Characterization of Contact Offenders and Child Exploitation Material Trafficking on Five Peer-to-Peer Networks

George Bissias°, Brian Levine°, Marc Liberatore°,
Brian Lynn°, Juston Moore°, Hanna Wallach°*, and Janis Wolak†
°College of Information & Computer Sciences, Univ. of Massachusetts Amherst, USA
*Microsoft Research New York City, USA
†Crimes Against Children Research Center, University of New Hampshire, USA
{gbiss, brian, liberato, blynn, jmoore, wallach}@cs.umass.edu, janis.wolak@unh.edu

Abstract: *We provide detailed measurement of the illegal trade in child exploitation material (CEM, also known as child pornography) from mid-2011 through 2014 on five popular peer-to-peer (P2P) file sharing networks. We characterize several observations: counts of peers trafficking in CEM; the proportion of arrested traffickers that were identified during the investigation as committing contact sexual offenses against children; trends in the trafficking of sexual images of sadistic acts and infants or toddlers; the relationship between such content and contact offenders; and survival rates of CEM. In the 5 P2P networks we examined, we estimate there were recently about 840,000 unique installations per month of P2P programs sharing CEM worldwide. We estimate that about 3 in 10,000 Internet users worldwide were sharing CEM in a given month; rates vary per country. We found an overall month-to-month decline in trafficking of CEM during our study. By surveying law enforcement we determined that 9.5% of persons arrested for P2P-based CEM trafficking on the studied networks were identified during the investigation as having sexually offended against children offline. Rates per network varied, ranging from 8% of arrests for CEM trafficking on Gnutella to 21% on BitTorrent. Within BitTorrent, where law enforcement applied their own measure of content severity, the rate of contact offenses among peers sharing the most-severe CEM (29%) was higher than those sharing the least-severe CEM (15%). Although the persistence of CEM on the networks varied, it generally survived for long periods of time; e.g., BitTorrent CEM had a survival rate near 100%.*

Keywords: Child pornography, Peer-to-peer, Internet, Child sexual exploitation

1 Introduction

Possession and distribution of child exploitation material (CEM), also known as child pornography, is illegal for many reasons. CEM depicts child victims being sexually abused and exploited, and its proliferation causes lasting damage to the victims who must cope with knowing their images are being used for sexual purposes in a trade that is worldwide and beyond their control (Bazelon, 2013; Svedin & Back, 1996; von Weiler et al., 2010). Peer-to-peer (P2P) filesharing networks are perceived to be responsible for a large proportion of the growth in the availability of online CEM (Koontz, 2005). Although there are other means for obtaining CEM on the Internet, P2P networks are based on free software that is relatively simple to use for anyone with a computer. Users can easily find CEM because most networks have easy-to-use search capabilities. Law enforcement and policy makers believe that millions of CEM files are available through P2P networks, although little research has attempted to measure actual numbers (Hurley et al., 2013; Wolak et al., 2014).

This proliferation of CEM circulated via P2P networks has generated questions by law enforcement agencies working to combat the problem, and by policy makers and others who are working to protect victims, understand offenders, and reduce the amount of online CEM. First, although there has been speculation about the numbers of persons using P2P networks to distribute or access CEM, research is lacking that

We are grateful for the outstanding work of law enforcement in this area and for their assistance with this article, including: Corporal Robert Erdely of the Indiana County, Pennsylvania, District Attorney's office; Thomas Kerle of Fox Valley Technical College; Joseph Versace of the Ontario Provincial Police; and the Violent Crimes Against Children Section of the FBI. This work was supported in part by Grant No. 2011-MC-CX-0001, awarded by the Office of Juvenile Justice and Delinquency Prevention, U.S. Department of Justice. Points of view or opinions in this document are those of the authors and do not necessarily represent the official position or policies of the U.S. Department of Justice or the persons and agencies listed in these acknowledgements.

DOI: 10.1016/j.chiabu.2015.10.022
Preprint. To appear in *Elsevier Child Abuse & Neglect.*

systematically measures numbers of P2P peers sharing CEM, how numbers of such peers vary among different networks, and whether numbers are increasing over time. Reports of CEM trafficking are often based on counts of IP addresses, which is an inaccurate metric. Further, because the Internet and P2P networks are not limited by national borders, CEM distribution is inherently an international issue; measures of prevalence must be worldwide. Second, although there is evidence that a proportion of CEM distributors are also child molesters (Bourke et al., 2014; Seto et al., 2011), the question of what proportion of such "dual offenders" are identified in recent investigations of CEM distribution on P2P networks remains a concern to law enforcement agencies with limited resources, and who have questions about how best to use those resources to protect children. Third, there are concerns that easy access to online CEM may spur offenders to seek more extreme images, such as those that depict infants and toddlers or show sadistic abuse. Some have hypothesized that offenders caught with such severe images are more likely to also be child molesters, but there is not much research that addresses this question. Finally, advocates and policy makers who are particularly focused on the experiences of victims portrayed in CEM are concerned about the number of images that are online, and about how long such images remain accessible once uploaded to a P2P network.

These important questions have rarely been addressed because they are difficult to research. P2P networks are vast and used to trade many different types of material, including movies, music, and legal adult pornography. To examine research questions about CEM on P2P networks, researchers have to be able to identify CEM files, distinguish CEM by severity, distinguish and geolocate the computers that share those files, and measure the public activity of those computers for substantial lengths of time for a variety of P2P networks. Further, access to law enforcement data is required to address questions about offline child molestation among P2P CEM traders.

2 Literature Review

In this section, we review a number of related works that have characterized CEM file sharing, examined the ratio of offenders that are both online distributors of CEM and offline molesters of children, and automatically classified pornographic content.

Characterization of CEM on P2P networks. There have been several past works quantifying public CEM trafficking, For example, Hurley et al. (2013) analyzed CEM trafficking on the Gnutella and eDonkey networks from October 2010 to September 2011. The focus of that work was on the discovery of improved strategies for investigators to reduce the prevalence of CEM, the characterization of particularly aggressive peers on these networks, and the efficacy of the anonymizing tool Tor (Dingledine et al., 2004) in preventing effective investigation of peers. Wolak et al. (2014) presented a related analysis of the same data set, with a focus on the proliferation of CEM by peers located solely in the United States. They found that these peers accounted for a small fraction of the unique content available worldwide.

Other past work has characterized peers on P2P networks via their queries for CEM (and not offered content). For example, Latapy et al. (2013) collected queries on the eDonkey network for 10 weeks in 2007 and 28 weeks in 2009, each time from a single server. They used this data to characterize the fraction of queries to the server related to child sexual exploitation. Fournier et al. (2014) presented a comparative study on the prevalence of CEM-related queries on the eDonkey and Kad networks. In a related study, Steel (2015) examined the patterns of queries on web search engines for CEM, with two key findings: a substantial fraction of such queries were performed using mobile devices; and that search engine efforts to selectively block such queries were effective, decreasing the success rate of these searches.

Unfortunately, none of these works present data for CEM possession on Gnutella and eMule from 2012 onward, and none provide data on Ares, Gnutella2, or BitTorrent for any dates. None differentiate the proliferation of CEM images involving infants and toddlers or sadistic acts. Finally, there are no results that compare countries. All such information are valuable to law enforcement agencies managing strategies and resources to thwart these crimes.

One reason that past work may not have estimated the number of peers in Ares and BitTorrent is that these networks lack per-peer identifiers called GUIDs. These two networks present only IP addresses, which are frequently re-assigned by ISPs and re-used by many peers over time, inflating population counts as we show within. Liberatore et al. (2014) did examine the inequality between IP addresses and GUIDs, but in the context of a proposed digital forensics system that is analogous to marking bills used in undercover drug purchases.

Studies of contact offenders. A series of past works have examined the rate of contact offenses among CEM possessors at various stages of the legal process. For example, in a meta-analysis of research examining rates of contact offenses among CEM possessors, Seto et al. (2011) found that 12% had an officially known history of sexual offenses against minors at the time of arrest. Such estimates, including the study we present, are almost certainly low because they are based only on what investigators discover during the course of their investigations and arrests. Additional information may be revealed through later events, including psychological evaluation and treatment. Research based on self-reports by offenders charged with CEM possession has found much higher contact offense rates; in the same Seto et al. article, they found that 55% had admitted to contact sexual offenses.

Bourke and Hernandez (2009) found that of 155 participants that had received treatment at the Butner Federal Correc-

tional Institute, 85% admitted that they had committed hands-on sexual abuse, though only 26% had documented histories of contact offenses at the time of sentencing. In a separate study, Bourke et al. (2014) used a sample of 127 suspects with no known history of hands-on offenses, in which 4.7% admitted to sexually abusing at least one child. Through the use of polygraphs, an additional 52.8% of the study sample disclosed information about hands-on abuse that they had perpetrated.

It is useful and important to repeat these studies because investigative processes and venues where CEM is traded change over time. Further, none of these previous works distinguish dual offender rates per P2P network and none distinguish rates by the severity of content shared. These differences help shape strategies by law enforcement that must manage limited resources.

Classification of content. Various past works have examined the problem of automatic content classification. However, none apply such methods to the separation of Severe CEM (i.e., depicting sadistic acts and infants or toddlers) from other CEM.

For example, Panchenko et al. (2012) used language analysis in distinguishing pornographic galleries from control text drawn from random encyclopedia articles. Munson and Tsymbalenko (2001) proposed using filenames to automatically classify content in the context of web image searches. Rowley et al. (2006) presented a system using face detection and summary features to detect pornographic image files found by a web crawler. A more sophisticated system was presented by Deselaers et al. (2008), using an approach based upon a bag-of-visual-words (that is, higher-level features than skin tone detection) to improve classification results.

Peersman et al. (2014) developed a system for discovering new CEM, and integrated it into the iCOP toolkit. Their system is based upon features of files: *n*-gram substrings of the files' names, and various automated analyses of media (e.g., images and video), such as skin-tone detection. The works of Ulges and Stahl (2011) and Schulze et al. (2014) are complementary to that of Peersman et al., and examine in more detail the problems of media analysis in the presence of false positives (e.g., adult pornography) and more sophisticated approaches to prevent it. Inches and Crestani (2012) described the results of a 2012 competition for automated detection of sexual predators in online conversations based upon text analysis.

3 Purpose of the Study

Law enforcement tasked with addressing child exploitation implicitly pose a series of hypotheses on a daily basis: given limited resources, an effective method of rescuing children from contact offenders is one that focuses on a particular P2P network; there is a correlation between the severity of CEM content shared and the likelihood of detecting a contact offender; and content that is introduced on any one network tends to survive the churn of participants and even network decimation. These are fundamental hypotheses but also challenging to validate. The purpose of this study is to focus on the following related but more limited questions and hypotheses that are addressable given the observations available to us.

