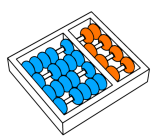
Measurement and Analysis of Child Pornography Trafficking on P2P Networks

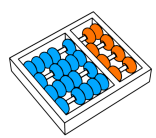
Ryan Hurley, Swagatika Prusty, Hamed Soroush, Robert J. Walls Jeannie Albrecht, Emmanuel Cecchet, Brian Neil Levine Marc Liberatore, Brian Lynn, Janis Wolak

Instituto de Computação - Unicamp

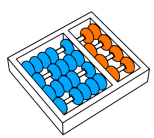
15 de Outubro de 2013



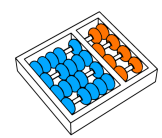
Scheduling



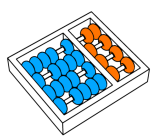
Schedule




UNICAMP

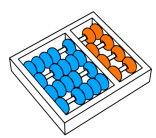


Schedule



- ✓ Works properly and is evaluated under the goal of the investigations.
- ✓ We follow that principles, (basic principles) rather than isolated characterization of the users.
- ✓ We will review the USA[1] Law under the constraints of criminal investigations for Children Pornograph.

[1] Fourth Amendment and related jurisprudence

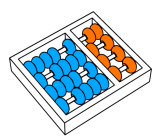


Works properly and is evaluated under the goal of the investigations

- ✓ That means: The criminal investigation is increasingly advanced and with more development tools for this one.
- ✓ There are always more groups that works to find and to discover and try to prevent Child Pornography (CP).
- ✓ But it is very difficult do prevent, because of the large scale growth in the worldwide web. There are over 1,8 milion CP in internet ”found on eMule” (we estimate much more).

Basic principles rather than isolated characterization of the users

- ✓ This means that we will not discuss about a particular user, but the market situation that involves this type of crime.
- ✓ In criminal investigations of the type we consider search warrants must specify this location, and not a person (not a user).
- ✓ Actions by the investigators are shortened by law ”Fourth Amendment and Related Jurisprudence”, where this means that the user has a protection on a electronic data.



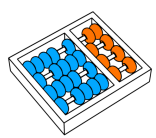
What is wrong with Fourth Amendment Jurisprudence?

The Third Party Doctrine

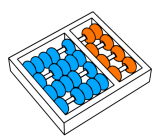
According to the Supreme Court's third party doctrine, personal information, once exposed to any third party, loses all Fourth Amendment protection. Some information exposed to third parties is protected by various statutes, but those can be inconsistent and outdated. The Electronic Communications Privacy Act (ECPA), for example, is notably out of date, leaving privacy protection of technology, as the Ninth Circuit put it, "a confusing and uncertain area of the law.". Some privacy interests that are currently unprotected under the Fourth Amendment. Konop ... also receive protection under the First Amendment – but that protection is far from comprehensive... (1967)

- ✓ The goal of the pre-warrant phase is not to make an arrest (a user, for example), but it is to obtain a judicially issued search warrant, for such cause (CP).
- ✓ This means, that we will look for a specify location, and not a person.
- ✓ Arrests in these criminal cases are typically not based on the network-acquired evidence. They are based on the fruits of the search and the person identified as possessing the contraband materials.

- ✓ Finally, we note that this follows a forensics model and not the traditional security attacker model.
- ✓ The techniques can be applied very successfully even though there exist many ways to defeat them.
- ✓ But many people do not attempt to hide them, only change the name of the file, as we know to hide the word ”sexually”, but a intentionally name to be ease to discover the file for another peer.



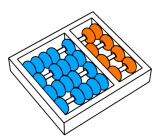
Schedule



- ✓ This study is based upon the analysis of a large number of observations of CP files on P2P networks.
- ✓ Also based upon the behavior of the peers that share these files.
- ✓ Most previous studies of P2P networks have taken place over just several days, or weeks, or a few months.

