.S. law enforcement officers nationwide to collect data on CP trafficking, all specifically trained on tools that we implemented for this purpose. Before presenting these results in more detail, we describe our measurement methodology.

## 2. FORENSIC MEASUREMENT

This study is based upon the analysis of a large number of observations of CP files on p2p networks, and the behavior of the peers that share them. In this section, we describe the sources of these datasets and provide salient details relevant to our analysis. In Section 5, we identify sources of bias in the data and potential limitations of our study.

**Background.** Unlike any other academic work we are aware of on p2p networks or Internet crime, this paper is based on data collected with the help of national and international law enforcement[3]. Starting in January 2009, we began deploying a set of forensic tools to investigators in the U.S. and internationally for online investigation of p2p CP trafficking.

Prior to our collaborative efforts, the standard method for online CP investigation was to make isolated cases: Leads were not shared among agencies or officers, other than by phone or email. Officers leveraged their own experience to prioritize suspects.

**Tools.** Our suite of tools called *RoundUp* [14], has enabled seamless sharing of *plain view*[4] observations of online CP and associated activities on various filesharing networks. The shared data, collected in order to make these cases, provide each investigator with a longitudinal view of CP offenders and provide a method of triage for selecting targets for further investigation; and of course, the data enable this study. Because over 2,000 investigators have been trained on our tool to date, and because it is in use by hundreds of investigators daily, the aggregate set of observations we have used for this study is incredibly detailed. The tools are still in use, and currently, law enforcement execute approximately 150 search warrants nationwide per month based on data collected using our tools. We do not present search warrant or arrest data in this study[5].

**Datasets.** Our datasets, summarized in Table 1, include law enforcement observations from Gnutella and eMule p2p networks. The Gnutella and eMule datasets span a one-year period from October 1, 2010 to September 30, 2011. Each record in these datasets corresponds to a law enforcement observation of a particular peer making available one or more FOI, and minimally contains date, time, IP address, application-level identifier, geographic location as determined by an IP geolocation database, and a file hash.

Most file sharing protocols include an application-level identifier unique to an installation of the application. In both Gnutella and eMule, these identifiers are persistent across users' sessions, and are referred to as *GUIDs* (globally unique identifiers). Peers on these networks are uniquely identified by their GUID, and we use peer and GUID interchangeably to identify unique running instances of the corresponding p2p software.

All FOI are uniquely identified using hash values; law enforcement manually classify files as FOI from a variety of sources. An enormous number of such FOI are shared on Gnutella and eMule. Respectively, there

[3]We detail lessons learned from working with law enforcement and the basics of digital forensics research elsewhere [30].

[4]Criminal procedure in the U.S., rooted in the Fourth Amendment, requires that investigations of p2p networks collect only *plain view* data ahead of a search warrant. Briefly stated, all data on a p2p network is in plain view because the source of the data has given up any reasonable expectation of privacy; see *U.S. v. Gabel*, 2010 WL 3927697 as a recent case that discusses plain view data in the p2p context.

[5]Our study's procedures were approved by our IRB. Rules for this project do not allow direct data collection on people who were arrested.

| Network | Date Range | Files | GUIDs | Observations |
|---|---|---|---|---|
| Gnutella (FOI only) | 10/1/2010 – 9/30/2011 | 139,604 | 775,941 | 870,134,671 |
| Gnutella Browse | 6/1/2009 – 9/30/2011 | 87,506,518 | 570,206 | 434,849,112 |
| eMule (FOI only) | 10/1/2010 – 9/30/2011 | 29,458 | 1,895,804 | 133,925,130 |
| IRC (no file data) | 6/2/2011 – 9/18/2011 | N/A | N/A | 7,272,739 |
| Ares (no file data) | 5/31/2011 – 9/18/2011 | N/A | N/A | 17,706,744 |

**Table 1: All datasets are observations of CP activity only, but IRC and Ares data do not contain information about files or GUIDs.**

are 139,604 and 29,458 known FOI shared by 775,941 and 1,895,804 GUIDs. Our tool searched only for FOI in a list containing about 384,000 entries; this list was updated several times over the course of this study. It is a small sample: the National Center for Missing and Exploited Children has reviewed more than 60 million child pornography images and videos[6].

In a limited fashion, we use two other datasets. Our IRC dataset, based on a more recent tool that we developed, covers a four-month period from June to September 2011. The IRC dataset is a log of IP addresses that were involved in public activity related to the sexual exploitation of children in public chatrooms; no file observations are in this dataset. We also use a dataset of CP-related activity on the Ares p2p network[7] collected using a tool we did not write, but collected by the same law enforcement officers responsible for all data in this paper. The Ares dataset contains only IP addresses and has no information about files shared, but addresses were only logged for peers that shared known FOI.

**Other Details.** Gnutella allows a peer to be *browsed* and thus investigators can enumerate all files shared by peers. Our *Gnutella Browse* dataset consists entirely of peer browses and includes all files a peer is sharing, not just FOI. Some Gnutella peers cannot be browsed; and we only collected FOI data from these peers. eMule does not permit browses to occur. Regardless, each of these datasets includes only peers that share one or more FOI; peers without FOI are not logged.

We draw a distinction between a time-limited view of a peer's shared files and the set of all files that a given peer was ever observed with. We define a GUID's *library* to be the set of files that were observed being shared by that GUID on a given day. A GUID's *corpus* is the set of all files shared by that GUID over the entire duration of the study. In both cases, we typically only include FOI, but we make it clear when a corpus or library includes non-FOI observed as the result of a browse.

**Limitations.** We discuss limitations and biases of our measurement process in Section 5.

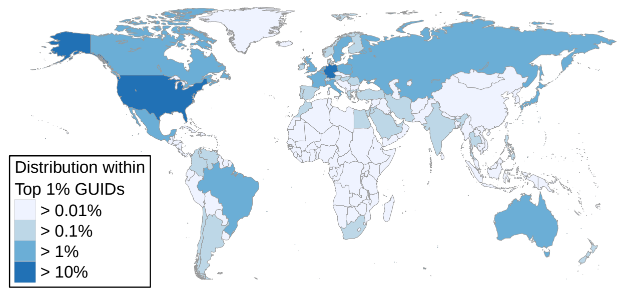



**Figure 1: CP traffickers are spread around the world. This logscale heat map shows the countries of the top 1% of Gnutella GUIDs by number of CP-related files shared over the year-long period of this study. A GUID, or globally unique identifier, corresponds to a unique installation of a Gnutella client. Our dataset contains a total of 775,941 such GUIDs.**

# 3. AVAILABILITY AND RESILIENCE

In investigating the trafficking of CP on p2p networks, the goal of law enforcement is to prioritize criminals whose arrest will have the greatest impact. But the strategy to achieve this goal depends upon the impact desired: Finding contact offenders who go otherwise unreported, finding those who create new CP, and decreasing the availability of FOI on the network are all priorities. In this section, we focus on strategies for reducing the availability of FOI.

Effective CP removal strategies are especially important as a means to prioritize law enforcement's limited resources and time. After online evidence is collected, many weeks of off-line processes are required at a minimum in each case until an arrest is made. Additional resources are required to go to trial. It is infeasible for investigators to arrest all users sharing CP and remove all FOI. Investigators need a triage strategy for deciding which small fraction of online leads to act upon.

An enormous set of perpetrators are active every day around the world. Figure 1 illustrates the geographic diversity of the top 1% of offenders on one p2p network. Even with unlimited resources, U.S. law enforcement can only partially impact file availability. Our results, discussed below, suggest the need for a coordinated international effort.

## 3.1 FOI Redundancy and Availability

Before we further discuss the implications of removing

[6]See `http://www.missingkids.com/missingkids/servlet/NewsEventServlet?&PageId=4604`.
[7]`http://aresgalaxy.sourceforge.net/`

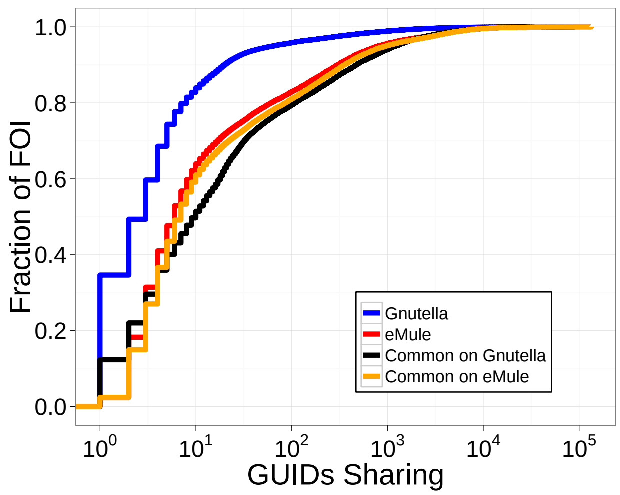



**Figure 2: Redundancy of FOI (files of interest) among multiple GUIDs as a CDF. Some files are seen on both networks, but the distribution of these observations is different. The "Common on Gnutella" line shows the CDF of these common files as seen on Gnutella, and similarly for the "Common on eMule" line.**

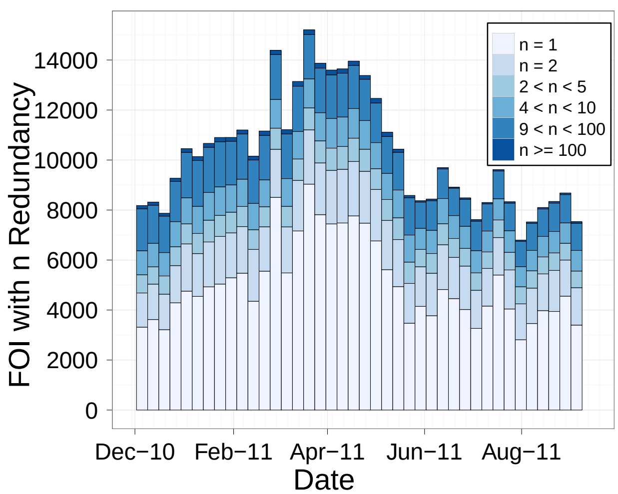



**Figure 3: The per-date $n$-redundancy of FOI across Gnutella as weekly averages. Most files are not highly redundant ($n = 1$) and thus vulnerable to removal from the network should the single peer possessing them be removed. The height of the stacked bars represents the weekly average number of unique files made available by peers.**

files, we characterize the redundancy and availability of FOI on Gnutella and eMule.

### *3.1.1 File Redundancy Across GUIDs*

Many FOI on Gnutella and eMule are not widely redundant among GUIDs within the same network. Figure 2 shows the relative *redundancy* of FOI, which is the number of GUIDs that possess and make available each file. The distribution is presented as a cumulative distribution function (CDF), which shows on the $y$-axis the fraction of FOI that are shared by *at most* $x$ GUIDs.