1. How many peers in these five networks are sharing known CEM, what are the trends, and how do the numbers vary by network and country?
2. What proportion of users arrested in the United States for trafficking CEM on P2P networks were identified during investigations as having committed contact offenses against children; how do the proportions vary by network; and is the proportion higher among peers sharing Severe CEM (i.e., depicting sadistic acts and infants or toddlers)?
3. What is the prevalence of known Severe CEM on these P2P networks over time?
4. How many known CEM files are being shared on these P2P networks over time; and what is the introduction and survival rates of CEM files on these networks?

These questions concern both law enforcement and victim advocates, and we address them in four analyses. Our first and second analyses inform law enforcement's management of limited training and enforcement resources according to factors such as jurisdiction, network, and the type of shared content. The third analysis measures the availability of Severe content, which is a factor in sentencing guideline discussions (United States Sentencing Commission, 2012) and, as our second analysis suggests, may be a factor in discovering contact offenders. The fourth analysis concerns the growing availability and survival of CEM, which is important for informing victim restitution, as well as informing law enforcement strategies aimed at reducing content availability.

We have worked with law enforcement to develop software that is used in proactive investigations of CEM trafficking on P2P networks. This software, provided at no cost, takes advantage of the fact that activity on P2P networks is public and highly visible. It detects CEM files already known to law enforcement from previous investigations, and we are able to distinguish content by severity. The software logs IP addresses and other public identifiers of computers that access CEM across five widely used P2P networks, the geographic region of the IP addresses, and dates and times the CEM was publicly shared. In particular, to address the questions raised in this article, we use three years of logs of public activity on BitTorrent (Cohen, 2003); eDonkey, including Kad (Kulbak & Bickson, 2005); Ares Galaxy (lap3k, n.d.); Gnutella (Klingberg & Manfredi, 2002); and Gnutella2 (Stokes, n.d.). Note that software for Ares was developed by Joseph Versace of the Ontario Provincial Police.

4 General Method

4.1 Distinguishing Users on P2P Networks

The eDonkey, Gnutella, and Gnutella2 networks distinguish instances of the P2P software by assigning a globally unique identifier (GUID) to the application when it is initially installed on a device. GUIDs remain consistent even as the IP address changes. While GUIDs are not one-to-one with people (who may control multiple computers), they provide a reasonable measure for differentiating devices that are trafficking CEM on a P2P network (Liberatore et al., 2014).

Unfortunately, BitTorrent and Ares do not use GUIDs. Multiple observations of a specific IP address do not necessarily indicate a distinct device or user over time because of several technologies that are commonly used. Network Address Translation (NAT) is a common mechanism by which multiple devices are contemporaneously routed through a single IP address. Dynamic Host Configuration Protocol (DHCP) is used by Internet Service Providers (ISPs) to temporarily assign an IP address to a device and then re-assign it to another device. Finally, like NAT, the effect of proxy servers, virtual private networks (VPNs), and the Tor network is to allow many participants to be masked by a single IP address. Although these technologies may not protect them from being observed by the public and law enforcement, it does affect the count of IP addresses. (For more on the limitations of Tor against law enforcement, see Hurley et al. (2013).)

4.2 Gathering Data on P2P Networks

In studying the five P2P networks, we collected data pertaining to files that had been identified by law enforcement agencies during trafficking investigations. The great majority, though not all of these files, met legal definitions of child pornography; for example, files that were part of a series of images depicting an initially clothed child who was sexually abused in subsequent images. We collectively refer to the identified files as *known CEM*. A unified set of known CEM was maintained across the five P2P networks. To identify files that have the same content (byte for byte), despite the same or different names, P2P networks make use of cryptographic hashing algorithms (e.g., MD4, MD5, and SHA-1).

To gather data on the P2P networks, we used instrumented P2P software to record over a billion observations of known CEM being trafficked on the networks. The instrumented software used the P2P networks' native discovery mechanisms for locating known CEM files, and for obtaining public information that distinguished remote clients (e.g., IP addresses and GUIDs). We did not always individually query each IP address, but often relied on the information provided by the P2P networks.

The instrumented software only used protocols as defined by each P2P network; we did not use special technology to gain unauthorized or special access to any peer participating in a P2P network. The instrumented software took as input cryptographic hash values identifying known CEM files, and the digests were used to restrict recorded observations exclusively to known CEM. When a known CEM file was detected, our instrumented clients would record the file's digest, time of observation, P2P network, and relevant network- and application-layer identifiers (e.g., IP address or GUID).

The size of our list of known CEM was not constant during our study, and files were added unevenly throughout. A file had to be in our known CEM list at the time of the observation for it to have been recorded. At the start of the study, in June 2011, the list included approximately 325,000 files uniquely identified by hash value; by the end of the data collection, in December 2014, it included over 2.5 million files. We are reporting on only a fraction of CEM available (NCMEC, n.d.).

Commercial geolocation software from MaxMind was used to map each IP address to a country and ISP (not without error). The geolocation software does not provide street addresses or any other personally identifying information. If a peer used anonymization technology to mask its actual IP address, such as a proxy or VPN, we make no attempt to identify the peer's actual IP address.

5 Analysis 1: How Many Peers in the Five P2P Networks are Sharing CEM?

In this section, we quantify how many peers were sharing CEM files on the five P2P networks. The most straightforward method would be to count distinct IP addresses, but unfortunately doing so tends to over-represent the number of peers. GUIDs provide a more accurate count, but not all P2P networks use GUIDs. Below, we present an estimate of the number of GUIDs on the two networks where they are absent, BitTorrent and Ares, by using per-ISP GUID-to-IP ratios, realizing a more representative per-network count of peers.

5.1 Method

We identified the country of origin by geolocating the IP address. If the IP address had been obfuscated through the use of an intermediary, such as a proxy, the intermediary's location was instead logged. A P2P network must have identified an IP address as either sharing or attempting to locate one of our known CEM files for it to have been logged.

We estimated the GUID counts for BitTorrent and Ares by first calculating the ratios of GUIDs-to-IP addresses actually observed for eDonkey, Gnutella, and Gnutella2 on a *per-ISP basis*. We calculate the per-ISP ratios because we expect different ISPs to employ different strategies for allocating IP addresses; e.g., a cellular ISP versus a business ISP. Our estimate of BitTorrent or Ares GUIDs present for a given ISP in a given month is equal to the calculated GUID-to-IP ratio for the ISP multiplied by the number of IP addresses observed

in that month for BitTorrent and Ares, respectively. The total number of GUIDs that month is the sum of all ISP-based GUID estimates. Our method for calculating the estimated ratio for each ISP, and an estimation of the error, is described in the supplementary materials.

For the figures presented in this and subsequent sections, points show the number of observations and shaded lines are a least-squares fit to the points using the LOESS algorithm. The width of the LOESS lines represent a 95% confidence interval.

5.2 Results

The number of distinct IP addresses worldwide sharing known CEM across the five networks was 1.7 million in December 2014, which is down from 3.2 million in September 2012. These numbers correlate with but greatly overestimate the number of unique installations of software sharing known CEM on these networks. We estimate the total number of GUIDs sharing known CEM on these networks as 840,000 in December 2014, down from 1.3 million in September 2012. Under an assumption that GUIDs and users are one-to-one (though we expect there are more GUIDs than users), we estimate that about 3-in-10,000 Internet users worldwide were sharing known CEM on these five P2P networks. We also provide per-country statistics in this section. For example, 2-in-10,000 users in the United States, 11-in-10,000 users in Spain, and 13-in-10,000 users in Argentina were sharing known CEM on these five networks. Of about 840,000 P2P software installations worldwide in December 2014 sharing known CEM, about 56% were from just four countries: China, Brazil, Mexico, and the United States.

5.3 Discussion

Clients by IP addresses. Figure 1(left) and Figure 2(left) show, respectively, the number of unique IP addresses that were observed worldwide and in the United States, per month, for almost three years of observations. For 2014, an average of 1.9 million distinct worldwide IP addresses were observed per month. In the United States, we observed an average of 54,000 distinct IP addresses per month during 2014. Worldwide data is meaningful even when assessing the scope of CEM within a single country; a user in the United States can download CEM from users throughout the world.

The top line in each of the two graphs is the total number of distinct IP addresses for all five P2P networks. If an IP address was observed on more than one network, it is counted only once in the top line. Overall, for the five networks, the number of distinct IP addresses is dropping, with only BitTorrent increasing in worldwide popularity among CEM traffickers.

Clients by GUID. In Figure 1(right) we show the per-month average of GUIDs per network worldwide; Figure 2(right) shows the same data for the United States only. For the three networks supporting GUIDs—eDonkey, Gnutella, and Gnutella2—the graph shows actual data. For Ares and BitTorrent, we extrapolated from observed GUID-to-IP address ratios obtained from the other networks as described above. Worldwide, there were about half the number of GUIDs as IP addresses. In the United States, however, the number of GUIDs was 80% the number of IP addresses.

Similar to the IP data, the GUID graphs show an overall decrease in P2P clients. We estimate that the worldwide total number of GUIDs sharing known CEM was 840,000 in December 2014, versus 1.3 million in September 2012. Ares, BitTorrent, and eDonkey were the most popular, together comprising over 95% of the GUIDs on all five networks since September 2012. In the United States there was greater variance in the popularity of these top three networks; they constituted at least 81% of GUIDs since September 2012, but have contributed at least 89% of the GUIDs since January 2014.

Although the trends are downward, these five P2P networks are not the only means for obtaining CEM on the Internet. Moreover, the population of CEM traffickers for the five networks alone overwhelms the number of trained law enforcement agents who can address these crimes.

Table 1 shows per-country GUID counts for each network, as a per-month average during 2014. The right-most column shows the percentage of the country's population of Internet users (based on Internet Live Stats, 2014) in each country if GUIDs were one-to-one with users. The table is limited to the top 10 countries ranked by total number of observed and estimated GUIDs. (Additional countries can be found in Table 6 of the supplementary material, ranked by per-capita population.) The relative popularity of P2P networks was not uniform across countries. For example, 87% of the P2P clients in Italy were observed using eDonkey, while 95% of the clients in Mexico were observed using Ares. In China, 65% were observed using BitTorrent and 35% were observed using eDonkey. These disparities should be factored into per-country strategies for addressing CEM on P2P networks.

6 Analysis 2: What Proportion of US Users Arrested for CEM Trafficking Were Identified as Sexually Offending Against Children Offline?

We surveyed law enforcement investigators to assess the percentage of cases they investigated on P2P networks that resulted in the detection, *during the investigation*, of a CEM trafficker who had also committed contact sexual offenses against children. In this section, we present the results of the survey, including the rates for each P2P network. For the BitTorrent network, we also present the number of detected contact offenders by severity of the known CEM they were observed sharing.