- ✓ This study is comprises of a thousand of observations per day for a full year.
- ✓ This duration is specially critical in criminal investigations.
- ✓ Scientific studies of crimes are often submitted as supporting facts during trial and sentencing.

- ✓ This study focus is on *files of interest* (FOI).
- ✓ These files includes child pornography (CP) images, as well as stories, child erotica and collections associated with this kind of crimes.
- ✓ Only content with hashing values matching a list put together by law enforcement by visual inspection was logged.



Background

- ✓ This paper is based on data collected with the help of national and international law enforcement.
- ✓ Starting in January 2009, they began deploying a set of forensics tools for online investigations.

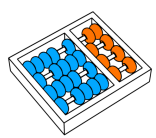
Background

- ✓ Prior to these collaborative efforts, the standard method for online investigation of CP was to make isolated cases.
- ✓ Leads were not shared among agencies or offices, other than by phone or e-mail.
- ✓ Officers leverage their own experience to prioritize suspects.

Tools

- ✓ A suite of tools, called *RoundUp*[2] (deployed by the researchers) has enabled sharing of plain view observations of online CP and associated activities on various networks.
- ✓ The shared data provide each investigator with a view of CP offenders and a method of triage for selecting targets (and enable this study).
- ✓ The tools are still in use, and law enforcement execute approximately 150 search warrants nationwide per month.

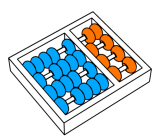
[2] Strengthening forensic investigations of child pornography on P2P networks[2]



Datasets

Table 1: Datasets

Network	Data Range	Files	GUIDs	Records
Gnutella (FOI only)	10/1/2010–9/18/2011	139,604	775,941	870,134,671
Gnutella Browse	6/1/2009–9/18/2011	87,506,518	570,206	434,849,112
eMule (FOI only)	10/1/2010–9/18/2011	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



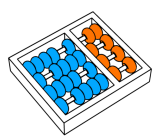
Other details

- ✓ Gnutella allows a peer to be browsed, so investigators can enumerate all files shared.
- ✓ Gnutella Browse dataset consists peer browses, not just FOI, but some Gnutella peers cannot be browsed (client configuration).
- ✓ eMule does not permit browses, so each of these datasets includes only peers that share FOIs;

Other details

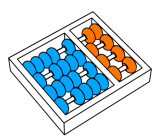
- ✓ A GUID's library is the set of files that were observed being shared by that GUID on a given day.
- ✓ A GUID's corpus is the set of all files shared by that GUID over the entire duration of the study.

Schedule



- ✓ Law enforcement's limited resources and time.
- ✓ Need triage, focusing on greater impact.
- ✓ Goal: Decrease the availability of FOI.