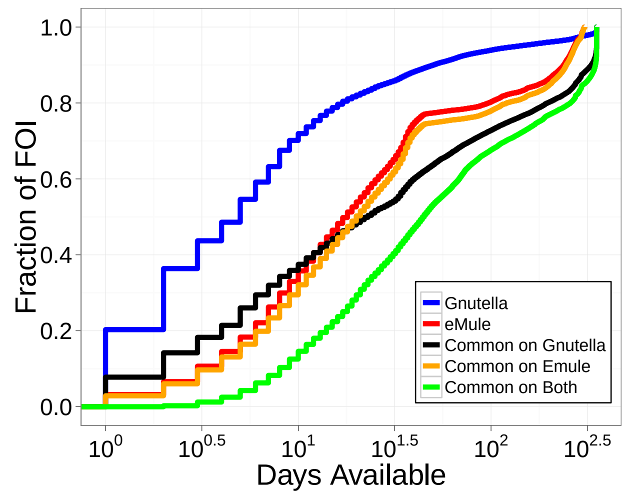



**Figure 4: CDF showing the days available per FOI (during 353 days for Gnutella and 329 days for eMule). Some files are seen on both networks, but the distribution of these observations is different. The "Common on Gnutella" line shows the CDF of these common files as seen on Gnutella, and similarly for the "Common on eMule" line. The "Common on Both" line shows these common files available on either network on any given day.**

For example, 90% of files on Gnutella were shared by at most 20 GUIDs; 99% of files were shared by at most 1,167 GUIDs; and 99.9% of files were shared by at most 9,129 GUIDs.

Figure 2 also shows the relative redundancy for the subset of FOI appearing on both networks. The set of files common to both networks is significantly more redundantly shared on each network than the set of all files on each network. There is high degree of FOI overlap among the two networks: 26,136 of the FOI on the eMule network (nearly 89%) were also seen on the Gnutella network, and 97% of Gnutella GUIDs were observed with at least one file that can be found on the eMule network.

Figure 3 shows the level of file redundancy on a weekly basis in Gnutella. As the stacked bars show, most files are only $n = 1$ redundant on a given week (indicating that only one copy exists in the network), with a small percentage shared by more than $n = 9$ GUIDs, and very few shared by $n = 100$ GUIDs. The overall low redundancy of most files suggests the strategy of prioritizing the investigation of users who possess a large amount of less redundant FOI in order to remove it from the network and prevent its proliferation. An easily intuited proxy for this measure is to target GUIDs who possess large corpora. Since most FOI are relatively less redundant, the GUIDs with the largest libraries likely have the most FOI with low redundancy.

### *3.1.2 File Availability Across Days*

We say a file is *available* on a given day if at least one peer is sharing that file on that day. This approach is simple in that it does not take into account bandwidth and reachability considerations, which are difficult to measure globally. We do not expect this simple definition to limit the applicability of our results, as the assumption of high bandwidth and reachability is conservative from the perspective of law enforcement.

Figure 4 plots the availability of FOI as a CDF on a semi-log scale. Gnutella files tend to have lower availability than eMule, with 80% of files available for more than one day; about 30% are available for more than 10 days; and about 5% of files are available for more than 100 days. Generally, files that are available for a single day are unique to a specific GUID; files that tend to have longer availability are possessed by many GUIDs, not all of whom are online on a given day. Again we see that the files that are common to both networks are more available than is typical on each individual network. By examining files available in common both networks, we see that these common files are significantly more available: about 30% are available for more than 100 days.

Although on average individual files are not often available for long periods of time, the number of files available at a given time is high. The height of the stacked bars in Figure 3 quantifies the number of FOI available per week on average on Gnutella. We have also calculated that on a daily basis, an average of 9,712 distinct files are available, with a peak of 32,020 files.

### 3.2 Law Enforcement Strategy

Our *law enforcement model* is as follows. Investigators have a global, historical view of GUIDs and their associated corpora, including both known FOI and other files. Investigators look to reduce FOI *availability*, by arresting the users that correspond to peers and removing their corpora from the network. Investigators aim to remove files from the network completely.

Content can be removed from these networks only by arresting users and taking their shared libraries offline, as the protocols and implementations inhibit falsifying or polluting content. Our goal is to find out which peers should be removed such that we minimize the number of files that are available at least one day.

In Appendix A we show that this problem is at least NP-Hard. Therefore, we evaluate a number of greedy heuristics aimed at reducing the availability of CP by removing peers. Our evaluation consists of removing subsets of peers from the data and examining the effect on availability. Specifically, we examine the following four heuristics:

- removing peers that were *observed* most often, i.e., largest number of days observed;
- removing peers with the largest *corpus* size;
- removing peers with the largest *contribution* to availability (as defined below); and
- removing peers selected randomly, as a baseline.

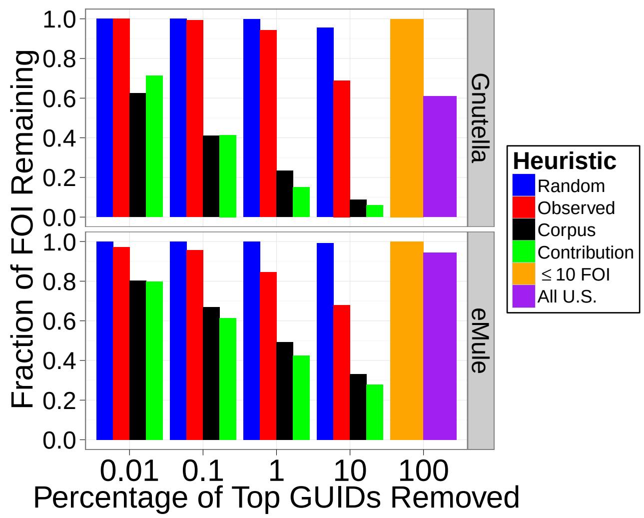



**Figure 5: The remaining fraction of FOI available at least one day given a percentage of GUIDs removed according to different heuristics: random, number of days observed, corpus size, and contribution to file availability on Gnutella and eMule. We removed the top 0.01, 0.1, 1, and 10 percent of GUIDs according to each heuristic. In Gnutella, the Corpus and Contribution heuristics achieve equal results when 0.113% of GUIDs are removed. Also shown is the impact of removing 100% of peers with 10 or fewer FOI, and 100% of peers in the U.S.**

For an arbitrary file on an arbitrary day, $n$ peers possess that file. We say that each peer provides a *file-contribution* of $\frac{1}{n}$th of that file. A peer's *contribution* to file availability is the sum of the file-contributions of the files in their corpus over the duration of the study.

An alternative measure of availability is *daily redundancy*, which is the number of peers that share a file on a specific day. The algorithm to optimally reduce the maximum redundancy over all files shared is simple: remove the peers with the largest corpus size first. It is unclear that minimizing redundancy, unless it is to zero (which is equivalent to unavailability), is useful or effective. To evaluate the effect of reducing redundancy to a small value, we would require reachability, bandwidth, and propagation models of the underlying p2p overlays. Thus, we do not consider daily redundancy further.

#### 3.2.1 Comparison of the Efficiency of Heuristics

Figure 5 compares the effectiveness of each of the above heuristics. Interestingly, removing the peers that were seen the most often has almost no effect on the availability of FOI. Removing peers by either contribution or corpus size is most effective; these measures are correlated, so their similarity in performance is unsurprising.

The vast majority of files are shared only by a relatively small set of prolific GUIDs. Consider Gnutella (though similar trends hold for eMule): If we remove the top 0.01% of 775,941 GUIDs as determined by corpus size, only 59% of the known FOI remain available in the network. In other words, 41% of the unique files on the network are made available by a group of only about 80 GUIDs. In fact, the top 0.01% have 3,242 distinct FOI on average, with the top peer possessing about 25,000 FOI. However, note that most of the files are only available for a relatively short amount of time; as Figure 4 shows, only 28% are available for more than 10 days during our study. Some of this is due to the relatively low number of days these high-contributing GUIDs were observed; this also explains the failure of the observed days heuristic.

These prolific GUIDs have a worldwide presence. Removing them requires tremendous multi-national cooperation as we discuss below.

### 3.2.2 Impact of Geography on Availability

GUIDs sharing CP are located all over the world, as shown in Figure 1. Our data are mostly based on the efforts of U.S. law enforcement, and the files they are looking for are arguably tuned to U.S. perpetrators. As law enforcement agents are limited by jurisdiction, the locational diversity of these users provides a resistance to the straightforward approach of prioritizing them. Only a small majority of top Gnutella GUIDs (by corpus size)—57 out of 100—are located in the U.S. The rightmost bar ("All U.S.") in Figure 5 shows the reduction in availability if we restrict our removal to U.S. GUIDs (that is, GUIDs with an IP located in the U.S.) only. Note that we remove *all* such GUIDS in our analysis, a clearly infeasible approach in practice. Just 30% of files are unavailable (internationally) after removing all GUIDs in the U.S.; removing just the top 0.01% internationally (a group of about 80 GUIDs) has a similar effect, suggesting the utility of a coordinated international approach.

Within the U.S., the problem is similarly large in scope. The top 5% of GUIDs in the U.S. comprises a set of 14,410 GUIDs, each with a corpus of at least 40 known FOI. Due to the weeks of manual effort required for each arrest, the limited resources in the U.S. allow for 3,100 arrests per year for both offline and online offenses [28].

### 3.2.3 Impact of Low-Sharing GUIDs on Availability

A large portion of GUIDs have comparatively few files. As shown in Figure 6, about 82% have 10 or fewer FOI. There are several reasons peers may appear to have few files. They may have files that are CP, but are not yet known to be FOI. They may be downloading FOI and not subsequently sharing them. They may have downloaded the files incidental to other activities. Finally, they may simply be sharing a smaller library. We expect removal of such low-sharing users to impact file

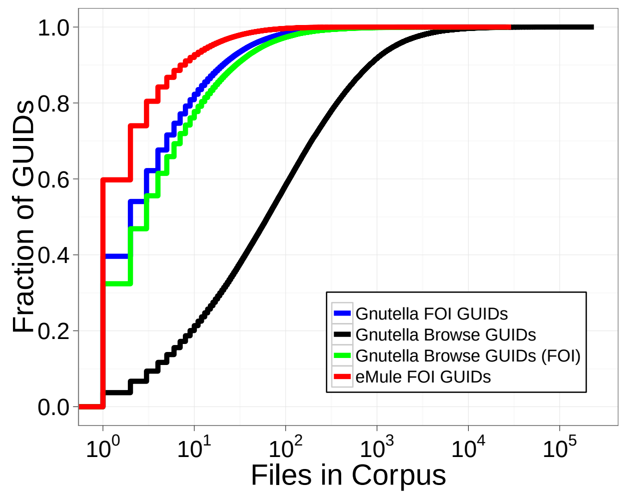



**Figure 6: CDF showing the corpus size per GUID, for various measurement types. The black line ("Gnutella Browse GUIDs") show the corpus size distribution for all files seen at GUIDs whose libraries were browsed, and the corresponding green line shows the distribution of FOIs within those browses. The other two lines show all FOI observed in any manner. (n.b., eMule does not allow browses.) Most GUIDs have very few files in their corpus. We give a week-by-week breakdown of Gnutella library sizes in Appendix D.**

availability significantly, since very many peers possess few files. Contrary to our expectations, removal of these GUIDs sharing few files has essentially no effect on file availability, as shown in the second-rightmost bar in Figure 5 ("$\leq$ 10 FOI"). This result provides further evidence that file availability depends primarily on those GUIDs with the largest corpora, though it does not consider the contribution to redundancy that these low-sharing GUIDs provide.