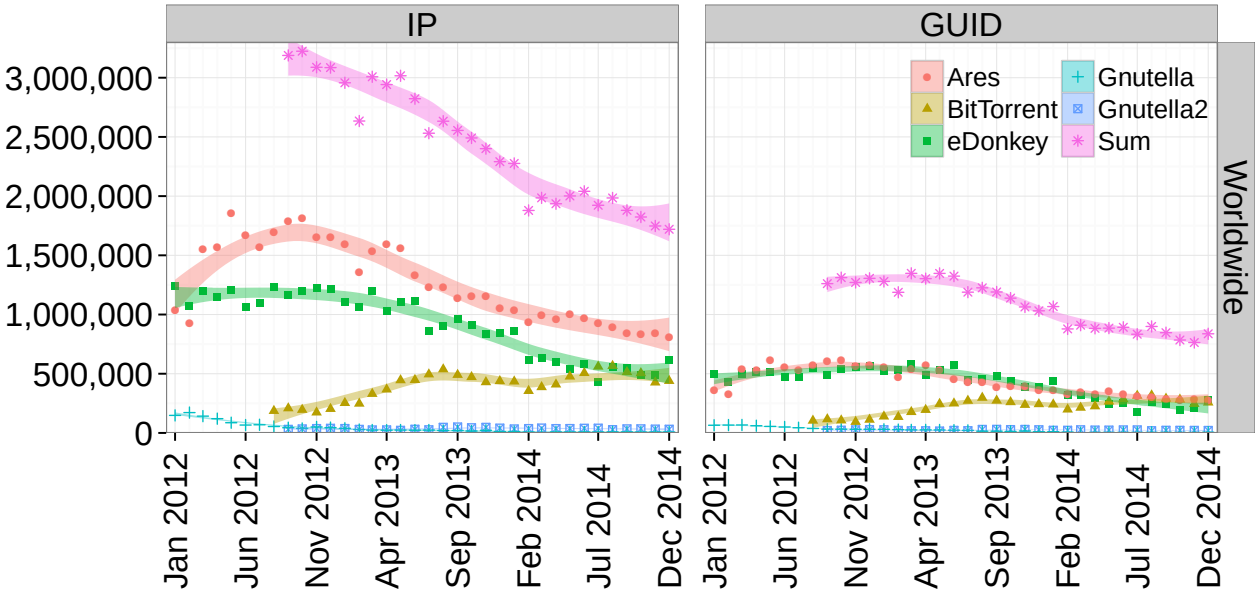

Figure 1. Worldwide number of observed IP addresses (left) and GUIDs (right) sharing known CEM in five P2P networks. For all figures, points are exact counts or estimates; shaded lines represent a LOESS-based fit to the points, where width is the 95% c.i. Ares and BitTorrent GUID values are estimates. Values for Sum do not start until all networks appear in the plot.

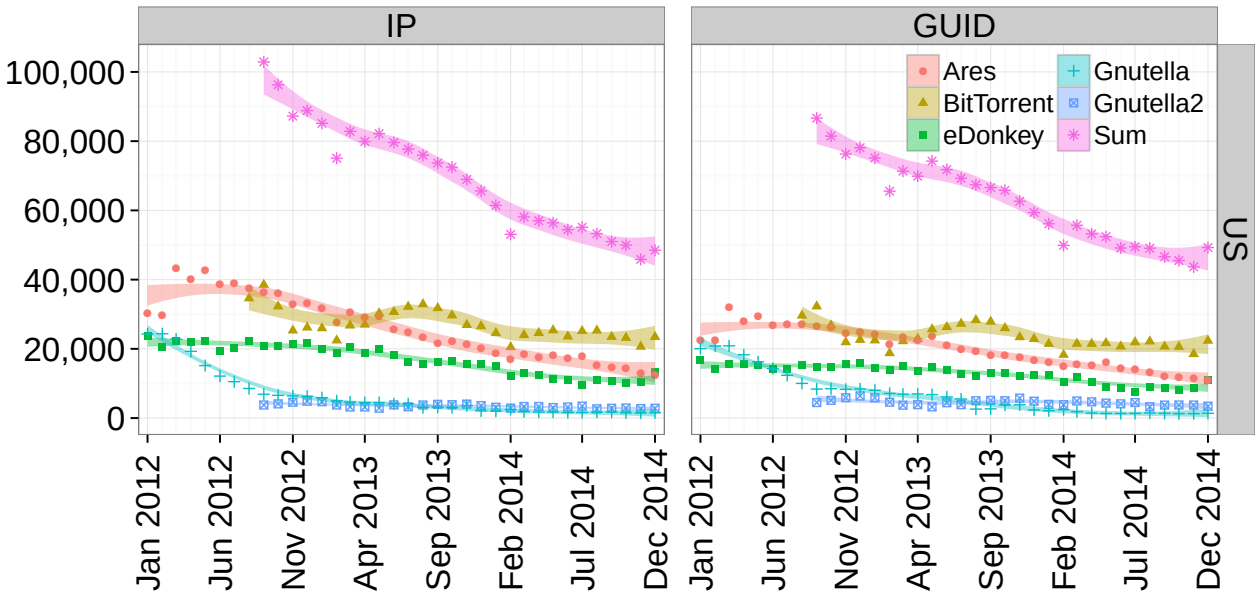

Figure 2. The same values as Figure 1 for peers geo-located to the United States only. See also Figure 7 in supplementary materials.

Table 1
Per-month average of distinct GUIDs per country sharing CEM.

Country[a]	eDonkey	Gnutella	Gnutella2	Ares[b]	BT[b]	Total[c]	% Pop.[c]
Germany	11,864	78	2,012	1,940	2,760	18,656	0.03
Colombia	551	11	28	19,568	240	20,398	0.08
France	16,880	164	1,710	3,006	2,662	24,423	0.04
Argentina	2,573	13	82	29,693	693	33,053	0.13
Italy	33,352	43	174	2,939	1,667	38,175	0.10
Spain	18,201	14	58	19,187	1,139	38,599	0.11
USA	9,667	1,555	4,031	13,824	20,934	50,010	0.02
Mexico	2,038	46	242	57,456	884	60,666	0.12
Brazil	10,695	234	2,398	61,164	4,076	78,566	0.07
China	92,954	8	352	200	163,668	257,181	0.04
All regions	247,398	4,788	19,504	290,302	244,916	806,909	

[a] Only countries with at least 5 million Internet users are listed. [b] BitTorrent and Ares counts are estimates as described in the supplementary materials. [c] This column uses a sum total of GUIDs that double-counts entries appearing in multiple networks.

Table 2
Survey results of US P2P CEM distribution arrests that also involved identifying contact offenders.

	Round 1 10/2008–6/2011		Round 2 7/2011–9/2012		Round 3 9/2012–7/2013		Total 10/2008–7/2013	
Cases and surveys	%	n	%	n	%	n	%	n
Cases opened		6,946		5,306		5,812		18,064
Ineligible[a]		89		200		208		497
Eligible cases	100	6,857	100	5,106	100	5,604	100	17,567
Surveys not completed[b]	27	1,855	23	1,183	34	1,908	28	4,946
Duplicate cases	1	84	1	34	0	12	1	130
Completed surveys	72	4,918	76	3,889	66	3,684	71	12,491
Not contact offender[c]	93	4,586	89	3,462	88	3,258	91	11,306
Contact offender[c]	7	332	11	427	12	426	9	1,185

[a] Ineligible cases were opened for training or other non-investigative purposes. [b] Includes no response, unable to reach investigator, refusal, case could not be identified, pending at close of survey.
[c] Percent of completed surveys.

6.1 Method

Surveys. We obtained a list of P2P CEM trafficking arrests from a database law enforcement investigators used for managing P2P case deconfliction. The information we collected from the deconfliction database included a case identifier number; IP address and its location; investigator name, department, telephone number and email address; date and time the case was opened; P2P network and GUID (if applicable and known); and in some cases, outcomes (e.g., search warrant executed, arrest, contact offender identified).

To augment the information in the deconfliction database, we surveyed investigators about their cases. We performed three rounds of surveys, covering cases opened roughly (with some overlap) in the months of October 2008 through June 2011, July 2011 through September 2012, and September 2012 through July 2013, respectively. We emailed letters to agency heads and investigators explaining the study. In the emails to investigators, we identified their cases and included a short survey asking them to specify any cases that resulted in the detection of a contact offender.

We defined contact offender to include past or current commission of sex crimes against minors involving physical contact (e.g., sexual molestation or assault) and other criminal sexual interactions (e.g., online enticement, production of child pornography). If an investigator did not respond to our initial request, we sent reminders two and four weeks after the initial mailing. If an investigator did not respond after the second request, we followed up with a phone call. Responses were recorded in a locally maintained database.

We obtained approval from our Institutional Review Boards (IRB). All required IRB procedures and mandated privacy regulations were followed by the researchers.

Data processing. We examined whether the proportion of CEM traffickers identified as contact offenders varied by P2P network The investigators specified a P2P network in about 10% of the survey responses. Using the P2P field from the deconfliction database and additional information that was provided, such as the GUID or IP address, we were able to identify the network that was investigated in most of the cases; but we were unable to identify the network in 390 of 12,491 cases (3%). For a very small percentage of the cases (about 0.1%), investigators identified multiple networks; in those instances we attributed the case to each network. Almost 90% of the cases were either Gnutella (61%) or Ares (28%). The large number of Gnutella and Ares cases is in part due to the number of law enforcement agents trained on each network.

File severity. In the BitTorrent P2P network, peers use *torrents* to advertise files that they are sharing. Torrents are not content themselves; they describe a collection of files using hash values of the content. In order to identify CEM shared on BitTorrent, law enforcement must first locate torrents then evaluate their relationship to known CEM. We examined whether offenders who trafficked torrents with more severe content (i.e., images of sadistic content, and infants or toddlers) constituted larger proportions of identified contact offenders.

It is a significant challenge to obtain, for a given set of CEM, a labeling of the content by severity. Such files are contraband and only law enforcement can evaluate the content. We were therefore grateful to obtain from law enforcement a set of content labels for the torrents that were the basis of many of the cases in our survey. Because not all torrents used in this article were labeled, we could perform our analysis on only a subset of our data. Specifically, we analyzed 317 BitTorrent cases (identified by IP address), where the IP address was observed sharing torrents that had been assigned a severity level by law enforcement.

Law enforcement determined content and assigned labels based on the entire collection of files described by a torrent. The labeling was designed and implemented by law enforcement independent of our study and its goals; we were unable to inspect images or make use of a typology from past work. The basis of the labels was visual inspection. The labeling system consisted of eight distinct labels or levels, ranging from Level 1 (least severe content, but still CEM) to Level 8 (most severe). In addition to the level assigned to a torrent, law enforcement agents added descriptions of the content referenced by the torrent. This was used to ensure the consistency of levels across torrents. Though we did not view the content, we did review the descriptions.

Level 8 torrents constituted a very small percentage of the overall torrents. Each Level 8 torrent consisted of a sizable collection of CEM depicting the sexual assault of infants or toddlers. Levels 5–7 were assigned mostly to torrents whose files depicted prepubescent children engaged in sexual acts, with sadistic acts depicted in several of the Level 7 torrents. A very small number of the files in the Level 8 torrents also appeared in a few of the Level 5–7 torrents, but the emphasis of those torrents was not on infants or toddlers. Torrents labeled as Levels 1–4 were deemed the least severe, and generally included child nudity (including those with an emphasis on genitals), masturbation, sexual play, or adolescent sex.

