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

University of Massachusetts Amherst
Williams College
University of New Hampshire
P2P networks are the most popular mechanism for trading child pornography (CP).
Active peers on a single day in April.
There is a dearth of information about CP activity on p2p networks.

So

We partnered with law enforcement to quantify and analyze the extent of CP trafficking.
Using our tools, our law enforcement partners observed millions of peers, from over 100 countries, sharing hundreds of thousands of CP files.
Project Stats Since 2009:

Our tools are in daily use by law enforcement in all 50 states.

Over 4,000 investigators trained.

Over 6,700 search warrants issued.

Over 4,500 arrests for CP possession.
104 children rescued from sexually abusive situations during the last twelve months.
Roadmap

• Motivation

• Dataset

• How do we remove content?

• How do we prioritize peers?

• Summary and conclusion
Dataset Overview

<table>
<thead>
<tr>
<th>Network</th>
<th>Date Range</th>
<th>Files</th>
<th>Peers</th>
<th>Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>eMule</td>
<td>Oct. 2010 – Sept. 2011</td>
<td>29,458</td>
<td>1,895,804</td>
<td>133,925,130</td>
</tr>
</tbody>
</table>

- Collected by law enforcement, using tools we provided
- Spans a one year period (Oct. 2010 – Sept. 2011)
- Gnutella and eMule
- CP files verified by hash value
- Peers identified by their GUID
Police gathered the data (using our tools).
We do not have access to any contraband.
Law enforcement verified that each peer shared CP through visual inspection and cryptographic hash value.
We focus primarily on U.S. law enforcement.
How do we remove content?
Lots of files...

139,604 distinct CP files on Gnutella

29,458 distinct CP files on eMule

9,000 available daily
Most of these images are of young children.
Filenames are not obfuscated. Often they are graphic and contain ages, genders, and descriptions of acts.
How resilient is the network?

<table>
<thead>
<tr>
<th>Network</th>
<th>Date Range</th>
<th>Files</th>
<th>Peers</th>
<th>Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>eMule</td>
<td>Oct. 2010 – Sept. 2011</td>
<td>29,458</td>
<td>1,895,804</td>
<td>133,925,130</td>
</tr>
</tbody>
</table>

1. **Redundancy**: the number of peers sharing a file.

2. **Availability**: the amount of time that any peer is sharing the file.
Redundancy CDF

Availability CDF

Blue line is Gnutella
Red line is eMule
Gnutella: 10% of CP Files are shared by more than 20 peers.

Blue line is Gnutella
Red line is eMule
Redundancy CDF

Blue line is Gnutella
Red line is eMule

Gnutella: 10% of CP Files are shared by more than 20 peers.

eMule: 25% of CP Files are shared by more than 20 peers.

Availability CDF
**Gnutella:** 10% of CP Files are shared by more than 20 peers.

**Gnutella:** 30% of CP Files are available for more than 10 days. 5% for more than 100 days.

**eMule:** 25% of CP Files are shared by more than 20 peers.
Gnutella: 30% of CP Files are available for more than 10 days. 5% for more than 100 days.

Gnutella: 10% of CP Files are shared by more than 20 peers.

eMule: 25% of CP Files are shared by more than 20 peers.

eMule: 60% of CP Files are available for more than 10 days. 20% for more than 100 days.
How can **police** remove content from the network?
Which peers should be removed such that we remove the most files?

NP-Hard problem so use heuristics:

- **Observed**: number of days a peer is online.
- **Corpus**: the number of files a peer is sharing.
- **Contribution**: the amount a peer contributes to the network.
Removing Content by Removing Peers

The figure shows the fraction of files remaining after removing a certain percentage of top peers. The x-axis represents the percentage of top peers removed, ranging from 0.01 to 100. The y-axis shows the fraction of files remaining. Two networks are compared: Gnutella and eMule. The heuristics used include random, observed, corpus, and contribution methods. The graph indicates that as the percentage of top peers removed increases, the fraction of files remaining decreases, with some heuristics performing better than others in terms of content removal efficiency.
Removing Content by Removing Peers

**Graph:**
- **Y-axis:** Fraction of Files Remaining
- **X-axis:** Percentage of Top Peers Removed

**Legend:**
- **Blue:** Random
- **Red:** Observed
- **Black:** Corpus
- **Orange:** Contribution
- **Pink:** All U.S.

**Networks:**
- **Gnutella**
- **eMule**
The Observed heuristic is hardly better than Random.
The Observed heuristic is hardly better than Random.