## 4. COMPARING AGGRESSIVE PEERS

In Section 3, we show that strategies for removing content from the entire ecosystem must target offenders from all countries. In the absence of a unified effort—and no such collaboration exists—investigators need a triage strategy. In this section, we characterize triage metrics for local investigators. Ideally, investigators would target the most *dangerous* offenders: those that are personally, physically abusing children. Unfortunately, such information is not available, sometimes not until months or years after arrest [27].

In lieu of that ideal, local investigators can target peers that are the most aggressive offenders: peers contributing most to the ecosystem. Specifically, we define *aggressive offenders* as those that are online for the longest duration and share the largest number of FOI. Similarly, investigators may target offenders that are conduits between p2p network communities (e.g., by sharing on both eMule and Gnutella), or offenders that seek to escape detection and justice by using Tor or network relays. Offenders that belong to any of these

| Identifier | Network | |
|---|---|---|
| | Gnutella | eMule |
| All GUIDs | 775,941 | 1,895,804 |
| Multi-Network GUIDS | 84,925 (11%) | 147,904 (7.8%) |
| Tor GUIDs | 3,666 (0.47%) | 16,290 (0.86%) |
| Tor GUIDs (> 2 days) | 2,592 (0.33%) | 11,998 (0.63%) |
| Relayed GUIDs | 76,478 (9.9%) | 78,223 (4.1%) |
| Top 10% Observed | 84,235 (11%) | 190,797 (10%) |
| Top 10% By Corpus | 77,782 (10%) | 189,951 (10%) |
| Top 10% By Contr. | 77,595 (10%) | 189,581 (10%) |

**Table 2: Sizes of each GUID subgroup. Definitions of each subgroup appear in this section.**

categories exhibit greater evidence of *intent* [11] beyond the average case, which is an important practical legal concern.

We quantify the aggressiveness of six subgroups of peers sharing FOI. We characterize the contribution of each subgroup to the duration of CP availability and the amount of CP content. The subgroups are:

- The top 10% of GUIDs sharing largest corpora;
- The top 10% of GUIDs seen sharing FOI the most number of days;
- The top 10% of GUIDs ranked by the *contribution* metric defined in Section 3.2;
- The set of GUIDs sharing FOI on at least two p2p networks (linked by IP address);
- GUIDs that use a known Tor exit node;
- GUIDs sharing FOI that use an IP address that we infer to be a non-Tor relay.

Our results show that the subgroup of the top 10% of peers ranked highest by our contribution metric is more aggressive than other subgroups. Further, all these subgroups are more aggressive than a group that consists of all peers that we observed. The exception is the subgroup of GUIDs using non-Tor relays, as we explain below. The differences of each subgroup to the set of all GUIDs are significant ($p < 0.001$).

Below we provide characteristics of each subgroup, and details of the behavior of each. For example, we show that GUIDs using Tor to share FOI use Tor irregularly, and therefore their true IP addresses are easily identifiable. Subsequently, we compare all subgroups side-by-side.

## 4.1 Peer Subgroups

The size of each subgroup is shown in Table 2. The size of the top 10% by corpus and observed days subgroups are slightly larger than 10%. This variability is due to ties in the ranked lists of GUIDs. We include all such GUIDs to avoid arbitrary tie-breaking.

### 4.1.1 Top 10% Groupings

Users can aggressively participate in p2p networks in two primary ways: by contributing a large number of files or a large amount of time. For example, one peer may share 100 files for a single day, and another may share a single file for 100 days. In the first, case, the content is large but other peers have only a limited time to take advantage. In the second case, the content is small but other peers will find it easier to gain access to the content. It is vital for investigators to address both types of aggressive behavior; the contribution metric balances these two concerns.

For these reasons, we create three subgroups corresponding to the 10% of GUIDs with the largest corpora of files ($\mathcal{F}$), the 10% with the most days observed online ($\mathcal{D}$), and the top 10% of GUIDs when ranked by the contribution metric ($\mathcal{C}$). There is substantial but not overwhelming overlap among these subgroups. The overlap in Gnutella, as defined by Jaccard's index, $J(A, B) = \frac{|A \cap B|}{|A \cup B|}$, is $J(\mathcal{C}, \mathcal{F}) = 0.51$ and $J(\mathcal{C}, \mathcal{D}) = 0.28$; the eMule subgroups overlap similarly.

### 4.1.2 Multi-Network Peers

Law enforcement are interested in users that are active on multiple p2p networks. Such users are more aggressive in terms of assisting in the distribution and availability of content to two communities, possibly acting as a bridge. We identify the set of GUIDs in Gnutella that are active in another network by finding all IP addresses in our Gnutella dataset that also appear in any of our eMule, Ares, or IRC datasets; and correspondingly in eMule for those that appear in any of the Gnutella, Ares, or IRC datasets. We refer to GUIDs in these sets as *multi-network GUIDs.*

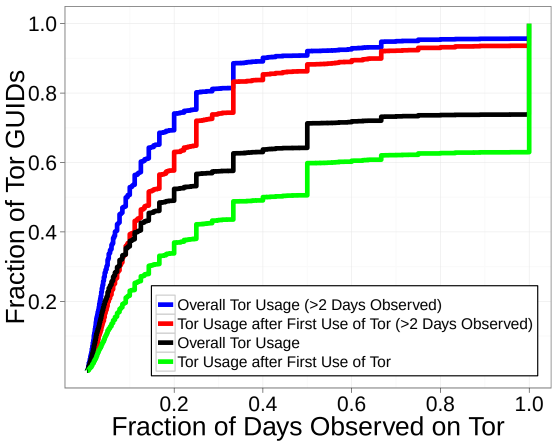
The total number of IPs addresses, private IPs (as per RFC 5735), and Tor exit nodes that we observed for each of these networks is shown in Table 3. Gen-

| Network | IP Addresses | | |
|---|---|---|---|
| | total | private | Tor |
| Gnutella | 3,025,530 | 32,195 | 7,357 |
| eMule | 5,643,350 | 1,256 | 21,025 |
| Ares | 1,714,894 | 225 | 1,799 |
| IRC | 88,658 | 245 | 746 |

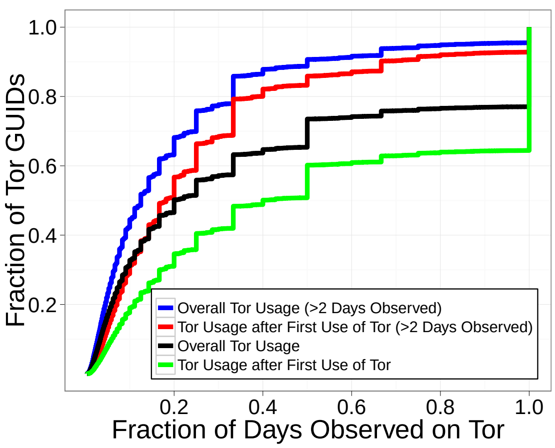
**Table 3: Number of IP addresses per network observed sharing FOI. In the case of IRC, the IP addresses correspond to clients observed in public chat rooms related to child sexual exploitation.**

| Networks $A$ | $B$ | %$A$ | $A \cap B$ | %$B$ |
|---|---|---|---|---|
| Gnutella ∩ | eMule | 6.8% | 199,824 | 3.1% |
| | IRC | 0.1% | 3,562 | 4.1% |
| | Ares | 1.0% | 30,596 | 1.8% |
| eMule ∩ | IRC | 0.1% | 4,654 | 5.3% |
| | Ares | 0.9% | 56,921 | 3.3% |
| IRC ∩ | Ares | 2.1% | 1,813 | 0.1% |
| Intersection of all | | | 308 | |

**Table 4: Overlap of IP addresses across multiple networks, excluding Tor IPs and private IPs. A small but significant set of IPs were seen across multiple networks, indicating particularly active users.**




**(a)** Gnutella




**(b)** eMule

**Figure 7: CDF of Tor usage per GUID for Gnutella and Tor. GUIDs do not use Tor consistently after first being observed at a Tor IP. In both networks, under 40% of Tor GUIDs consistently used Tor after first being observed using it. When considering only Tor GUIDs seen on >2 days (which comprise about 70% of all Tor GUIDs), the rate falls to below 10%.**

erally, private IPs are the result of sub-optimally or misconfigured end-user applications, as opposed to indicating privacy awareness. In contrast, Tor use indicates privacy-aware users. Table 4 shows the size of each pairwise network overlap. For all such intersections, we first remove private IPs and Tor exit nodes (as listed in the Tor consensus files[8]). Of all network pairs, the Gnutella-eMule overlap is the largest.

The union of all three intersections comprises our 84,925 GUID multi-network subgroup for Gnutella. We perform a similar calculation for eMule, resulting in 147,904 GUIDs.

### *4.1.3 Peers that Use Tor*

Peers that use Tor are of interest to law enforcement because they are actively masking their identities, thwarting investigations of this crime. Tor does not filter application-level data: GUIDs are passed through to investigators, and thus appear in our dataset as well. We define a GUID as a *Tor GUID* if it was ever once observed as having an IP address listed as a Tor exit node in the Tor consensus for the date of the observation. When a Tor GUID's IP is a known Tor exit node, as listed in the Tor consensus for the date of the observation, we say that the GUID is *using* Tor. As Table 2 shows, this set is not large on either network: 3,666 GUIDs for Gnutella and 16,290 GUIDs for eMule.

It is striking that the vast majority of Tor GUIDs do not use Tor consistently, which makes it possible to detect their true IP address. In Figure 7, we show the CDFs of *overall Tor usage*. In both networks, only about a quarter of the Tor GUIDs used Tor every time they were observed. More significantly, for these GUIDS, under 40% consistently use Tor after their first use of Tor.