6.2 Results

Of the 18,064 entries in the deconfliction database, 17,567 were for actual P2P CEM arrests. We surveyed the law enforcement investigators who had made the arrests, and asked the investigators in which of the cases the investigation identified a contact offender (for that case or previously). The surveys covered cases opened between October 2008 and July 2013. Combining all three rounds of the survey (Table 2, columns 8 and 9), of the 17,567 cases where an arrest was made, we received 12,491 responses (71%). We found in 1,185 out of the 12,491 completed surveys (roughly 9.5%), investigators reported that the CEM possessors were also contact offenders. In other words, approximately 9.5% of persons arrested for CEM possession were identified during the course of their investigation as having committed a past or present contact sex crime against a child.

We found differences in the proportions of contact offenders among the networks. For cases where we were able to identify the P2P network, contact offenders were identified in 593/7,640 (8%) of Gnutella arrests, 352/3,537 (10%) of Ares arrests, 13/107 (12%) of Gnutella2 arrests, 76/492 (15%) of eDonkey arrests, and 70/337 (21%) of BitTorrent arrests (Table 3).

In 317 of the 337 BitTorrent cases, we were able to associate the IP address with one or more torrents that had been assigned a severity level by investigators. Of these 317 cases, 67 (21%) were identified as contact offenders during investigation. We evaluated the percentage of identified contact offenders by severity level. Table 4 shows the results by torrent severity levels. The center column is the number of cases we observed for each severity level. The third column shows the number of contact offenders identified within the level, and the percent of that severity level's population identified as contact offenders at the time of arrest.

The difference in percentages is significant (p-value=0.02) between peers sharing the most severe (Level 8) content (28.8%) depicting sexual assault of infants or toddlers, and peers sharing less severe (Level 1–4) content (15.4%) depicting scenarios of child nudity without a focus on infants and toddlers or sadistic acts. Offenders can appear in multiple rows of the table; for example, if an offender shares Level 4 and Level 8 torrents, they'll appear in counts for the first and last rows.

Table 3
Contact offenders found by investigators by network.

		Contact offenders		Less than[a]	
Network	Surveys	n	%	eDonkey	BitTorrent
Gnutella	7,640	593	7.8	Yes[*]	Yes[*]
Ares	3,537	352	10.0	Yes[*]	Yes[**]
Gnutella2	107	13	12.1	No[****]	Yes[***]
eDonkey	492	76	15.4		Yes
BitTorrent	337	70	20.8		
Unidentified	390	98	25.1		

[a] States whether the percentage of contact offenders identified for the network is statically less than the offenders identified for the eDonkey or BitTorrent networks.
[*] p<0.01 [**] p<0.02 [***] p<0.03 [****] p>0.05

6.3 Discussion

Contact offenders. In addition to targeting a serious crime, CEM investigations are a valuable means of detecting contact offenders. Many such offenders who likely would go undetected are revealed through investigations of CEM possession. Even when CEM investigations fail to uncover contact offenses, they document offenders' proclivities and thus strengthen cases if CEM possessors are investigated for contact offenses at later dates. But it is also likely that many investigations are not identifying contact offenders that have been arrested for CEM trafficking. We found that 9.5% of those arrested were identified during the investigation as contact offenders. This is different from saying that 9.5% of those arrested were contact offenders. Bourke and Hernandez (2009), Bourke et al. (2014), and Seto et al. (2011) found offender rates of 85%, 53%, and 55% respectively. It is unrealistic to expect all contact offenders to be detected during an investigation; but it does raise the question as to what information and resources could make law enforcement more effective in identifying contact offenders and their victims.

Findings by network. Three networks—BitTorrent, eDonkey, and Gnutella2—each had percentages of identified contact offenders higher than the average with 21%, 15%, and 12% respectively (Table 3); the average of the three networks was 17%. But each of these networks had much smaller sample sizes. We cannot conclude that any of these networks have a higher percentage of contact offenders in their overall population of users because investigators did not take random samples, and not all contact offenders are identified during the investigation. Rather, the higher percentages more likely indicate the effectiveness of investigators' approaches to finding contact offenders. The "less than" column shows the list of networks where the percentage column is statistically different (i.e., p-values are less than 0.02 in all cases listed according to a one-sided two-sample permutation test of proportions).

The percentage of identified contact offenders in the unidentified network category is high at 25%, but would have no real effect if identified as Gnutella or Ares (which is likely the case). It is also worth noting that the percentage of identified contact offenders increased in each successive round of surveys: 7%, 11% and 12%, respectively. This increase may be due to increased effectiveness of law enforcement at identifying contact offenders.

Table 4
A comparison of surveys for cases in BitTorrent where the level rating of the CEM content was known.

		Contact Offenders	
Content Category	Surveys	n	%
Level 1–4 CEM	91	14	15.4
Levels 5–8 CEM	215	47	21.9
Levels 6–8 CEM	200	46	23.0
Levels 7–8 CEM	139	33	23.7
Level 8 CEM[a]	80	23	28.8

[a] The difference in percentages between contact offenders sharing Level 1–4 and Level 8 content is significant (p-value=0.02); differences between other levels are not.

Relationship of contact offending to content of CEM files. Table 4 suggests that an increase in severity corresponds to an increase in the percentage of identified contact offenders; however, we examined only a relatively small sample for this preliminary result. Only the difference between the percentage of identified contact offenders sharing Level 1–4 content (15.4%) and Level 8 content (28.8%) is a statistically significant difference (p-value=0.02, using a one-sided permutation test of proportions). To our knowledge, we are first to report data on a possible link between the type of CEM content (distinguished by age of the child) and identification of contact offenses. We hope that further study by us and others will confirm these results. In particular, these cases involved varied processes from different agencies and juris-

dictions; Bourke et al. (2014) showed, for example, that the use of a polygraph can affect results. Another explanation of the differences between Levels 1–4 and Level 8 content could be that investigators work harder to uncover contact offenses at the time of arrest when content is more severe.

7 Analysis 3: What Percentage of Files Available Are Severe Content?

In this section, we characterize the number and percentage of severe files on the five networks using a data set of files manually categorized by law enforcement. Additionally, we present a classifier that can automatically segment CEM by severity by evaluating filenames, and use it to characterize files of unknown severity on each network.

7.1 Method

Severe CEM files. To develop our method, we obtained a set of cryptographic hash values corresponding to known CEM files, along with their associated filenames, which were curated by our law enforcement partners for reasons external to this article. Each hash value was labeled with one of the following (mutually exclusive) classes: Age Difficult, Child Exploitation Material, Non-Pertinent, Infant/Toddler CEM, Obscenity, or Sadistic CEM. All but the Non-Pertinent and Obscenity classes would generally be considered child pornography.

As with the levels in Analysis 2, we chose the Severe category to consist of files from either the Infant/Toddler or Sadistic classes. In total there were 14,965 manually labeled Severe files. The Not Severe category is made up of the remaining classes, which included 925,995 manually labeled files. Note that Not Severe files still include files considered CEM. The labeling of torrents was done independently of the manual categorization used here, and consistent definitions are not assumed. We matched this set of manually labeled CEM hash values to observations logged by the tools described in the General Method section. Since the same CEM file may be associated with multiple filenames, we constructed a *filename set* for each file. On average, each filename set contained 1.48 distinct filenames.

Automatic file categorization. Most of the files in our known CEM list have not been manually labeled by severity, and labeling all files through visual inspection is time consuming because of the size and sensitive nature of the files. As an alternative, we chose to automatically label files using their filenames rather than content. We considered all distinct filenames during labeling by converting the filenames in each filename set to lowercase, removing file extensions, and then concatenating them (separated by spaces) into a single string called the *filename string*. Each file's filename string was subsequently divided into (non-unique) word *tokens* by splitting on non-alphanumeric characters. We showed the 25 most frequent word types (i.e., those associated with the largest numbers of tokens) for the Severe and Not Severe categories to our law enforcement partners, who confirmed that these types correspond with domain knowledge about these categories.

In order to automatically separate Severe from Not Severe files, we experimentally evaluated the performance of three classifiers: logistic regression, and multinomial and Bernoulli naïve Bayes (Manning & Schütze, 1999). These classifiers differ in the way they model the data, and hence in the criteria that they use for classification. In the end, we used only logistic regression (LR), which performed best among the three. The details of the performance of our LR classifier is presented in supplementary materials.

We analyzed the P2P network observations recorded between August 2011 and December 2014 to determine the prevalence of Severe known CEM files on the five networks. We started with the subset of CEM files that had previously been manually labeled as Severe or Not Severe by law enforcement. These manually labeled files were used to train our LR classifier, which was subsequently applied to all remaining (unlabeled) CEM files. The classifier output was used to estimate the probability-of-severe for each unlabeled file; and because LR produces probabilities, they can be summed to yield an estimate of the expected number of severe files. For that reason, using the classifier to estimate the count of severe files is a significantly easier task that determining whether a specific file is severe. Indeed, the sums were quite accurate; the *mean absolute percentage error*, calculated as the average of values ($|estimated - actual|/actual$) across 50-fold test-and-train validation of the labeled set was just 0.97%.

7.2 Results and Discussion

Figure 3 shows the number of Severe files over time for both the *Complete* (red) and *Manual* (blue) sets. For the Complete set, the Severe file count is the number of manually labeled files plus the sum of probability-of-severe for all classifier-labeled files. For the Manual set it is just the total number of manually labeled Severe files. Note that Severe file ratios (the plot on the right) are equivalent to the Severe file count divided by total known CEM files. The total known CEM files include only manually labeled files for the Manual set, and both manually labeled and classifier-labeled files for the Complete set.

For Manual files, Figure 3(left) shows that BitTorrent had the greatest count of manually labeled Severe files (blue curve) by December 2014. In contrast, the same figure shows that eDonkey had the most Severe files when considering Complete counts (red curve). In every network, the classifier revealed as many as 750 Severe files each month that were not in the Manual set. In some cases, specifically Gnutella, Gnutella2, and Ares, these additional Severe files more than double the total number of labeled (manual or classifier) Severe files for some months. The Complete and Manual curves

have both generally increased from January 2012 until December 2014 across all networks except Gnutella, where the number has steadily declined. This implies that the number of Severe files has tended to increase in most networks even without the addition of classifier-labeled files.

Although the count of Severe files has declined in the Gnutella network, Figure 3(right) shows that those Severe files comprise the greatest fraction relative to all known CEM files across all networks. The BitTorrent and eDonkey networks have experienced no significant change in the fraction of Severe files between January 2012 and December 2014. In fact, because the number of Severe and Not Severe files fluctuated nearly in unison in the Complete and Manual sets, both show very similar Severe ratios. For Gnutella2 and Ares, there is a significant decrease in the fraction of Severe files over time. In early 2012, Severe files comprised 25% of Ares CEM files. By late 2014, this percentage had decreased to around 7%. Over the same time period, the percentage of Severe files on Gnutella2 also decreased significantly, though less dramatically, from roughly 15% to just over 10%.