Schedule



File Redundancy Across GUIDs

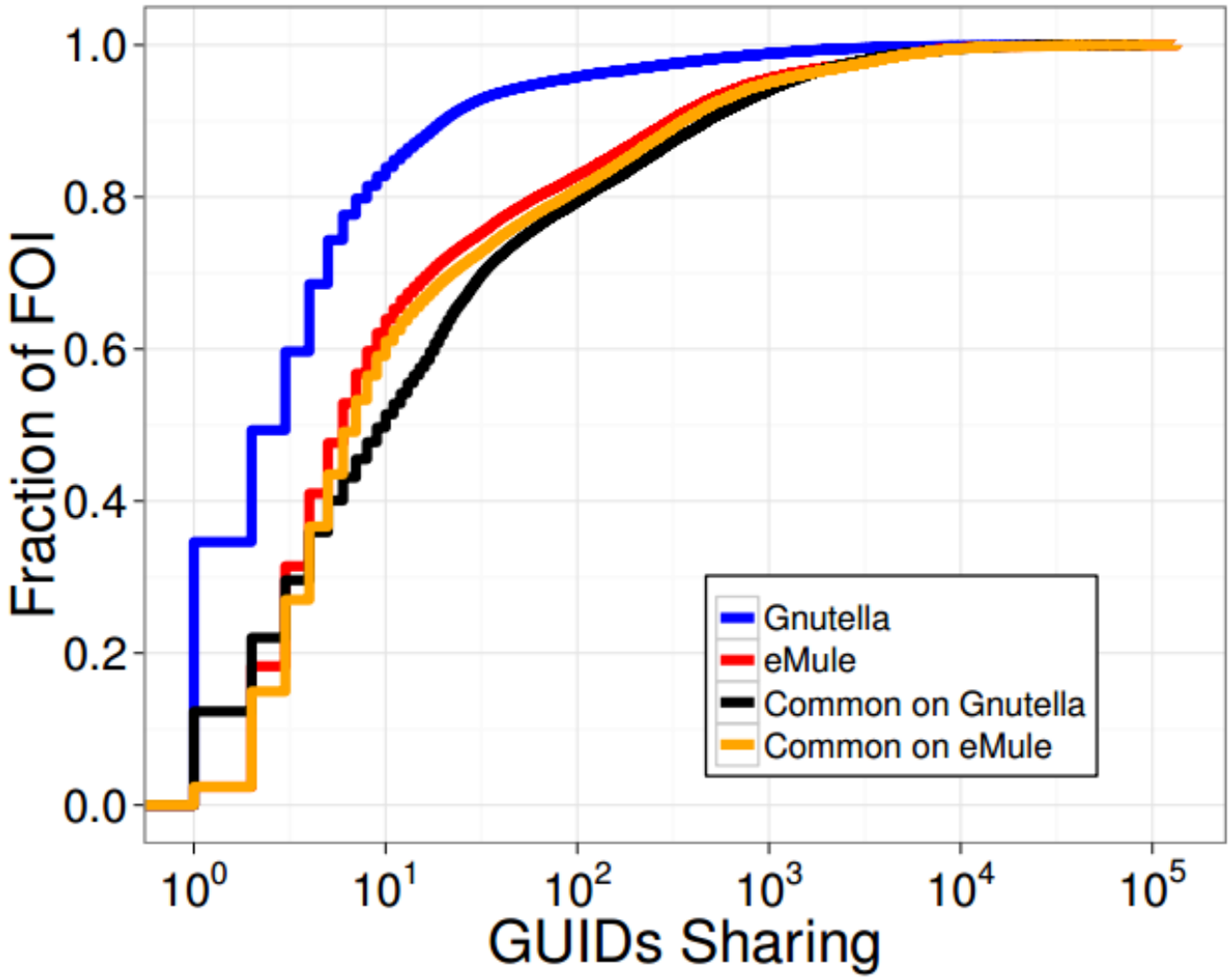
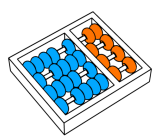


Figure 1: Count of how many IDs are sharing the same file

File Redudancy Across GUIDs

- ✓ Low redudancy of FOIs across the GUIDs:
  - 90% of Gnutella files, shared by at most 20 Ids.
- ✓ Files in common between networks have more redudancy.
- ✓ Apparently, a good strategy would be to prioritize the users with less redundant FOI.