When we examine these 40% of nodes that used Tor consistently, we found that most were observed on the Gnutella and eMule networks for only one or two days. Therefore, we recomputed the distribution of Tor usage for the subset of Tor GUIDs observed three or more days, which is over 70% of all Tor GUIDs. We again also computed the CDFs of Tor usage after first using Tor. The resulting CDFs are the upper lines in Figure 7. In sum, over 90% of GUIDs using Tor for more than two days on eMule and Gnutella are easily linked back to a non-Tor IP address, one that is most likely their real location. (GUIDs seen only one day may all be the same user on different days; see Section 5 for a discussion of user aliasing in our dataset.)

This irregular use could be due to ignorance of how Tor works, careless configuration of the computer, or frustration with the lower throughput of Tor. It is well known that Tor's *design* does not offer technical protection to p2p users because it does not hide identifying application-level data [16]. In contrast, we provide the first empirical evidence that Tor *users* do not use the software consistently, even among those with a strong reason to so. Regardless of the quality of Tor's security, this evidence strongly suggests that its usability (its interface, its effects upon perceived speed, or some other factor) is lacking. We conclude that the use of Tor, as observed in practice, poses only a small hurdle to investigators. Given this result, reports by the Tor developers that "Journalists use Tor to communicate more safely with whistleblowers and dissidents"[9] should give one pause, without evidence that those groups are

[8]Consensus files contain the list of IPs addresses acting as exit nodes on a daily basis; see `https://metrics.torproject.org/data.html`

[9]Quoted from `https://www.torproject.org/about/overview.html.en`

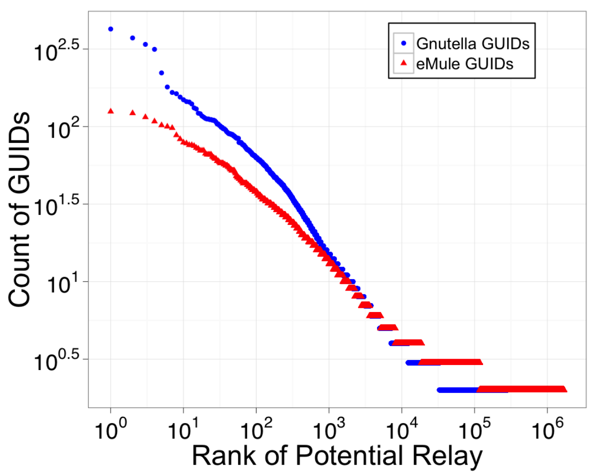



**Figure 8: Rank-order plot of the number of GUIDs seen at each potential relay IP.**

| | GUID Group | Mean Value (99% CI) Corpus Size | | Days Obs. | |
|---|---|---|---|---|---|
| | All | 10.9 | (10.7, 11.1) | 5.2 | ( 5.2, 5.2) |
| Gnutella | Tor | 43.9 | (39.0, 49.6) | 23.4 | (21.8, 25.1) |
| Gnutella | Relayed | 18.9 | (18.3, 19.5) | 4.8 | ( 4.7, 4.9) |
| Gnutella | Multi-Network | 25.9 | (24.9, 27.0) | 10.8 | (10.6, 11.0) |
| Gnutella | Top 10% Obs. | 41.8 | (40.7, 43.0) | 28.7 | (28.5, 29.0) |
| Gnutella | Top 10% Corp. | 75.9 | (74.3, 77.7) | 16.2 | (16.0, 16.5) |
| Gnutella | Top 10% Contr. | 69.1 | (67.6, 70.9) | 19.5 | (19.3, 19.8) |
| | All | 4.3 | ( 4.3, 4.4) | 4.1 | ( 4.1, 4.1) |
| eMule | Tor | 21.2 | (19.9, 22.5) | 17.4 | (16.9, 18.0) |
| eMule | Relayed | 9.2 | ( 8.9, 9.6) | 5.5 | ( 5.4, 5.6) |
| eMule | Multi-Network | 10.8 | (10.6, 11.0) | 9.5 | ( 9.4, 9.7) |
| eMule | Top 10% Obs. | 23.5 | (23.2, 23.8) | 22.3 | (22.2, 22.4) |
| eMule | Top 10% Corp. | 27.8 | (27.4, 28.5) | 18.7 | (18.6, 18.8) |
| eMule | Top 10% Contr. | 25.8 | (25.4, 26.5) | 19.0 | (18.9, 19.1) |

**Table 5: The expected value and 99% confidence interval of each characteristic for each subgroup of GUIDs. Each subgroup's mean differs from the mean of the "All" group. Each such difference is statistically significant ($p < 0.001$), as determined by a computational permutation test ($R = 10,000$). Confidence intervals are computed by bootstrap ($R = 10,000$).**

significantly more tech-savvy than the peers we study.

### *4.1.4 Peers that Use Suspected Relays*

The final subgroup we identify is a set of peers that are using IPs that we suspect are relays (other than Tor exit nodes). To create this subgroup, we first collected the set of non-Tor IP addresses used by GUIDs that also used a Tor exit node. Figure 8 shows a rank-order plot of these IPs. We discard the IPs that hosted fewer than four GUIDs (267,035 in the case of Gnutella, and 1,671,419 for eMule), and we nominate the remaining IPs as potential relays. Finally, we create the subgroup of *relayed GUIDs* as the set of GUIDs seen using the potential relays. We cannot validate these GUIDs as having definitely used relays; for example, it may be the potential relays are IP addresses that get reassigned frequently. However, we consider their use of these shared IPs sufficient to define them as a distinct set.

| | GUID Group | Library Size, Month-over-Month increase | decrease | varied | no change |
|---|---|---|---|---|---|
| Gnutella | All | 6.4% | 6.0% | 8.3% | 79.3% |
| Gnutella | Tor | 11.5% | 9.4% | 29.1% | 50% |
| Gnutella | Relayed | 5.4% | 4.5% | 5.5% | 84.6% |
| Gnutella | Multi-Network | 9.2% | 8.7% | 16.1% | 66% |
| Gnutella | Top 10% Obs. | 13.9% | 12% | 47.7% | 26.4% |
| Gnutella | Top 10% Corp. | 12.2% | 11.5% | 36.8% | 39.5% |
| eMule | All | 5.5% | 4.5% | 7.5% | 82.5% |
| eMule | Tor | 10.4% | 8.7% | 35.9% | 45% |
| eMule | Relayed | 6.6% | 5.4% | 9.1% | 78.9% |
| eMule | Multi-Network | 8.9% | 7.8% | 19.6% | 63.7% |
| eMule | Top 10% Obs. | 15.3% | 11.8% | 57.4% | 15.5% |
| eMule | Top 10% Corp. | 16.5% | 11.7% | 54.4% | 17.4% |

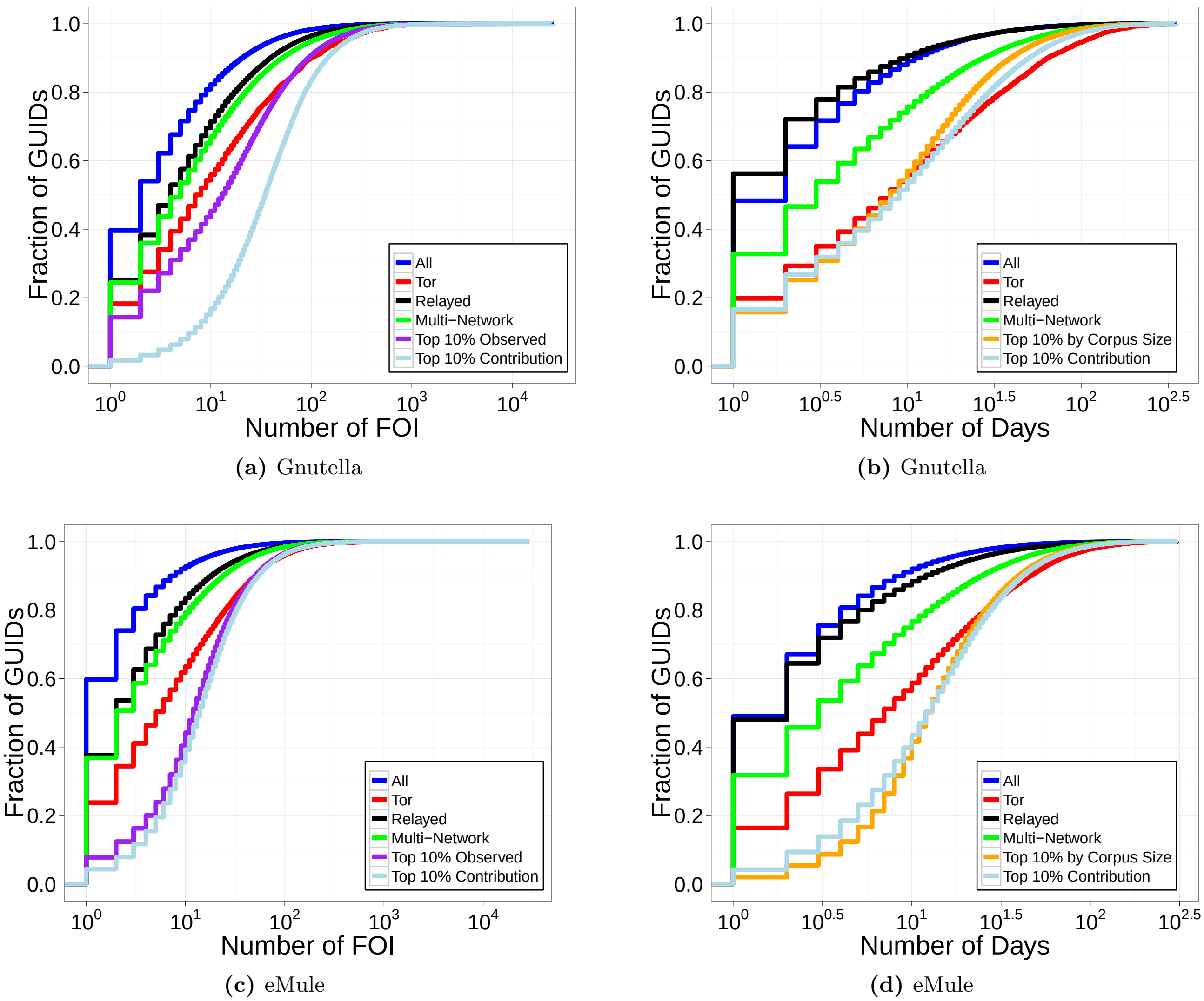
**Table 6: Month-to-month changes in GUID libraries. "Increase" or "decrease" means that the GUID's library (not corpus) size for the month monotonically increased or decreased over time; "varied" means the GUID's library size fluctuated.Tor and relayed GUIDs were generally more active in modifying their libraries than ordinary GUIDs; further, they tended to increase their library size consistently over time.**

## 4.2 A Comparison of Peer Behavior

There are substantive and statistically significant differences among the subgroups in terms of per-GUID corpora and number of days observed. These differences can be seen in Figure 9 and are summarized in Table 5. In particular, the subgroups generally have a larger corpus size and more days observed online than the set of all GUIDs. The three top-10% subgroups show this effect most strongly, but the Tor subgroup and multi-network subgroups show similar effects. Notably, these latter two subgroups are selected independently of corpus size and days online. This result confirms a hypothesis that tech-savvy groups, whether through Tor or multi-network use, are more aggressive.