8 Analysis 4: For How Long Does Known CEM Persist on P2P Networks?

In this section, we examine the availability of known CEM over time on the five P2P networks. We quantify how many known CEM files are available on each network on a per-month basis. We also quantify how long CEM stays on the networks once introduced; and whether the availability of files is increasing despite the decrease in clients discussed in Analysis 1.

8.1 Method

We used the data that was gathered as described in the General Method section. Since our data set was limited to observations of known CEM, the measurements are a lower bound. We consider a file to be available on a network in a given month if it was observed to be associated with at least one IP address during that month. CEM files are identified by their cryptographic hash value, as described in the General Method section, and thus files that differ only slightly (for example, due to cropping or resizing) are counted as separate files.

8.2 Results

Worldwide, the total count of distinct known CEM files was 161,000 during the last month of our study, December 2014. We observed 122,000 distinct known CEM files being shared by clients located in the United States during this month. While the total number of known CEM files publicly advertised for download across all P2P networks is increasing, within individual networks, the longevity and month-to-month availability of known CEM is more variable. For the three most popular networks identified in Analysis 1—Ares, BitTorrent, and eDonkey—at least 80% of the known CEM files initially observed in the first month of study were still available in December 2014. However, we have begun to observe a decline in Ares' month-to-month file availability. Our data set did not allow us to determine when files were initially introduced to a network, but an examination of active BitTorrent torrents indicates that a majority may have been introduced to the network in 2010 or earlier.

8.3 Discussion

Absolute counts of known CEM files. In Figure 4, we show the number of unique known CEM files made available on the five P2P networks. The "All networks" total counts a file only once, even if the file was observed on multiple networks during the same month. BitTorrent quickly dominates the "All networks" category after it enters the data set. Figure 4 shows both monthly (left) and cumulative (right) counts. The number of known CEM files shared on all networks increased considerably from September 2012, the first month with data for all networks, to December 2014. In the United States, the total number of known CEM files shared each month across all networks increased from about 42,000 to 122,000 over this time period. Worldwide, the number increased from roughly 59,000 to 161,000. There are at least two explanations for this consistent increase.

The first explanation is that, over time, the set of known CEM files has increased by nearly a factor of eight. The dates of these additions to the set are not always made known to us, though we can infer them on the basis of large upticks in the per-month counts. However, because the per-month file introduction trends vary between the networks, it appears that there is not a one-to-one increase in files added to the known CEM list and observed CEM files.

The second explanation for the consistent increase in the number of known CEM files observed over time is that users are adding new content to the networks. In fact, users often add a file to a network before investigators become aware of it; new lists of identified CEM are generated in part from the contents of computers seized after P2P CEM trafficking investigations. Unfortunately, we don't know when files were first shared to a network. Accordingly, these figures represent a rough lower bound on the number of CEM files shared.

Long-term file survival and introduction rates. In this section, we examine the survival rate of known CEM across our entire set of observations. We define *survival rate* across a time period as the fraction of files observed at the start of the time period also available at the end of the time period. Figure 5(left) shows that for the three most popular P2P networks—BitTorrent, Ares, and eDonkey—the survival rate across the period of the study equals or exceeds 80%. The survival rates are lower for the less popular Gnutella and Gnutella2 networks, at about 12% and 49%, respectively. The

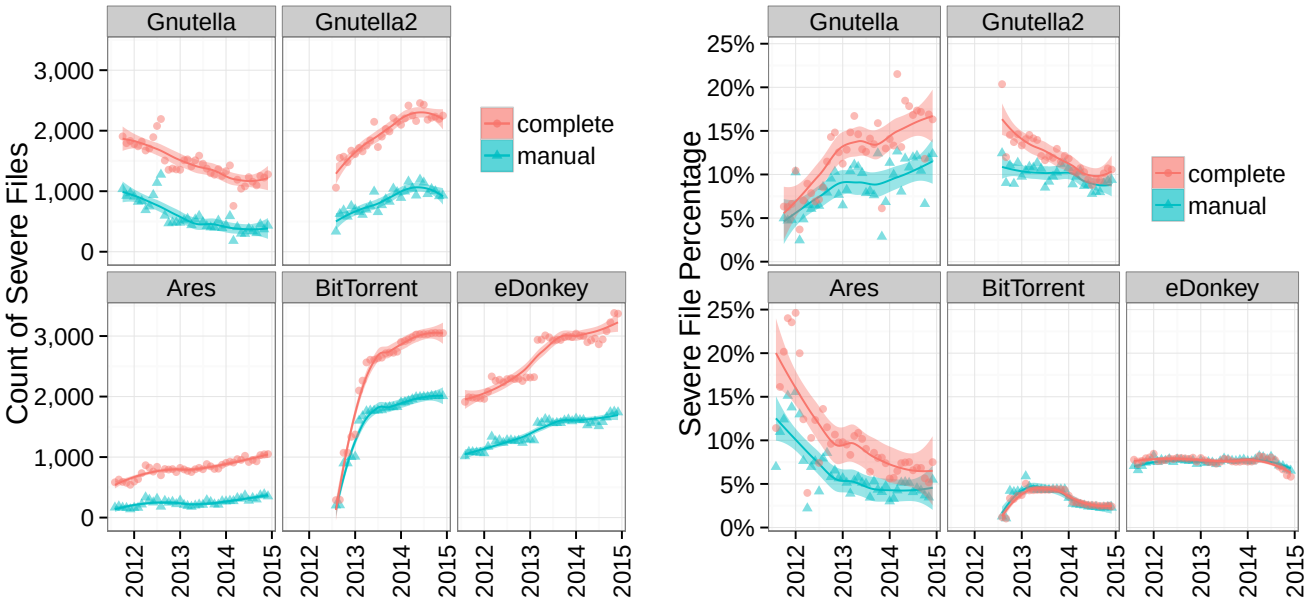

Figure 3. Severe file counts (left) and Severe file ratio (right). The latter reports the same count as the former but normalized by the total file count. Total file count for the Manual set is the count of all manually labeled files, and for the Complete set it includes both manually labeled and classifier-labeled files. Severe file counts are the count of files manually labeled as Severe for the Manual set, and the same manual count plus the sum of probability-of-severe for the Complete set.

"US" data refers to files initially observed in the United States but that are still available somewhere in the world in the last month of observations. The low survival rate for Gnutella is almost certainly due to the sharp decline in users, and the vigorous law enforcement investigations of CEM traffic on Gnutella (Liberatore et al., 2010). As libraries of files that were only possessed by a few peers leave the network, those files become unavailable. As the number of peers on a network shrinks, the magnitude of the expected loss in file availability with each departing peer grows.

The removal of CEM from the Gnutella network does not mean that these CEM files are disappearing from P2P networks, as we show in Figure 5(right). Despite 12% of the originally observed CEM files on Gnutella remaining on Gnutella (between October 2011 and December 2014), nearly 60% of those files were still available on at least one of the five P2P networks worldwide (in December 2014).

In Figure 6(left) and (right), we examine the per-month survival and introduction rates in our observations, respectively. Survival rate is defined as above, and we define *introduction rate* across a time period as the fraction of files available at the end of the time period that were not available at the start of the time period. For the "All networks" data, we considered a file to have been introduced if it was observed on at least one network.

Files meet the criteria for introduction to a network in one of several ways. A CEM file might appear for the first time on a network, for example when a user obtains it from another source then starts sharing it on the network. Alternately, a CEM file might be added to the known CEM list, and thus start being observed on the networks from that point on. Finally, a CEM file might have been reintroduced after being made unavailable for a month or more.

Figure 6(left) shows via the LOESS-based fit that the survival rate trend for each network over 2013 and 2014 is relatively stable; except for Ares, which has begun to show a decline. BitTorrent CEM stays near 100% survival, with eDonkey survival rates near 90%. Ares has decreased from about 75% to 65%. The per-month survival rate for known CEM on both Gnutella and Gnutella2 reflects the lower number of users on these networks. Just as our measurements of unique known CEM files are dominated by BitTorrent (see Figure 4), the survival rate across all networks is similarly influenced by BitTorrent's rate.

Figure 6(right) plots the introduction rates. The values are influenced by the introduction of new known CEM to our data set. For example, BitTorrent introductions happen in a few bursts, corresponding to the initial prototyping of the BitTorrent investigations and stepwise rollout of new torrents. We also note that new content may have been introduced by peers but not appear in our data set. Even so, eDonkey exhibits a 10% introduction rate, while Ares has a rate of 25%. There are several months where the introduction rate across all networks is oddly high; we suspect these correspond to the

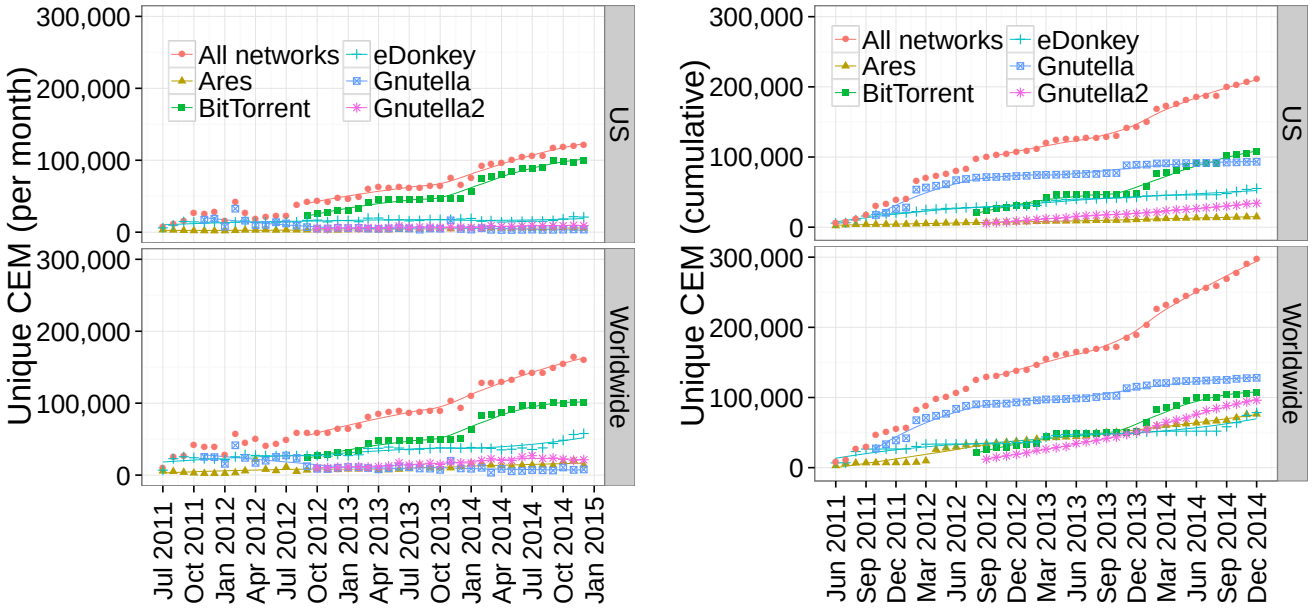

Figure 4. The number of unique known CEM files publicly advertised on P2P networks per month (left) and cumulatively (right). The "All networks" data is the number of unique files across all networks.