File Availability Across Days

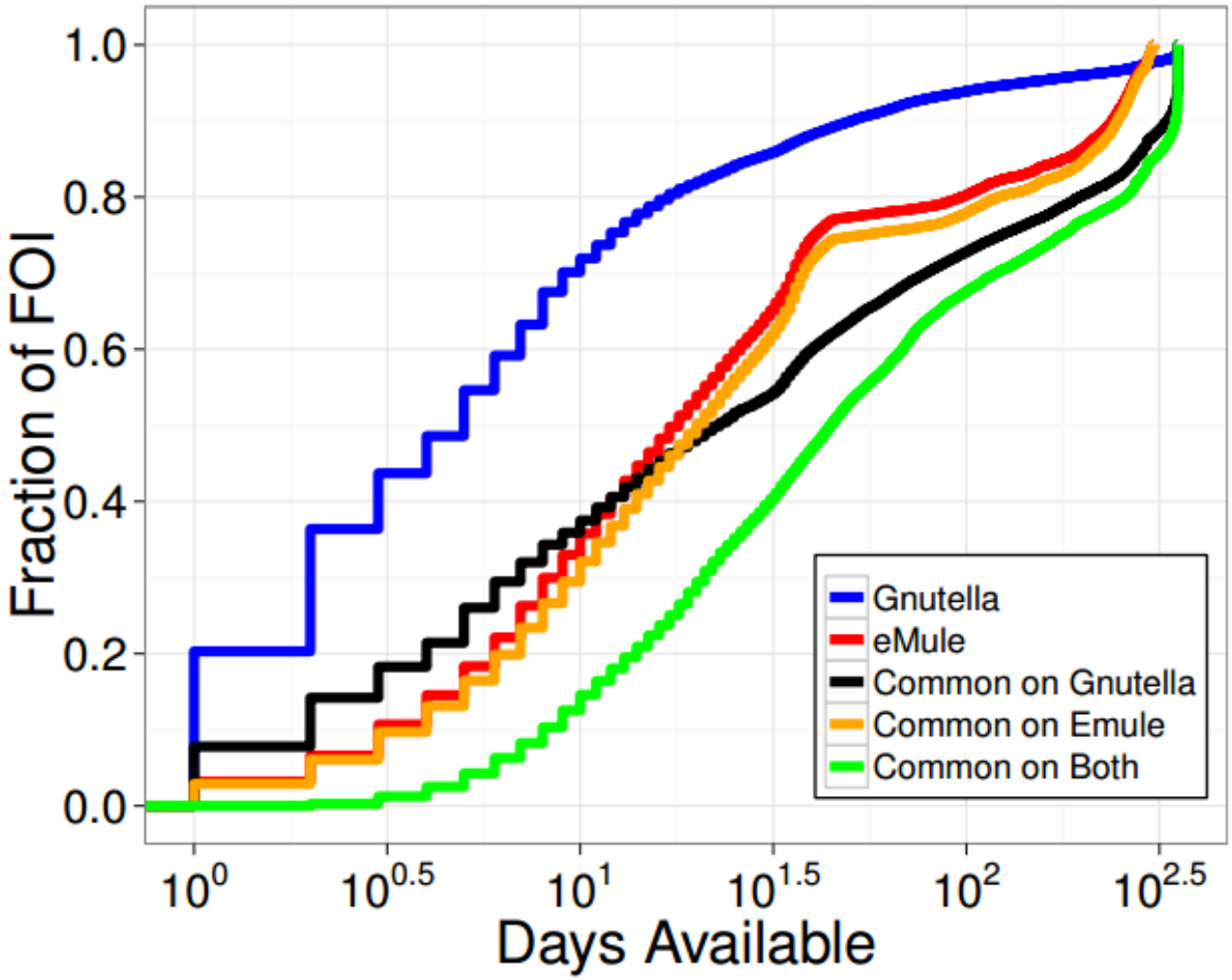
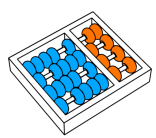


Figure 2: Count of days a file has been found online

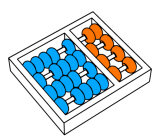
File Availability Across Days

- ✓ Gnutella have a lower availability than eMule:
  - Only 30% of Gnutella files are available more than 10 days .
- ✓ Files available fewer days, tend to be also less redundant.
- ✓ Common files also have a greater availability.

Schedule



- ✓ We know that the strategies for removing content from the entire ecosystem (the internet) must target offenders from all countries.
- ✓ We do not have of a unified effort, and no such collaboration exists.
- ✓ Investigators need a triage strategy.
- ✓ The better were if the investigators have target to catch the more dangerous criminals, but such information is not available.



- ✓ In lieu of that ideal, investigators can take peers that are offensive in the net.
- ✓ Peers that show evidence the target of the intent the user.
- ✓ This includes peers that are online for the longest duration.
- ✓ Peers that share the largest number of ”FOI” (File of Interest).
- ✓ Offenders by P2P network, as we know: eMule, Gnutella...or offender that seek to escape detection with the use of TOR.