The set of GUIDs in the top 10% contribution subgroup represent a combination of the other aggressiveness metrics. This result can be viewed by comparing CDFs in the figure, or by comparing means in the table. For example, the top 10% contribution subgroup's mean corpus size is higher than the top 10% observed, and its mean number of days observed is higher than the top 10% corpus subgroup. The contribution metric could easily be parameterized to weight observations more heavily, though we do not show such results here.

The relayed subgroup in general has larger number of FOI than the all group, and appears online more days on average than the all group in eMule. However, the relayed subgroup shows fewer days observed online than the all group in the Gnutella network. This result suggests that either this subgroup, as we've defined it, is not aggressive, or that our process for identifying non-Tor relays is faulty. It may also be that the peers in the relayed subgroup are more successful at aliasing themselves as different GUIDs that appear on the network




**(a)** Gnutella **(b)** Gnutella

**(c)** eMule **(d)** eMule

**Figure 9: Characterizations, as CDFs, of per-GUID corpora and days observed. The top GUIDs by contribution are among the most aggressive in both networks. Tor GUIDs are also significantly more active, and Multi-network GUIDs are active as well. Relayed GUIDs show a more mixed result: while otherwise more aggressive than the all group, they are observed less often on the Gnutella network.**

fewer number of days each. In Section 5, we examine the general problem of peer aliasing in this data set.

In Table 6, we show a final comparison: the changes in library size over time for each subgroup. Tor GUIDs are much more likely to increase their library size month-to-month than relayed or multi-networked GUIDs, though relayed GUIDs also show this effect. In fact, Tor GUIDs are comparable in their behavior to the generally more aggressive top 10% subgroups. This correlation is unsurprising, in that one-third of the Tor GUIDs in each network are also members of these groups.

# 5. MEASUREMENT LIMITATIONS

Our data, corresponding to law enforcement observations, present a limited, non-random sampling of peers and files; we are cognizant of potential bias and describe several limitations in this section.

## 5.1 Collection Bias

First, while our data are extensive and detailed, we do not necessarily have information on all peers that are sharing CP. Law enforcement often locate new peers by a keyword search and verify that a candidate file is a FOI using hash values; there is no systematic spidering of the network. Similarly, we do not necessarily have information on all CP content being shared in the network nor by any specific peers. Instead we have data on previously verified, that is, *known* FOI; this verification was done manually by law enforcement, and we identify verified files by hash value.

Some studies [8,9,24,25] choose to label search results as CP if the result contained a CP keyword. That approach is necessary in absence of file hashes of known FOI. We forgo that approach, in part because of the high chance of mislabeling non-CP files as CP. We have many observations of files that have names claiming to be CP but which are not known FOI; we do not treat such files as CP (or FOI) in our data. Similarly, we have observations of CP files that do not have any CP terms in the filename. Hence, our measurements of the amount of CP shared by peers are conservative lower bounds. Further, the set of known FOI was identified by U.S. investigators. It is likely biased towards files and filenames shared by traffickers in the U.S. Traffickers in other countries are likely underestimated by our study.

Second, many of our records are the result of keyword or hash searches that do not provide a complete listing of a peer's currently shared files. Ideally, all of our records would be associated with a browse, in other words, a complete listing of the peer's current files. eMule does not support browse functionality at all, and investigators do not browse all Gnutella peers on all days. For example, a peer may be identified as having file $A$ on day 1 and day 3, but that file is not seen on day 2 because the appropriate keyword or hash search was not run. As a result, we may be underestimating the amount of CP content possessed by each peer as well as the number of days they are online. Further, Gnutella peers with more files are more likely to be found in a keyword search, and subsequently browsed. Similarly, peers that are online more often are also more likely to be found using a search. The result is that we might be underestimating the number of peers that are rarely online and have few files.

Third, we use the GUIDs[10], defined in the Gnutella and eMule protocols to uniquely identify peers, to distinguish between different observations. A user may be associated with, or *aliased* to, more than one GUID. For example, a user might have one or more installations of the Gnutella client software with each installation showing up as a different GUID. Hence, the number of GUIDs in these networks serves as a rough upper bound on the number of users. We explore this issue below.

Finally, some peers are removed from the network. Before, during, and after the collection of the datasets we analyze, law enforcement were and are active in investigating and arresting CP traffickers. When arrests of users of p2p networks occur, the peers corresponding to the arrested users are removed from the network. We do not know which peers correspond to such removals, and we do not take these removals into account in our analyses.

[10]We do not use IP addresses for this purpose. The challenge of IP aliases that DHCP, NAT, and similar mechanisms present to traditional network measurement are largely irrelevant to law enforcement. A subpoena of billing records and fruits of a subsequent search warrant can dealias IP addresses reliably.

## 5.2 User Aliasing

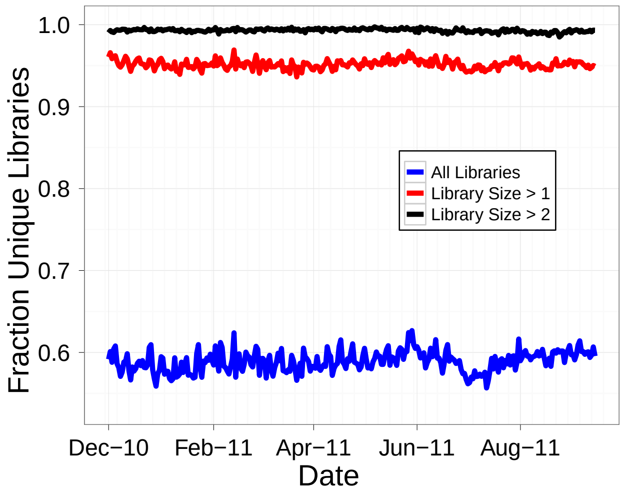
The relationship between p2p network GUIDs and real users is not one-to-one in our dataset. In fact, it is possible for multiple distinct GUIDs to correspond to a single user in our dataset. We refer to this phenomenon as *user aliasing*. In this section, we examine observable user aliasing and attempt to quantify its effects upon the analyses in the previous sections. In sum, we find that GUIDs that share at least three FOI any given day generally have distinct libraries. In Gnutella, we can compare all files shared by a GUID, and in that case users sharing a library of at least two files are generally distinct on a given day. We also find little evidence to suggest users are changing their GUIDs and then continuing to share the same library or a portion of it later that day.

The true user aliasing rate in our data (that is, the relative frequency of two GUIDs in our data set representing one person) is unknowable to us. However, the reasons for deliberate aliasing can be enumerated: *(i)* if a user has two computers (or multiple accounts on a single computer), each with an installation of Gnutella, he will control two unique GUIDs; and *(ii)* a user may reinstall or upgrade their p2p client on a single computer or otherwise modify their GUID over time. We have no way of detecting the first case from only network data; however, the second case can be detected if the user does not alter what files they are sharing, as the file library acts as a kind of signature for the user. It is this latter case that we evaluate in the remainder of this section.

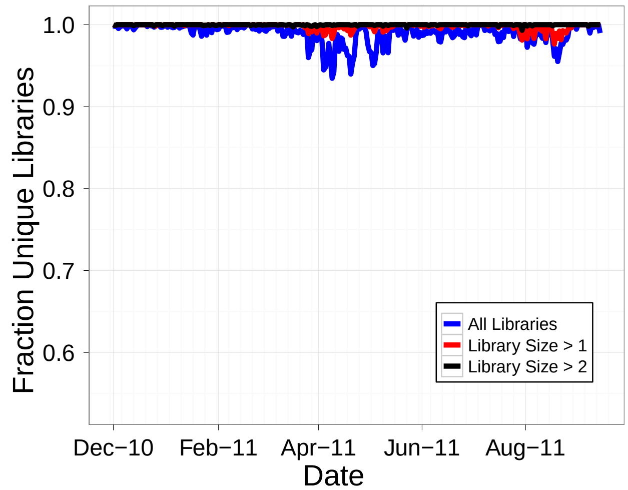
Most users, as identified by GUIDs, are seen with very small libraries of a single file or two. This fact is illustrated in Figure 6 in Section 3 (and in a week-by-week breakdown in Appendix D). We posit that such small libraries are not particularly differentiable. By excluding them, we can determine a lower bound on the user aliasing of type *(ii)* that may be occurring.

**Analysis and Results.** We computed day-by-day similarities between Gnutella libraries to determine a lower bound on user aliasing, or alternatively, an upper bound on the number of unique libraries present in the dataset. Generally, we found most libraries to be distinct. Parallel results to hold for eMule, though our view of eMule user libraries is limited due to the lack of browse functionality in that system.

Figures 10(a) and (b) show a comparison of Gnutella GUID libraries, plotting the fraction of GUIDs with libraries that are a unique collection of files. In Figure 10(a) a comparison is made of just the files of in-

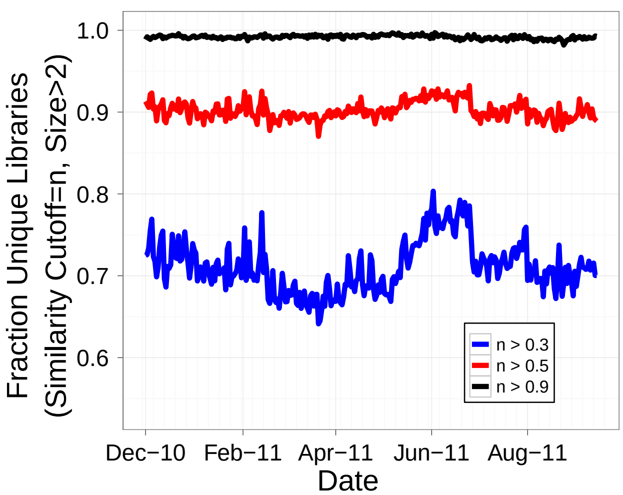



**(a)** Partial view of libraries based only on FOI.

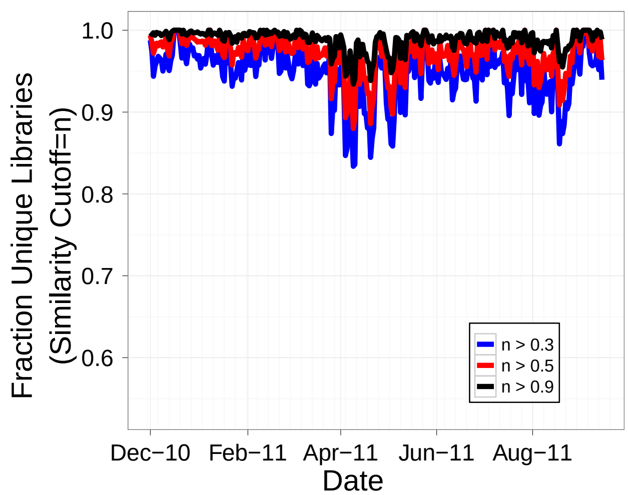



**(b)** Complete view of libraries.

**Figure 10: Fraction of GUIDs with unique libraries on specific days, where uniqueness is defined as libraries that completely match. When considering libraries of at least two FOI, approximately 95% are unique. When considering complete libraries, over 93% are unique.**




**(a)** Partial view of libraries based only on FOI.




**(b)** Complete view of libraries.

**Figure 11: Fraction of GUIDs with a unique library, where uniqueness is defined as there being no other library with a similarity greater than n. The similarity of two libraries is defined their Jaccard index, $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$. On most days, 90% of libraries have no more than half their files in common.**

terest at each GUID; Figure 10(b) compares all files in the library of each GUID (from a Gnutella browse request). GUIDs that have tens of files (or more) are easy to distinguish from other GUIDs.