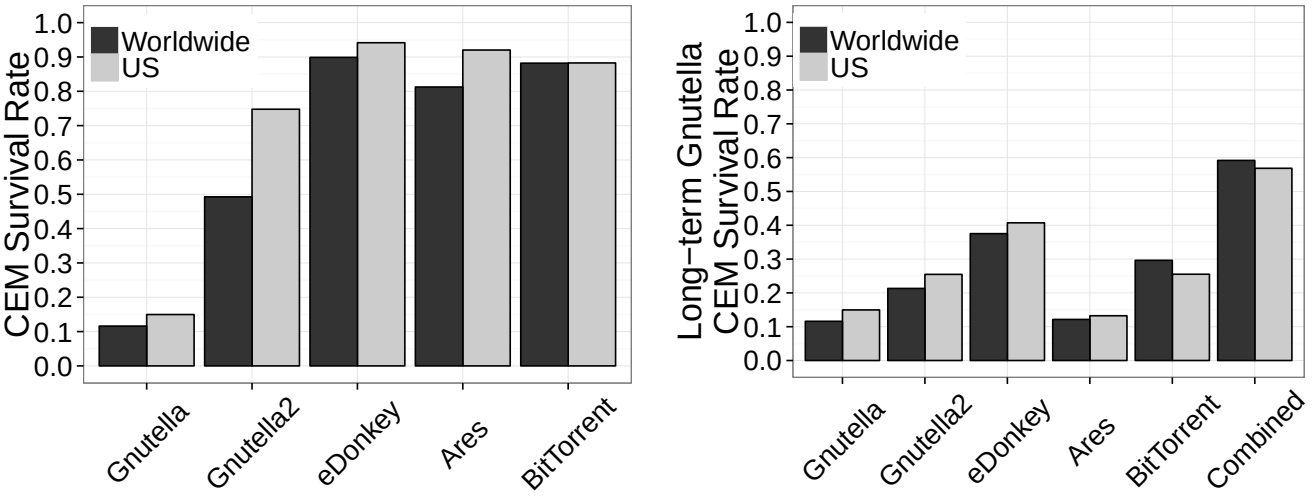

Figure 5. (left) Fraction of known CEM from the first month of observation in each network still available in the last month (Dec. 2014). (right) Fraction of CEM from the first month of observation in Gnutella that are still available on each network in the last month.

addition of new hashes corresponding to known CEM files. The smallest networks (Gnutella and Gnutella2) have what appear to be paradoxically high introduction rates; however, given the small user base, absolute changes in the number of peers (and their corresponding libraries) lead to relatively large changes in rates.

BitTorrent torrent survival. Once a torrent is defined, its description of the constituent files is immutable. For the torrent to remain active, at least a portion of its files must continue to be shared on the network. A torrent may also contain optional, mutable fields, such as the creation date. A creation date is not completely reliable due to its mutability, but it can be used to get a general understanding of when a

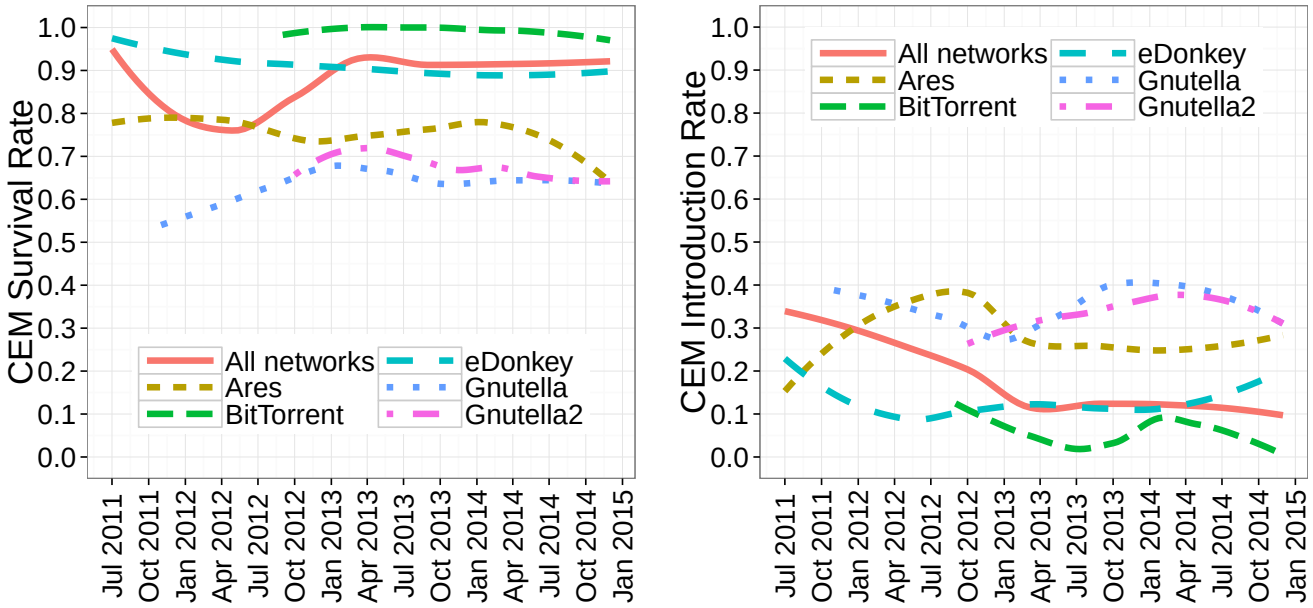

Figure 6. The per-month *survival rate*(left) and *introduction rate*(right) of CEM publicly advertised on P2P networks.

torrent, and therefore the files referenced by the torrent, may have been introduced to the BitTorrent network.

We examined the torrents used in our observations. About half of the torrents that were still active had creation date fields, with the earliest two from 2002 (BitTorrent was introduced in 2001). Over 60% of the active torrents with creation dates were from 2010 or before. This statistic implies that a large percentage of the known CEM files observed on the BitTorrent network were introduced at least two years before we began our observations.

9 Summary and Concluding Discussion

We estimate that over 840,000 peers shared CEM on the Ares, BitTorrent, eDonkey, Gnutella, and Gnutella2 P2P networks worldwide in December 2014. Although BitTorrent is increasing in popularity among CEM traffickers, the overall trend is downward. Our characterization of activity on each network is important for law enforcement who must allocate limited training and enforcement resources. While law enforcement efforts may be driving this trend in part, we expect that CEM traffickers are increasingly making use of technologies that may be more appealing; e.g., social networks, cellular messaging apps, and darknets. These venues require new investigative techniques to address CEM distribution and to discover contact offenders, as well as to characterize the scope of such criminal activity.

Our survey of United States law enforcement found that, within BitTorrent, where law enforcement applied their own measure of content severity, the rate of contact offenses among peers sharing the most severe CEM (28.8%) was higher than those sharing the least severe CEM (15.4%) (p-value=0.02). These results point to one or both of the following possibilities: possession of Severe files are causally linked to contact offenses; or discovery of Severe files by law enforcement more often invokes a process where contact offenses are discovered during arrest (such as the use of polygraphs). If the former, then a deeper understanding of this link is necessary, and greater emphasis on investigative tools that uncover such content online is needed. If the latter, then an examination of the efficacy of various arrests processes is needed. An improved sequel to our study of possession and contact offenses would ideally hold all cases to a consistent arrest process (e.g., always using a polygraph examination and a consistent process for determining contact offenses), would make use of integrated data collection prospectively rather than retrospectively, and would ideally follow the offenders for several years afterwards. Our study shows the prevalence of CEM involving infants and toddlers, victims who lack an ability to report their situations. These results indicate the importance of deploying proactive approaches to discovering contact offenders. Additionally, in the United States, possession of CEM involving children under 12 is a sentencing enhancement, and our data is relevant to those considering revisions to this policy that distinguish younger ages.

Survival rates of CEM files on the P2P networks we studied is above 60% each month. Our results show similarly that despite Gnutella's collapse in popularity, its content survives on other networks years later. These quantified results are important for victims that seek restitution based on the lasting

damage that distribution of CEM causes (Bazelon, 2013). While Gnutella's collapse is due to a copyright-related law suit (Halliday, 2011), these observations predict that a law enforcement strategy that aims to shut down networks will not significantly reduce content availability.

In future work, we seek to address other limitations of our study. For example, our work on severity is limited to the classification systems used by law enforcement; other classification systems would allow direct comparison to past studies. Similarly, our estimates of CEM availability and user populations are based on only CEM already identified by law enforcement; many more CEM files are available and introduced to these networks each day.

10 References

Bazelon, E. (2013, January 27). The price of a stolen childhood. *New York Times Magazine*.

Bourke, M. L., Fragomeli, L., Detar, P. J., Sullivan, M. A., Meyle, E., & O'Riordan, M. (2014). The use of tactical polygraph with sex offenders. *Journal of Sexual Aggression*, *21*(3), 354–367. doi: 10.1080/13552600.2014.886729

Bourke, M. L., & Hernandez, A. E. (2009). The butner study redux: A report of the incidence of hands-on child victimization by child pornography offenders. *Journal of Family Violence*, *24*(5), 183-191. doi: 10.1007/s10896-008-9219-y

Cohen, B. (2003, February). Incentives build robustness in BitTorrent. In *Proc. Intl. Workshop on Peer-to-Peer Systems.*

Deselaers, T., Pimenidis, L., & Ney, H. (2008, December). Bag-of-visual-words models for adult image classification and filtering. In *Proc. Intl. Conf. on Pattern Recognition* (pp. 1–4). doi: 10.1109/ICPR.2008.4761366

Dingledine, R., Mathewson, N., & Syverson, P. (2004, August). Tor: The second-generation onion router. In *Proc. USENIX Security Symposium* (pp. 303–320).

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, *27*(8).