There are 6 (six) sub-groups of peers offenders:

1. The top 10% of GUID's of largest corpora.
2. The top 10% of GUID's of sharing FOI the most numbers of days.
3. The top 10% of GUID's ranked contribuition metric (the same we saw in last topics).
4. The top 10% of the set of GUID's linked by ip adress sharing FOI.
5. The top 10% of GUID's that use a know TOR exit node.
6. The top 10% of GUID's sharing FOI that use a IP adrres and we infer that is a non TOR relay.

We can see this result under the tables:

Table 2: Sizes of each GUID subgroup

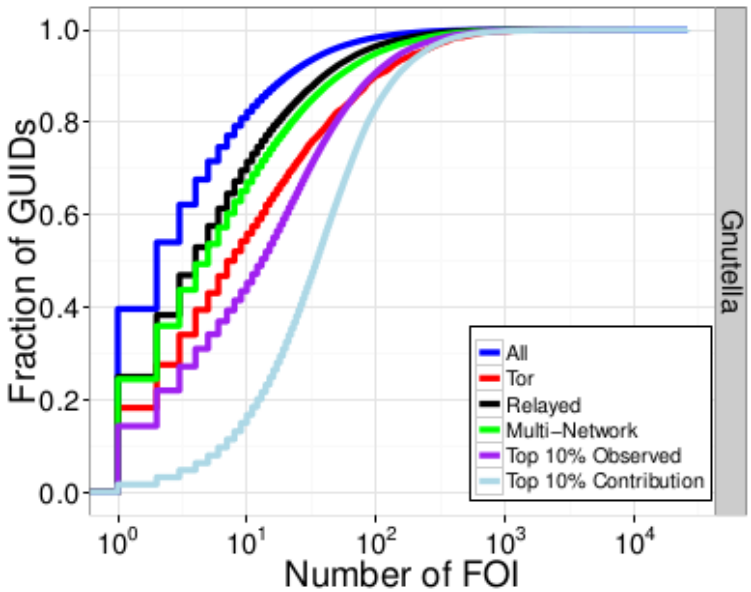
	Network	
Identifier	Gnutella	eMule
All GUIDs	775,941	1,895,804
Multi-Networks GUIDs	84,925 (11%)	147,904 (7,8%)
TOR GUIDs	3,666 (0.47%)	16,290 (0.86%)
TOR GUIDs (>2 days)	2,592 (0.33%)	11,998 (0.63%)
Relayed GUIDs	76,478 (9.9%)	78,223 (4.1%)
Top 10% Observed	84,235 (11%)	190,797 (10%)
Top 10% by Corpus	77,782 (10%)	189,951 (10%)
Top 10% by Contribution	77,595 (10%)	189,581 (10%)

Table 3: Numbers of IP addresses per network sharing FOI

	IP Addresses		
Network	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

- ✓ 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 it irregularly, and therefore their true IP addresses are easily identifiable.