Figure 10(a) shows that in general, GUIDs with a single file are easily aliased with other GUIDs with the same single file: only about 58% of GUIDs have unique libraries on a given day of our dataset. Among the 40% of Gnutella GUIDs that have two or more FOI, over 95% have unique libraries. Among the 25% of GUIDs with three or more FOI, over 99% have distinct libraries.

Fewer aliases are present when comparisons can be made of the complete libraries, as is possible with Gnutella browse information, by including all files, not just FOI. This is illustrated in Figure 10(b). Note that GUIDs with a single FOI typically possess more than one file, and thus they are more likely to be unique. Typically, GUIDs seen with two or more files in their library had a unique library about 95% of the time; GUIDs with three or more files were unique over 99% of the time. eMule does not allow user libraries to be browsed, and so distinguishing GUIDs in Gnutella is easier.

The above data suggest that in our study, we can treat GUIDs as uniquely distinguishable when their libraries contain at least two FOI or when we consider all files that they share. The analysis also suggests that users are rarely if ever changing their GUID and appearing on the same day with the same library. They would appear as aliases if so, and if this was common, the fraction of unique libraries would be lower.

Based on a similar analysis, we also make the claim that there is no compelling evidence that many users are changing GUIDs appearing on the network that day and preserving only *most* of their shared libraries. Figure 11 quantifies the uniqueness of partial and complete libraries using the Jaccard index: $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$. In Figure 11(a), we see that for GUIDs with at least three FOI, approximately 90% of libraries have no more than half their files in common on most days of our study. In Figure 11(b), we compare all files in each GUID's library, not just FOI. Here we see on most days, 85% of libraries have no more than 30% of their files in common.

A limitation of our calculations above is that we compare GUIDs only within a day's time. We haven't computed equivalence or similarity among GUID libraries across multiple days because the computation is too lengthy to handle in a reasonable timescale for our dataset. We plan to compute approximate similarity and equivalence for these larger timespans in future work.

## 6. RELATED WORK

**Ecosystems & Underground Economies.** Our work is similar in theme to a body of work exploring characteristics of ecosystems rooted in network-based services. Many of these works employ economic principles to help explain some of the apparent phenomena associated with security, privacy, and online criminal activity. Acquisti and Grossklags [1] examine the seeming inconsistency between a user's expressed interest for privacy and their willingness to abandon that interest for the sake of short-term benefits; those results may explain the irregular use of Tor by the peers in our dataset. Analyzing a dataset of IRC messages collected over seven months, Franklin et al. [6] characterize and categorize the participants and services existing in the underground economy of criminal activities. Other related work includes analysis of spam value chains [10, 13], freelance labor and its application to web service abuse [18], and the ecosystem supporting pay-per-install services commodifying malware distribution [2].

**Content Availability in P2P Systems.** A large body of related work on p2p systems investigates availability, performance, and issues related to the use of incentives [3, 7, 17, 19, 20, 32]. Unlike our work, these studies mostly focus on understanding and analyzing the unique properties of p2p networks and their users' behavior, and do not specifically target CP or separate aggressive subgroups. Other studies target the legality of p2p networks and investigate ways to prevent or reduce the piracy of copyrighted material using methods such as item poisoning and network pollution [4, 15], which reduces the perceived availability of content.

**CP Trafficking in P2P Systems.** Prior studies of CP-related trafficking on the Internet have a limited scope. They are mostly indicative of the alarming presence of contraband rather than comprehensively quantifying how the files are being shared [9, 12]. Analyzing results from a three-month study of the *isoHunt* BitTorrent indexing service, Prichard et al. [21] conclude that CP is consistently shared and warn against the potential normalization of CP among Internet subcultures. Internet Filter Review [23] reports on the existence of 100,000 websites offering CP and of 116,000 daily CP-related queries on Gnutella.

Hughes et al. [9] examine three days (one week apart each) of Gnutella traffic, finding that 1.6% of search traffic and 2.4% of response traffic is CP-related. Using natural language analysis on the same dataset, Hughes et al. [8] present an approach for automated detection of CP-related activities. Latapy et al. [12] analyze two datasets of keyword-based search queries issued by users of the eMule system. The first dataset is from 2007 and spans 10 weeks, the second is from 2009 and spans 28 weeks. In these datasets, about 0.25% of queries relate to child pornography and more than 0.2% of peers on the network are involved.

Steel [25] analyzes the supply of and demand for CP on the Gnutella network using a dataset of both queries and query hits collected over several weeks. They report that 1% of all queries and 1.45% of all query hits were CP-related, with the median age searched for being 13. They also report that while most of the available CP files are images, 99% of searches are for movies.

In our prior work [14], we analyze CP-related activity on Gnutella during a five-month period with no overlap with the study in this paper. We show that the correspondence between IP addresses and application-level identifiers is not one-to-one, and propose proactive methods of differentiating the end hosts. In contrast, our focus in this work is on reducing file availability and characterizing peer behavior.

Rutgaizer et al. [24] examine 67 days worth of queries and downloads of torrent files hosted by a popular torrent indexing site. The study relies only on data visible to the indexer and tracker, and lacks data and metadata to establish ground truth; unlike our work, the study is not on verified CP content.

## 7. CONCLUSIONS

Partnering with over 2,000 law enforcement investigators, we have analyzed child pornography trafficking on Gnutella and eMule over a year-long period. CP trafficking over p2p networks is widespread. Law enforcement do not have sufficient resources to pursue all CP possessors, and instead investigators must focus their efforts on arresting those criminals that are most dangerous to children and society. In addition to catching child abusers, law enforcement aim to eliminate CP from these networks and thereby stop the continued victimization of the children in the imagery.

We have shown that prioritizing removal of offenders according to our contribution metric is effective in removing content from the network. Further, we have shown that attempts to reduce the trafficking of CP on p2p networks will be most effective when part of a cooperative international effort. Our work characterizing aggressive peers and their contributions to this crime provides law enforcement quantifiable guidance in prioritizing these many offenders. Tor thwarts network investigations, however, we observe that in practice offenders use Tor inconsistently. Over 90% of regular Tor users send traffic from a non-Tor IP at least once after first using Tor. Finally, data collection and analysis such as ours informs policy, enabling policymakers to make decisions grounded in fact.


## Acknowledgements

We are grateful for the participation of law enforcement in this project. This material is based upon work supported in part by the National Science Foundation under Grant Nos. CNS-1018615 and CNS-0905349, in part by the Office of Juvenile Justice and Delinquency Prevention under Grant No. 2011-MC-CX-0001, and in part by the National Institute of Justice under Grant No. 2008-CE-CX-K005.


## 8. REFERENCES


[1] A. Acquisti and J. Grossklags. Privacy and rationality in individual decision making. *IEEE Security & Privacy*, 3:26–33, January 2005.

[2] J. Caballero, C. Grier, C. Kreibich, and V. Paxson. Measuring pay-per-install: the commoditization of malware distribution. In *Proc. USENIX Security*, pages 1–13, Aug. 2011.

[3] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proc. ACM IMC*, pages 1–14, 2007.

[4] N. Christin, A. Weigend, and J. Chuang. Content availability, pollution and poisoning in file sharing peer-to-peer networks. In *Proc. ACM Electronic Commerce (EC)*, pages 68–77, 2005.

[5] J. Chu, K. Labonte, and B. Levine. Availability and locality measurements of peer-to-peer file systems. *Proc. SPIE*, 4868:310, 2002.

[6] J. Franklin, V. Paxson, A. Perrig, and S. Savage. An inquiry into the nature and causes of the wealth of internet miscreants. In *Proc. ACM CCS*, pages 375–388, Oct. 2007.

[7] K. Gummadi, R. Dunn, S. Saroiu, S. Gribble, H. Levy, and J. Zahorjan. Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. *SIGOPS Oper. Syst. Rev.*, 37:314–329, Dec. 2003.

[8] D. Hughes, P. Rayson, J. Walkerdine, K. Lee, P. Greenwood, A. Rashid, C. May-Chahal, and M. Brennan. Supporting law enforcement in digital communities through natural language analysis. In *Proc. Intl Workshop Computational Forensics*, pages 122–134, Berlin, Heidelberg, 2008.

[9] D. Hughes, J. Walkerdine, G. Coulson, and S. Gibson. Peer-to-peer: is deviant behavior the norm on P2P file-sharing networks? *IEEE Distributed Systems Online*, 7(2), Feb. 2006.

[10] C. Kanich, N. Weavery, D. McCoy, T. Halvorson, C. Kreibichy, K. Levchenko, V. Paxson, G. Voelker, and S. Savage. Show me the money: characterizing spam-advertised revenue. In *Proc. USENIX Security*, Aug 2011.

[11] O. S. Kerr. *Computer Crime Law*. West / Thompson Reuters, St. Paul, Minnesota, 2nd edition, 2009.

[12] M. Latapy, C. Magnien, and R. Fournier. Quantifying paedophile queries in a large p2p system. In *Prof. IEEE INFOCOM*, pages 401–405, April 2011.

[13] K. Levchenko, A. Pitsillidis, N. Chachra, B. Enright, M. Félegyházi, C. Grier, T. Halvorson, C. Kanich, C. Kreibich, H. Liu, D. McCoy, N. Weaver, V. Paxson, G. Voelker, and S. Savage. Click trajectories: End-to-end analysis of the spam value chain. In *Proc. IEEE Symp. Security & Privacy*, pages 431–446, Nov 2011.