Fournier, R., Cholez, T., Latapy, M., Chrisment, I., Magnien, C., Festor, O., & Daniloff, I. (2014). Comparing pedophile activity in different P2P systems. *Social Sciences*, *3*(3), 314–325. doi: 10.3390/socsci3030314

Halliday, J. (2011, May 13). Limewire settles file-sharing legal battle for $105m. *The Guardian*. Retrieved from http://www.theguardian.com/technology/2011/may/13/limewire-settles-filesharing-legal-battle

Hurley, R., Prusty, S., Soroush, H., Walls, R., Albrecht, J., Cecchet, E., ... Wolak, J. (2013, May). Measurement and analysis of child pornography trafficking on P2P networks. In *Proc. Intl. World Wide Web Conference.*

Inches, G., & Crestani, F. (2012, September). Overview of the intl. sexual predator identification competition at pan-2012. In *CLEF 2012 Evaluation Labs and Workshop Online Working Notes.* Retrieved from http://ceur-ws.org/Vol-1178/CLEF2012wn-PAN-InchesEt2012.pdf

Internet Live Stats. (2014, Retrieved March 16, 2015). *Elaboration of data by Intl. Telecommunication Union, U.N. population division July 1, 2014 estimate.* http://www.internetlivestats.com/internet-users-by-country/.

Klingberg, T., & Manfredi, R. (2002, June). *The Gnutella RFC, version 0.6.* http://rfc-gnutella.sourceforge.net/src/rfc-0_6-draft.html.

Koontz, L. D. (2005). *File sharing programs: The use of peer-to-peer networks to access pornography* (Tech. Rep. No. GAO-05-634). U.S. Government Accountability Office.

Kulbak, Y., & Bickson, D. (2005, January). *The eMule Protocol Specification.* http://www.cs.huji.ac.il/labs/danss/p2p/resources/emule.pdf.

lap3k. (n.d.). *Ares galaxy* [Computer Software]. http://aresgalaxy.sourceforge.net/.

Latapy, M., Magnien, C., & Fournier, R. (2013). Quantifying paedophile activity in a large P2P system. *Information Processing & Management*, *49*(1), 248–263. doi: 10.1016/j.ipm.2012.02.008

Liberatore, M., Erdely, R., Kerle, T., Levine, B. N., & Shields, C. (2010, August). Forensic investigation of peer-to-peer file sharing networks. In *Proc. DFRWS Annual Digital Forensics Research Conference.* doi: 10.1016/j.diin.2010.05.012

Liberatore, M., Levine, B. N., Shields, C., & Lynn, B. (2014, September). Efficient tagging of remote peers during child pornography investigations. *IEEE Transactions on Dependable and Secure Computing*, *11*(5), 425–439. doi: 10.1109/TDSC.2013.46

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Munson, E. V., & Tsymbalenko, Y. (2001). To search for images on the web, look at the text, then look at the images. In *Proc. Intl. Workshop on Web Document Analysis.* Retrieved from http://cgi.csc.liv.ac.uk/~wda2001/Papers/25_munson_wda2001.pdf

(NCMEC) National Center for Missing & Exploited Children. (n.d.). *Child Victim Identification Program.* http://www.missingkids.com/cvip.

Panchenko, A., Beaufort, R., & Fairon, C. (2012). Detection of child sexual abuse media on P2P networks: normalization and classification of associated filenames. In *Proc. LREC Workshop on Language Resources for Public Security Applications.* Retrieved from http://cental.fltr.ucl.ac.be/team/~panchenko/11_Camera_ready_paper.pdf

Peersman, C., Schulze, C., Rashid, A., Brennan, M., & Fischer, C. (2014, May). iCOP: Automatically identifying new child abuse media in P2P networks. In *Proc. IEEE Security and Privacy Workshops* (pp. 124–131). doi: 10.1109/SPW.2014.27

Rowley, H. A., Jing, Y., & Baluja, S. (2006). Large scale image-based adult-content filtering. In *Proc. Intl. Conf. on Computer Vision Theory and Applications* (pp. 290–296). (http://www-cgi.cs.cmu.edu/afs/cs.cmu.edu/user/har/Web/visapp2006.pdf)

Schulze, C., Henter, D., Borth, D., & Dengel, A. (2014, April). Automatic detection of CSA media by multi-modal feature fusion for law enforcement support. In *Proc. Intl. Conf. on Multimedia Retrieval* (pp. 353:353–353:360). doi: 10.1145/2578726.2578772

Seto, M., Hanson, R., & Babchishin, K. (2011). Contact sexual offending by men with online sexual offenses. *Sex Abuse*, *23*(1), 124–145. doi: 10.1177/1079063210369013

Steel, C. M. (2015). Web-based child pornography: The global impact of deterrence efforts and its consumption on mobile platforms. *Child Abuse & Neglect*. doi: 10.1016/j.chiabu.2014.12.009

Stokes, M. (n.d.). *Gnutella2* [Computer Software]. http://g2.doxu.org/.

Svedin, C. G., & Back, K. (1996). *Children who don't speak out: About children being used in child pornography*. Rädda Barnen.

Ulges, A., & Stahl, A. (2011). Automatic detection of child pornography using color visual words. In *Proc. IEEE Intl. Conf. on Multimedia and Expo* (pp. 1–6). doi: 10.1109/ICME.2011.6011977

United States Sentencing Commission. (2012, December). *Federal Child Pornography Offenses, Chapter 12: Findings, Conclusions, and Recommendations to Congress* (Tech. Rep.). Retrieved from http://www.ussc.gov/news/congressional-testimony-and-reports/sex-offense-topics/report-congress-federal-child-pornography-offenses

von Weiler, J., Haardt-Becker, A., & Schulte, S. (2010). Care and treatment of child victims of child pornographic exploitation (cpe) in germany. *Journal of Sexual Aggression*, *16*(2), 211–222. doi: 10.1080/13552601003759990

Wolak, J., Liberatore, M., & Levine, B. N. (2014, February). Measuring a year of child pornography trafficking by U.S. computers on a peer-to-peer network. *Elsevier Child Abuse & Neglect*, *38*(2), 347–356. doi: 10.1016/j.chiabu.2013.10.018

11 Supplementary Materials

Below are supplementary materials that provide details of our mathematical models, expanded methodology, and supplementary data via oversized tables and additional tables.

11.1 Geographic Breakdown of CEM Sharing

Table 5 shows the per-month average distinct IP addresses sharing known CEM per country. The table is meant to be compared with Table 6 to illuminate the differences from counting by GUID rather than IP address. Table 6 shows the per-month average distinct GUIDs observed per country sharing CEM for each P2P network; unlike Table 1, the countries are the top 35 ranked per-capita, rather than the top 10 ranked by total GUIDs.

11.2 Gnutella Historical GUID Counts

In Figure 7, we combine the October 2011 through December 2014 United States Gnutella GUID counts from our study with Gnutella GUID data collected by Hurley et al. (2013), beginning in January 2010. The Gnutella network was supported strongly by the company LimeWire. In March 2011 a permanent injunction was issued against LimeWire related to copyright violations, and their shutdown was followed by a precipitous decline in the network.

11.3 Calculation of GUID Estimates

We estimated the GUID counts for BitTorrent and Ares by first observing the ratios of GUIDs-to-IP addresses that we observed for eDonkey, Gnutella, Gnutella2 on a *per-ISP basis*. We calculate the ratios per-ISP because we would expect different ISPs to employ different strategies for allocating IP addresses; e.g., a cellular ISP versus a business ISP. For each ISP, we had 1 to 36 measurements of this ratio on a per-month basis. Let $G = g_1, g_2, \ldots, g_n$ be a set of counts of GUIDs observed per month for a given ISP, and let $I = i_1, i_2, \ldots, i_n$ be a set of counts of IP addresses observed per month for the same set of n months, where $1 \leq n \leq 36$. In both cases, we take the sum of eDonkey, Gnutella, and Gnutella2 counts. We estimate *each ISP's* GUID-to-IP ratio r_m during month m as the sum of distinct GUIDs observed with the ISP over the sum of distinct IP addresses for the ISP. For a given ISP, r is our estimate of its GUID-to-IP ratio given its set of r_m values.

$$r = \sum_{m=1}^{n} G_m / \sum_{m=1}^{n} I_m \tag{1}$$

Our estimate of BitTorrent and Ares GUIDs in a given month for a given ISP is the GUID-to-IP ratio, r, times the number of IPs observed using BitTorrent and Ares, respectively.

To account for a biased estimate, we compute r using bootstrapping. Specifically, we compute r as the mean of a set $r_1, r_2, \ldots$, where each r_j ($j = 1 \ldots$) is computed using a sample of the n values from G and I chosen with replacement. For each ISP, we compute r from only months with at least 50 GUIDs. If no such months exist, we use a ratio computed from all ISPs in the same country. If a country doesn't have any months with at least 50 GUIDs, we use a worldwide ratio. We use bootstrapping to compute the country and worldwide ratio estimates.

To quantify the error of our method, we computed ratios from all months but one, and for the remaining month, we compared the estimated sum number of GUIDs for eDonkey, Gnutella, and Gnutella2 to the actual sum. We did this across all 36 months and found that the estimates had a mean absolute percentage error (MAPE) of 4.8%. The MAPE is computed as the mean of ($|estimated - actual|/actual$) over all 36 months.

Figure 8 plots the per-month difference between our estimates of GUID counts for eDonkey, Gnutella, and Gnutella2 and actual counts. In this plot, the estimate for each month is based on a ratio computed from all other months, as described in the main text.

11.4 Evaluating Logistic Regression Classification of Severe Files

To evaluate the performance of our LR classifier, we used 50-fold stratified cross-validation, which essentially entailed performing 50 randomized experiments. For each experiment, we randomly divided the set of 940,960 manually labeled CEM files into non-overlapping training and testing sets, with the training set consisting of 90% of the files. The training set was used to train the LR classifier, while the testing set was used to measure its performance. Both sets had the same distribution of category labels (i.e., Severe and Not Severe) as

Table 5
2014 per-month average of distinct IP addresses observed sharing CEM.

Country[a]	eDonkey	Gnutella	Gnutella2	Ares	BT	% Pop.[b]
USA (40th)	11,674	1,747	3,033	16,235	23,649	0.02
Spain	36,184	21	82	42,541	2,359	0.23
Brazil	36,190	593	4,332	204,978	13,626	0.24
Peru	1,958	23	78	30,010	503	0.26
Ecuador	325	5	54	16,147	214	0.28
Dom. Rep.	387	6	5	15,008	148	0.31
Italy	100,261	97	384	8,928	4,981	0.31
Chile	1,985	8	61	34,193	930	0.32
Mexico	6,346	115	455	184,073	2,662	0.38
Argentina	7,738	38	221	102,597	2,154	0.45

Note: Median population percentage for countries shown is 0.031%; worldwide median is 0.071%
[a] Top 9 countries by population with at least 5 million Internet users, plus the United States. [b] The column estimating the percentage of the country's population sums the total IPs, but does not double-count addresses appearing in more than one network.