A comparison of Peer Behavior



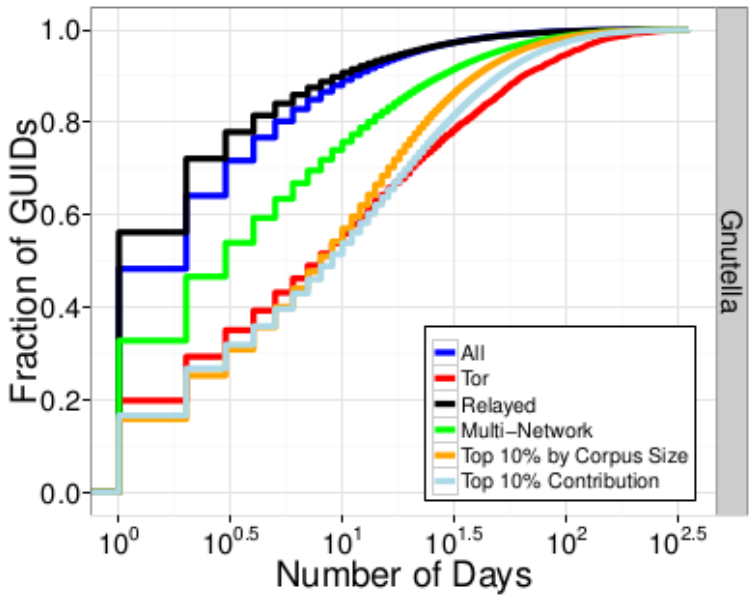
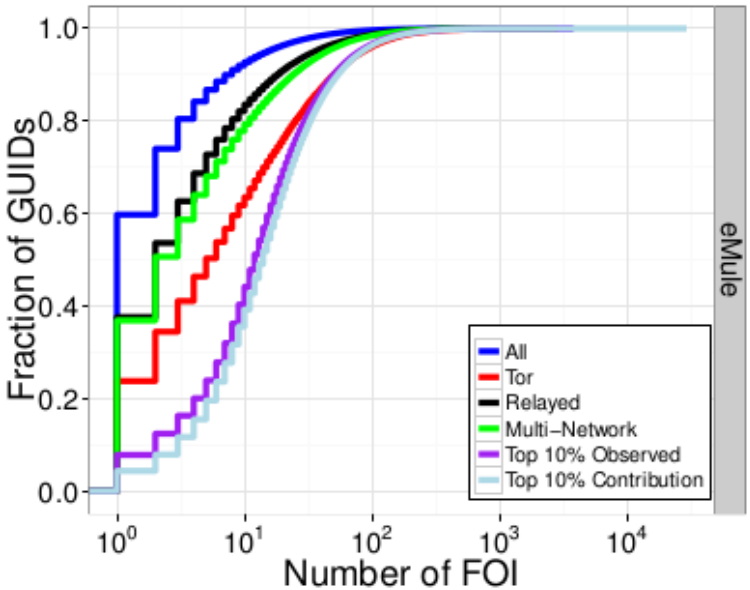


Figure 3: Cumulative distribution function (Gnutella)

.

A comparison of Peer Behavior



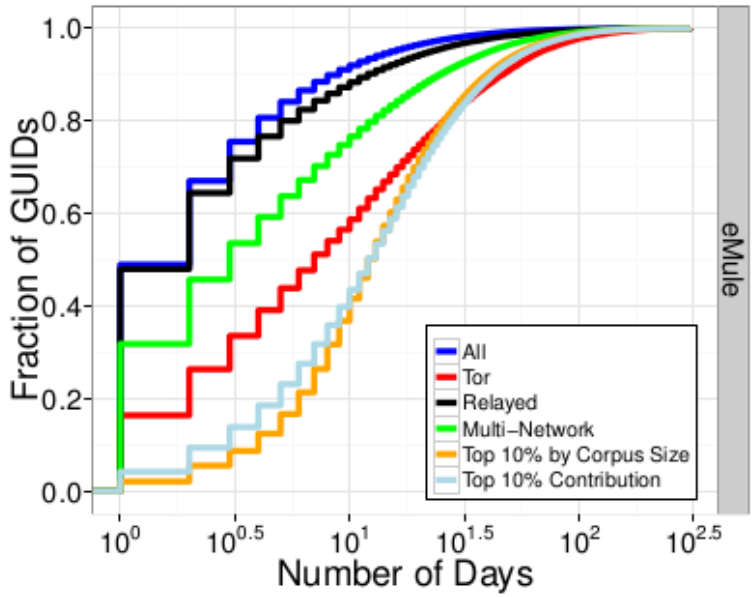


Figure 4: Cumulative distribution function (eMule)

.