[14] M. Liberatore, B. Levine, and C. Shields. Strengthening Forensic Investigations of Child Pornography on P2P Networks. In *Proc. ACM CoNext*, Nov 2010.

[15] X. Lou and K. Hwang. Collusive Piracy Prevention in P2P Content Delivery Networks. *IEEE Transactions on Computers*, 58(7):970–983, July 2009.

[16] P. Manils, A. Chaabane, S. Le Blond, M. Kaafar, C. Castelluccia, A. Legout, and W. Dabbous. Compromising Tor Anonymity Exploiting P2P Information Leakage. In *Proc. HotPets*. (See also `http://blog.torproject.org/blog/bittorrent-over-tor-isnt-good-idea`), July 2010.

[17] D. Menasche, A. Rocha, B. Li, D. Towsley, and A. Venkataramani. Content availability and bundling in swarming systems. In *Proc. ACM CoNext*, pages 121–132, 2009.

[18] M. Motoyama, D. McCoy, K. Levchenko,

S. Savage, and G. Voelker. Dirty jobs: the role of freelance labor in web service abuse. In *Proc. USENIX Security*, Aug 2011.

[19] J. Otto, M. Sánchez, D. Choffnes, F. Bustamante, and G. Siganos. On blind mice and the elephant: understanding the network impact of a large distributed system. In *Proc. ACM Sigcomm*, pages 110–121, Aug 2011.

[20] M. Piatek, T. Isdal, T. Anderson, A. Krishnamurthy, and A. Venkataramani. Do incentives build robustness in bit torrent. In *Proc. USENIX NSDI*, Apr 2007.

[21] J. Prichard, P. Watters, and C. Spiranovic. Internet subcultures and pathways to the use of child pornography. *Computer Law and Security Review*, 27(6):585–600, 2011.

[22] S. Rhea, D. Geels, T. Roscoe, and J. Kubiatowicz. Handling churn in a DHT. In *Proc. USENIX ATEC*, June 2004.

[23] J. Ropelato. Internet pornography statistics. `http://internet-filter-review.toptenreviews.com/internet-pornography-statistics.html`, 2007.

[24] M. Rutgaizer, Y. Shavitt, O. Vertman, and N. Zilberman. Detecting pedophile activity in bittorrent networks. In *Proc. PAM Conf.*, Vienna, Austria, March 2012.

[25] C. Steel. Child pornography in peer-to-peer networks. *Child Abuse & Neglect*, 33(8):560–568, 2009.

[26] D. Stutzbach and R. Rejaie. Understanding churn in peer-to-peer networks. In *Proc. ACM IMC*, pages 189–202, Aug 2006.

[27] United States Sentencing Commission. Public hearing on federal child pornography crimes. `http://www.ussc.gov/Legislative_and_Public_Affairs/Public_Hearings_and_Meetings/20120215-16/Agenda_15.htm`, February 15, 2012.

[28] U.S. Dept. of Justice. The National Strategy for Child Exploitation Prevention and Interdiction: A Report to Congress. `http://www.projectsafechildhood.gov/docs/natstrategyreport.pdf`, August 2010.

[29] S. Vinterbo. A Note on the Hardness of the k-Ambiguity Problem. Technical Report DSG-TR-2002-006, Decision Systems Group, Harvard Medical School, June 2002.

[30] R. J. Walls, B. N. Levine, M. Liberatore, and C. Shields. Effective Digital Forensics Research is Investigator-Centric. In *Proc. USENIX Workshop on Hot Topics in Security (HotSec)*, August 2011.

[31] J. Wolak, D. Finkelhor, and K. Mitchell. Child-Pornography Possessors Arrested in Internet-Related Crimes: Findings From the NJOV Study. Technical report, National Center for Missing & Exploited Children, 2005.

[32] C. Zhang, P. Dhungel, D. Wu, and K. Ross. Unraveling the BitTorrent Ecosystem. *IEEE Transactions on Parallel and Distributed Systems*, 22(7):1164–1177, July 2011.

# APPENDIX

## A. PROOF SKETCH OF NP-HARDNESS

We define the PEER REMOVAL problem as follows. Given a set of peers and the set of files they shared over $D$ days, remove at most $r$ peers such that the number of files available for at least one day is minimized. In other words, we minimize $\sum_i f_i$, where $f_i = 1$ if a file $i$ is available from at least one peer during at least one day of the $D$ days, and 0 otherwise.

We show that PEER REMOVAL is NP-hard by reducing the MINIMUM $k$-COVERAGE problem to it. The goal of this NP-hard problem [29] is to select $k$ sets from a collection of $n$ sets such that the cardinality of their union is the minimum.

Given any instance of the MINIMUM $k$-COVERAGE, we construct an instance of PEER REMOVAL in LSPACE as follows. Let each of the $n$ sets represent a peer with each element of the set representing a file owned by him. Removing $r = n - k$ peers such that availability is minimized would be the same as selecting $k$ peers such that the cardinality of the union of their corpora is minimized.

## B. MOBILITY

Each IP address in our dataset is linked to a city-level geographic location using results from a commercial service. We examined whether GUIDs that appeared in multiple cities are characteristically different from other GUIDs within a given network. The results in this section show that, unlike the peer groups analyzed in Section 4, multi-city peers do not stand out in terms of days online or library size.

GUIDs can appear in multiple cities for several reasons. First, the user may move to different geographic locations. Second, the user may use a remote host to which he has have access, or he may use a relay. We can't distinguish these two cases other than for IP addresses that we know or believe to be relays. Therefore, our set of multi-city GUIDs does not include GUIDs that we observed using Tor or a potential relay (see Section 4.1.4). Using this definition, multi-city GUIDs account for 81,496 GUIDs in the Gnutella network. (We elide an analysis of eMule in this section.)

Figure 12 shows the number of FOI held by multi-city GUIDs. For comparison, the same values for all GUIDs and for Tor GUIDs in Gnutella are repeated from Figure 9a. The data shows that multi-city GUIDs tend to have

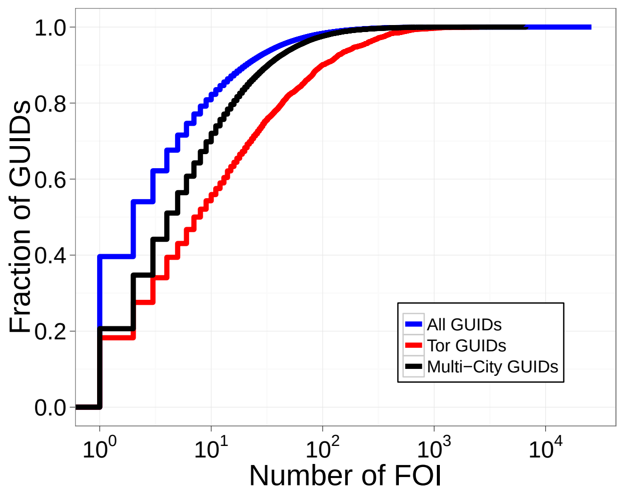



**Figure 12: CDF of the fraction of GUIDs versus number of files of interest. Multi-city GUIDs do not share significantly more FOI than the set of all GUIDs, and much fewer than Tor GUIDs. "All GUIDs" and "Tor GUIDs" data is repeated from Figure 9a.**

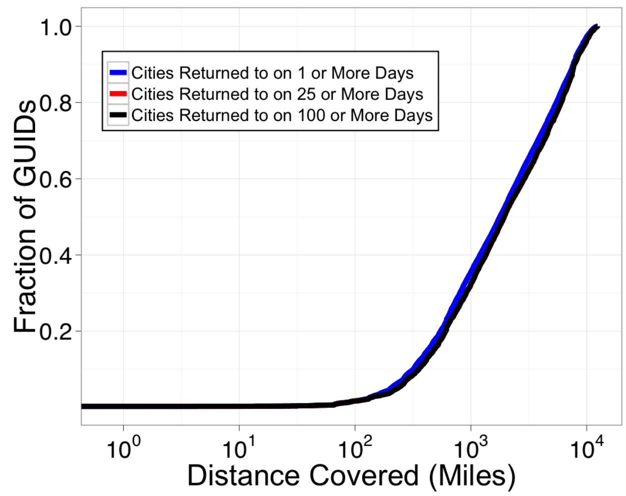



**Figure 13: CDF of GUIDs versus the distance covered in miles. Each line reduces the valid locations of a GUID by only considering locations where a GUID returned to a city more than $n$ times. Limiting by valid cities produces no differences in behavior.**

same fraction of FOI in their corpus when compared to all GUIDs in the network.

We also attempted to find interesting subsets of multi-city GUIDs that were repeatedly in different cities. First, we coarsely computed the geographic diameter of the cities associated with each multi-city GUID. The diameter is the geographic length of the diagonal of a rectangle that covers all cities that a multi-city GUID was seen at. Specifically, we used the haversine formula to compute distance. The red line in Figure 13 shows a CDF of the distance covered by multi-city GUIDs. We can filter the entire group by, for each GUID, ignoring IPs in cities that were visited fewer than $n$ times. Figure 13 illustrates that limiting a GUID to a subset of its cities yields no interesting subgroups.

Finally, we characterized multi-city GUIDs by the time spent in their *home city*, defined the as the city a GUID was observed in most. We found that 42,761 multi-city GUIDs (52%) are seen in their home city more than 50% of the time. Do GUIDs that are more nomadic contribute larger libraries to the network? We have no evidence to support such a claim, as shown in Figure 14. In that figure, the multi-city GUIDs are ordered by the fraction of time spent in their home city and then bucketed in 10% increments. Each bucket contains roughly 8,000 unique GUIDs and represents the number of files of interest in each multi-city GUID's corpus. The range of all boxes is between three and ten files, showing that no subset of multi-city GUIDs appear to contribute more significantly to the network.

In sum, most users stay relatively close to their home city, and the particularly multi-city GUIDs are no more aggressive than their single-city counterparts.