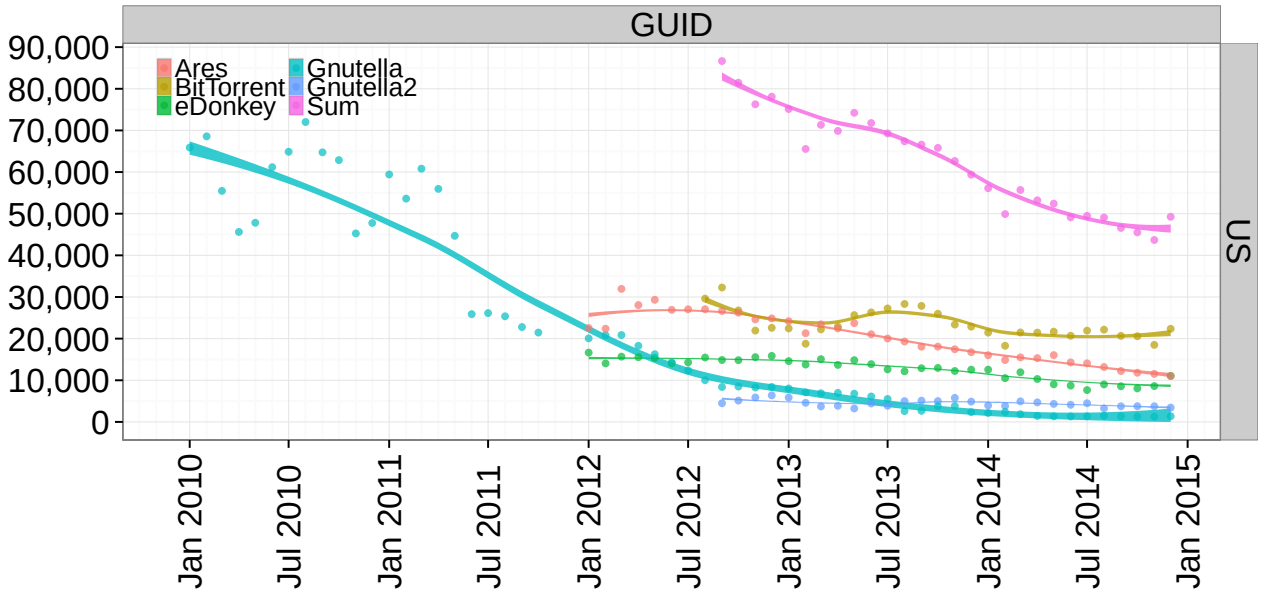

Figure 7. Five years of United States Gnutella GUID counts, including over two years prior to our study.

the full set of files. For all files in the testing set, the classifier returned the *probability-of-severe*, which is an estimate of the posterior probability that the given file actually belongs to the Severe category given the tokens it comprises. For each of the 50 experiments, we generated separate *precision–recall* (PR) and *receiver-operating-characteristic* (ROC) curves by varying the minimum threshold for probability-of-severe deemed sufficient to justify a classification of Severe. The threshold was varied between 0.0 (all files labeled Severe) and 1.0 (all files labeled Not Severe). For the PR curve, *precision* is defined as the fraction of test files labeled as Severe by the classifier that are actually Severe; in contrast, *recall* is the fraction of Severe files in the test set that are labeled as Severe by the classifier. The ROC curve plots the true positive rate versus the false positive rate, where the true positive rate is equivalent to recall, and the false positive rate is the fraction of Not Severe test files that are incorrectly labeled as Severe. All metrics, precision, recall, true positive rate, and false positive rate take their values in the range [0, 1].

Ideally a classifier would yield high precision with high

Table 6
2014 per-month average of distinct GUIDs observed sharing CEM.

Country[a]	eDonkey	Gnutella	Gnutella2	Ares[b]	BT[b]	% Pop.[c]
Algeria	305	21	81	326	166	0.01
Romania	609	13	51	311	689	0.01
USA	9,667	1,555	4,031	13,824	20,934	0.02
Turkey	754	112	93	6,038	771	0.02
Australia	1,252	125	536	865	2,079	0.02
Czech Republic	699	17	110	305	885	0.02
Austria	1,048	20	235	237	295	0.03
Germany	11,864	78	2,012	1,940	2,760	0.03
Hungary	711	47	132	91	976	0.03
UK	3,718	294	1,268	3,833	6,135	0.03
Hong Kong	587	4	531	49	391	0.03
Ukraine	1,938	8	197	40	2,461	0.03
Portugal	1,166	21	84	542	255	0.03
Saudi Arabia	397	87	155	3,819	677	0.03
Finland	644	10	66	188	726	0.03
Canada	2,856	302	891	2,195	4,529	0.03
Poland	3,214	39	212	4,362	1,048	0.03
Switzerland	1,425	30	173	345	542	0.04
China	92,954	8	352	200	163,668	0.04
Denmark	910	64	103	240	856	0.04
France	16,880	164	1,710	3,006	2,662	0.04
Belgium	2,024	70	573	811	734	0.04
Sweden	1,578	97	255	457	2,086	0.05
Dom. Rep.	72	2	6	2,603	25	0.05
Netherlands	3,306	194	990	1,548	3,791	0.06
Peru	505	16	25	7,203	118	0.06
Brazil	10,695	234	2,398	61,164	4,076	0.07
Colombia	551	11	28	19,568	240	0.08
Chile	707	8	22	10,163	320	0.10
Ecuador	139	9	14	5,912	92	0.10
Italy	33,352	43	174	2,939	1,667	0.10
Spain	18,201	14	58	19,187	1,139	0.11
Mexico	2,038	46	242	57,456	884	0.12
Argentina	2,573	13	82	29,693	693	0.13
Israel/Pal.	8,043	8	31	308	986	0.16

Note: Median population percentage for countries shown is 0.015%; worldwide median is 0.031%

[a] Top 35 countries by population with at least 5 million Internet users.
[b] BitTorrent and Ares counts are estimates as described in Appendix. [c] This column uses a sum total of GUIDs that double-counts entries appearing in multiple networks.

recall and a high true positive rate with a low false positive rate. In practice, however, there is often a tradeoff between precision and recall, and between true positives and false positives. Figure 9(left) shows precision–recall tradeoffs for all 50 randomized experiments. The figure indicates that it is possible to achieve at least 0.55 precision with 0.20 recall across all 50 experiments. Although a precision of 0.55 is relatively low in absolute terms, this value still improves greatly on baseline classification, which labels all files as Severe. In that case precision would be 0.02 and recall 1.0.

Figure 9(right) shows a similar tradeoff between true and false positive rates for each of the experiments. The ROC curve shows that true positive rates remain greater than 0.80 for all false positive rates greater than 0.20, which indicates

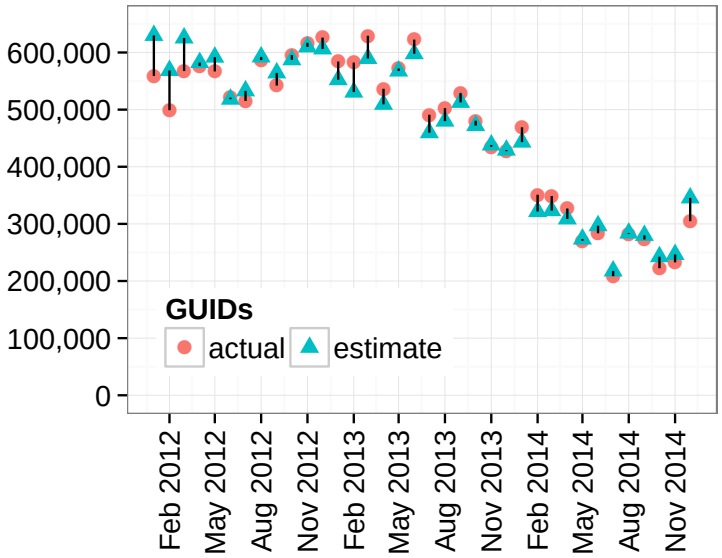

Figure 8. The per-month difference between our estimates of GUID counts for eDonkey, Gnutella, and Gnutella2 and actual counts.

that the classifier tends to be effective at differentiating between Severe and Not Severe files. Had the classifier been completely ineffective, we would have seen true positive rates equal to false positive rates for all probability-of-severe thresholds. The median *area-under-the-curve* (AUC) is 0.91. AUC can be interpreted as the probability that the LR classifier will return a higher probability-of-severe for a randomly chosen Severe file than a randomly chosen Not Severe file (Fawcett, 2006).

As noted in the main test, using the classifier to estimate the count of severe files is a significantly easier task that determining whether a specific file is severe. Across 50-folds, the MAPE of the labeled set was under 1%.

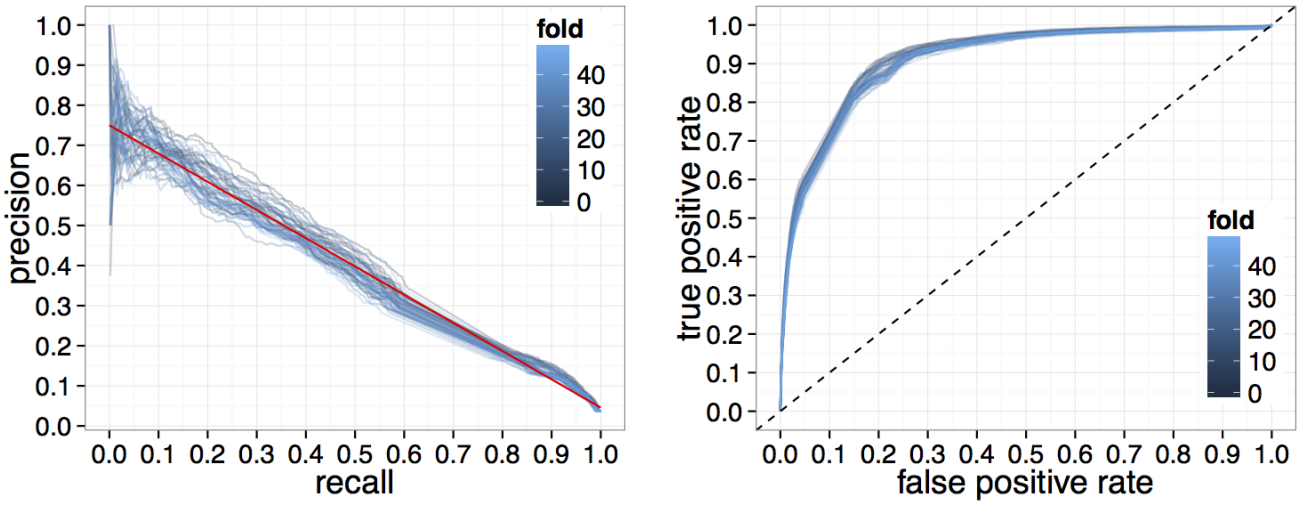

Figure 9. Precision-recall (left) and receiver-operating-characteristic (right) performance for 50-fold stratified cross validation of file category classification with logistic regression. The PR curves demonstrate the tradeoff between classifying all Severe files as being Severe (high recall) and ensuring the classification correctness of all files classified as Severe (high precision). The straight line in the PR plot is a linear best-fit for all points. The ROC curves validate the performance of the LR classifier with respect to the probability of correctly predicting a file as Severe; curves furthest away from the dashed line (where true positive rate equals false positive rate) indicate better performance.