Table 4: Characterization of GUIDs groups

	GUID Groups	Mean Value (99% CI)	
		Corpus Size	Days Observed
Gnutella	All	10.9 (10.7, 11.1)	5.2 (5.2, 5.2)
	TOR	43.9 (39.0, 49.6)	23.4 (21.8, 25.1)
	Relayed	18.9 (18.3, 19.5)	4.8 (4.7, 4.9)
	Multi-Network	25.9 (24.9, 27.0)	10.8 (10.6, 11.0)
	Top 10% Obs.	41.8 (40.7, 43.0)	28.7 (28.5, 29.0)
	Top 10% Corp.	75.9 (74.3, 77.7)	16.2 (16.0, 16.5)
	Top 10% Contr.	69.1 (67.6, 70.9)	19.5 (19.3, 19.8)

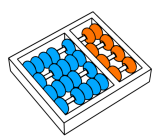
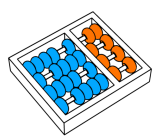


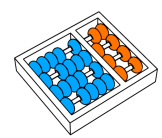
Table 5: Characterization of GUIDs groups

	GUID Groups	Mean Value (99% CI)	
		Corpus Size	Days Observed
eMule	All	4.3 (4.3, 4.4)	4.1 (4.1, 4.1)
	TOR	21.2 (19.9, 22.5)	17.4 (16.9, 18.0)
	Relayed	9.2 (8.9, 9.6)	5.5 (5.4, 5.6)
	Multi-Network	10.8 (10.6, 11.0)	9.5 (9.4, 9.7)
	Top 10% Obs.	23.5 (23.2, 23.8)	22.3 (22.2, 22.4)
	Top 10% Corp.	27.8 (27.4, 28.5)	18.7 (18.6, 18.8)
	Top 10% Contr.	25.8 (25.4, 26.5)	19.0 (18.9, 19.1)

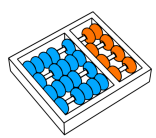
Schedule



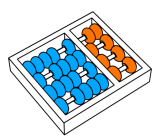

UNICAMP



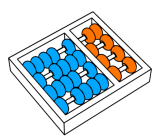
Schedule



- ✓ Files in U.S mainly. Others countries are underestimated.
- ✓ No total coverage of files of a certain GUID.
- ✓ Peers that are rarely online or have few files may have been missed.
- ✓ Greater number of GUIDs because of different installations.
- ✓ Peers may have been removed because of police action.

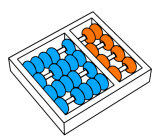


Schedule

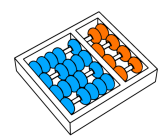


- ✓ Ecosystems and Underground Economies:
  - Economic characteristics of Network-based ecosystems.
  - May explain the irregular use of Tor encountered.
- ✓ Content Availability in P2P Systems:
  - Research of general use of P2P.
- ✓ CP Trafficking in P2P Systems:
  - Previous work are shallow on understanding how these files are being shared.

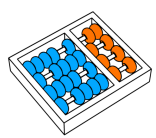
Schedule




UNICAMP



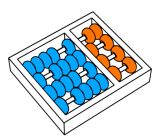
Schedule



Bibliography

Ryan Hurley, Swagatika Prusty, Hamed Soroush, Robert J. Walls, Jeannie Albrecht, Emmanuel Cecchet, Brian Neil Levine, Marc Liberatore, Brian Lynn, and Janis Wolak.
Measurement and analysis of child pornography trafficking on p2p networks.
In *Proceedings of the 22nd international conference on World Wide Web*, WWW '13, pages 631–642, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

Marc Liberatore, Brian Neil Levine, and Clay Shields.
Strengthening forensic investigations of child pornography on p2p networks.
In *Proceedings of the 6th International Conference*, Co-NEXT '10, pages 19:1–19:12, New York, NY, USA, 2010. ACM.



Thanks

Thanks!

Measurement and Analysis of Child Pornography Trafficking on P2P Networks

Ryan Hurley, Swagatika Prusty, Hamed Soroush, Robert J. Walls Jeannie Albrecht, Emmanuel Cecchet, Brian Neil Levine Marc Liberatore, Brian Lynn, Janis Wolak