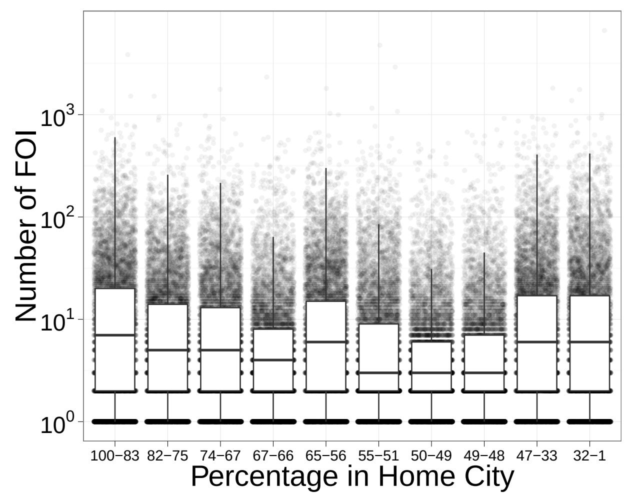



**Figure 14: A box plot showing the number of files that multi-city GUIDs are ever seen with. The groups are created by ordering the multi-city GUIDs by the fraction of time spent within their home city and then bucketing by 10% increments so that each box contains the same number of GUIDs. Each box represents the semi-interquartile range, with the middle line showing the median; underlying data is displayed as jittered semi-transparent points. No group is strikingly different from any other group.**

## C. CHURN

In this subsection, we evaluate the level of user *churn* in our data. Due to space limitations, we have placed these results in this appendix. Past works on characterization of churn [5, 7, 22, 26] reveal the highly dynamic nature of peer participation in p2p systems. However, there is noticeable difference in specific conclusions they reach, mainly due to the challenging nature of gathering unbiased data about peer participation [26]. To our knowledge, none of the previous studies evaluate data

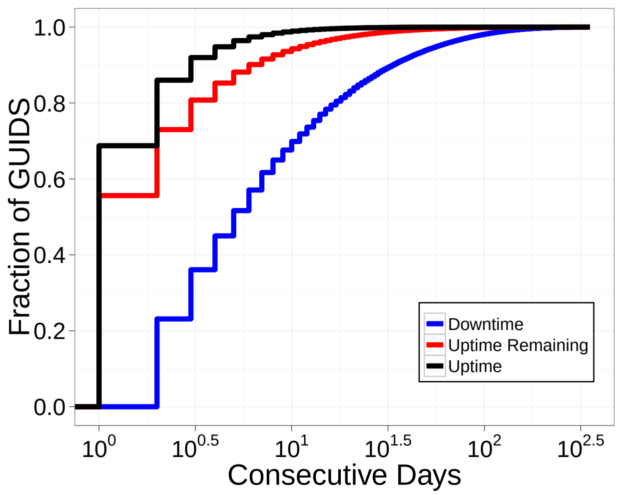



**Figure 15: CDF of consecutive days of up- and down-time for GUIDs, as well as per-day CDF of days remaining in each GUID's consecutive days seen. Downtimes are only counted when they occurred between the first and last day a GUID was observed. $> 68\%$ of GUIDs are not seen on consecutive days, and more than $> 51\%$ of GUIDs are not observed for 5 or more days at a time.**

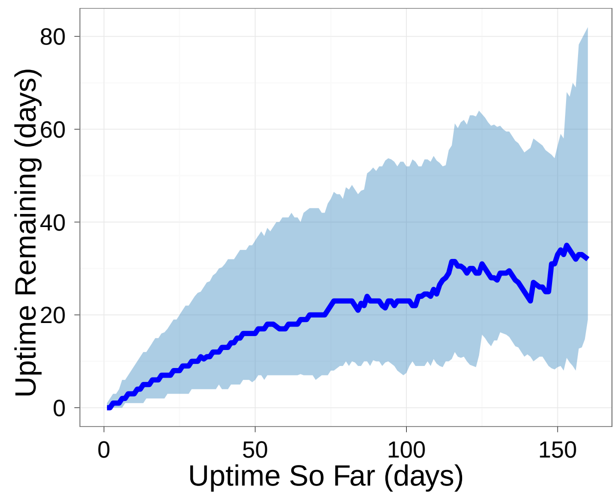



**Figure 16: Correlation of consecutive days observed at a given point to the median number of consecutive days remaining for each GUID. The shaded area represents the semi-interquartile range. Consecutive days observed are a good predictor of future uptime.**

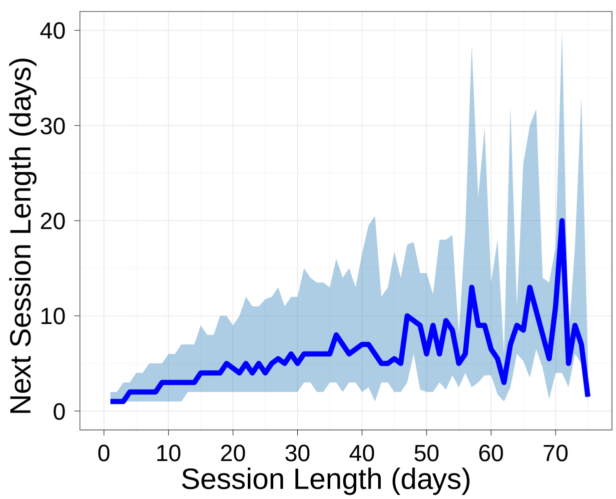



**Figure 17: Correlation of session uptime to the median uptime of the next session, where sessions are defined as consecutive days where a GUID was seen. The shaded area represents the semi-interquartile range. The correlation is weak, indicating that session lengths are likely independent.**

longer than a few weeks' duration. Nor have they evaluated churn of CP traffickers. Consequently, we find our analysis of churn insightful in spite of the limitations discussed in Section 2. In sum, we find in our dataset that while there is high churn, there are also many GUIDs which are consistently active in the network. We elide the analysis of churn in eMule.

Figure 15 quantifies the uptime and downtime of GUIDs with respect to consecutive days seen in the data. This graph shows that most GUIDs ($> 68\%$), when seen, are not seen on consecutive days, and that most GUIDs ($> 51\%$) are not observed for five or more consecutive days. This data implies that most GUIDs are only intermittently observed; however, at any given time, a significant fraction of GUIDs' uptimes are longer than a single day. Stutzbach et al. [26] also observe that while a randomly selected active peer is likely to have a long uptime, a randomly selected session is more likely to be short in the Gnutella network. Their analysis, however, is more granular yet covering a much shorter period.

Figure 16 shows the correlation of the number of consecutive days that a GUID is observed to the median number of consecutive days remaining for that GUID. The shaded area represents the semi-interquartile range. The results show that the number of consecutive days observed is a good predictor of future uptime. Previous studies [26] also show that while exhibiting high variance, uptime is on average a good indicator of the remaining uptime.

Figure 17 shows the correlation of session uptime to the median uptime of the next session, where sessions are defined as consecutive days where a GUID was seen. The shaded area represents the semi-interquartile range. The correlation is weak, indicating that session lengths are likely independent. This result is in contrast to what is reported by Stutzbach and Rejaie [26]. We attribute this difference mainly to the different file preferences of the users and longer observation period in our dataset.

## D. OTHER VISUALIZATIONS

In this appendix, we show several characterizations that expand upon figures in the main text.

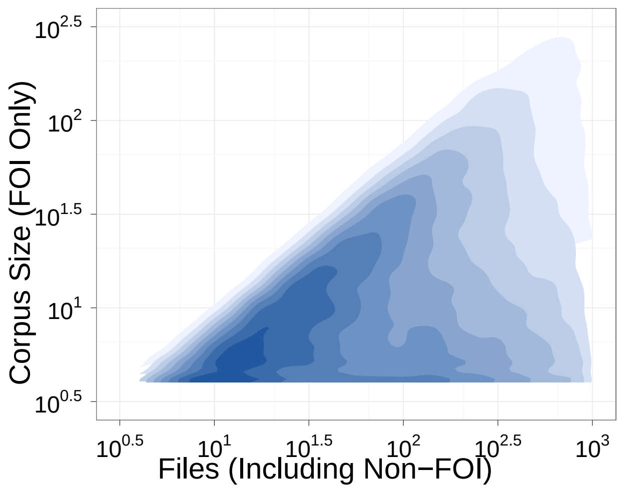
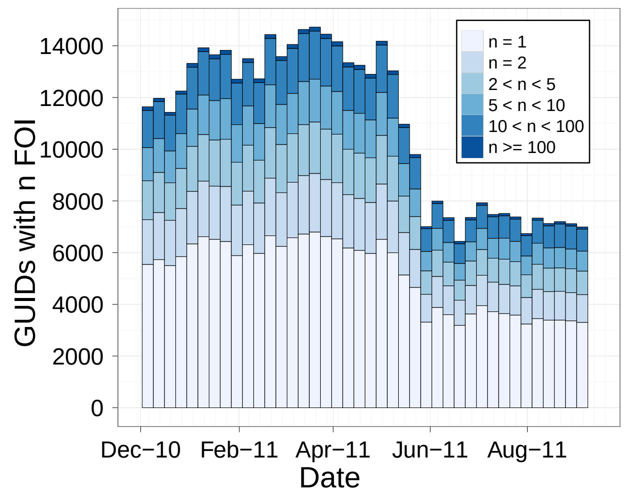
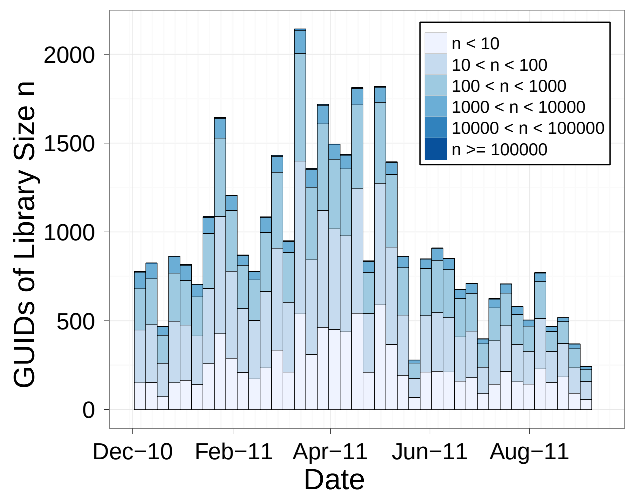



**Figure 18: Correlation between total number of shared files (FOI and non-FOI) and files of interest only, in Gnutella. Darker areas indicate a greater density of observations. A positive correlation between the quantities is present, but the large numbers of GUIDs with few FOI weaken it.**




**(a)** Counts of GUIDs possessing n files of interest.




**(b)** Counts of GUIDs with a library size of n.

**Figure 19: Number of GUIDs with (a) n files of interest, and (b) n total files, in Gnutella. Bar width represents one week of data.**

We evaluate the correlation between the corpus size of a GUID and the number of shared files of interest in Figure 18. In summary, the growth in total number of files (FOI and non-FOI) is weakly correlated with growth in the number of FOI; the fact that many peers possess one FOI weakens the correlation accordingly.

Figure 19 shows the library sizes of Gnutella GUIDs over time, rather than cumulatively (as in Figure 6). Figure 19(b) shows complete libraries, rather than only known FOI as in Figure 19(a). While we have fewer data points for complete libraries than for only known FOI, the data we do have indicate that complete libraries tend to be much larger on average, though with greater variability.