Contribution scales well, but requires more data.
Most files are shared by a relatively small set of peers.

The Observed heuristic is hardly better than Random.

Contribution scales well, but requires more data.
Removing Content by Removing Peers

![Graph showing the fraction of files remaining for different networks and heuristics as a function of the percentage of top peers removed.](image-url)

- **Heuristic:**
  - Random
  - Observed
  - Corpus
  - Contribution
  - All U.S.

- **Networks:**
  - Gnutella
  - eMule
CP trafficking is a global issue.
Location of Top Peers

[Map of the world showing the location of top peers, with regions shaded in different intensities of blue to indicate the distribution of top peers globally.]
U.S. law enforcement arrest about 3,100 offenders per year – less than 1% of the Gnutella population.

About 3,400 U.S. peers seen per day sharing known CP.
How do we prioritize peers?
Peer Overview

<table>
<thead>
<tr>
<th>Network</th>
<th>Date Range</th>
<th>Files</th>
<th>Peers</th>
<th>Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>eMule</td>
<td>Oct. 2010 – Sept. 2011</td>
<td>29,458</td>
<td>1,895,804</td>
<td>133,925,130</td>
</tr>
</tbody>
</table>

- **Gnutella**: 775,941 peers
  - 29% shared just one known file
  - 11% seen on other networks (est. by IP address)
  - Approx. 3,600 connected via Tor
- **eMule**: 1,895,804 peers
Peer Triage

• Peers that show more intent:
  • Peers that try to hide their actions (Tor)
  • Peers on multiple networks
  • Top 10% of peers by file contribution
• Others discussed in the paper
What is Tor?
Tor is free software and an open network that helps you defend against a form of network surveillance that threatens personal freedom and privacy, confidential business activities and relationships, and state security known as traffic analysis.

Learn more about Tor »
Tor hides a peer’s IP address, but it does not obfuscate the application level ID.
<table>
<thead>
<tr>
<th>Peer Group</th>
<th>Peers</th>
<th>Days (avg.)</th>
<th>Files (avg.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Gnutella</td>
<td>775,941</td>
<td>5.2</td>
<td>10.9</td>
</tr>
<tr>
<td>Tor</td>
<td>3,666</td>
<td>23.4</td>
<td>43.9</td>
</tr>
<tr>
<td>Multi-Network</td>
<td>84,925</td>
<td>10.8</td>
<td>25.9</td>
</tr>
<tr>
<td>Top 10% Contr.</td>
<td>77,595</td>
<td>19.5</td>
<td>69.1</td>
</tr>
<tr>
<td>Peer Group</td>
<td>Peers</td>
<td>Days (avg.)</td>
<td>Files (avg.)</td>
</tr>
<tr>
<td>--------------------</td>
<td>----------</td>
<td>-------------</td>
<td>--------------</td>
</tr>
<tr>
<td>All Gnutella</td>
<td>775,941</td>
<td>5.2</td>
<td>10.9</td>
</tr>
<tr>
<td>Tor</td>
<td>3,666</td>
<td>23.4</td>
<td>43.9</td>
</tr>
<tr>
<td>Multi-Network</td>
<td>84,925</td>
<td>10.8</td>
<td>25.9</td>
</tr>
<tr>
<td>Top 10% Contr.</td>
<td>77,595</td>
<td>19.5</td>
<td>69.1</td>
</tr>
</tbody>
</table>

**Key Point:** Each subgroup is more active in their trafficking, having more CP and uptime than the average peer.
Tor is not a significant barrier to investigation.

90% of frequent Tor peers can be linked back to a non-Tor IP address.
Criminal Investigations

- Bound by U.S. Law
- Plain-view only
- Warrants and Arrests
Summary. We split our analysis into two parts: content and peers.

Content. It is difficult to remove content -- too many peers, too much geographic diversity, and too few resources.

Peers. Law enforcement can focus their limited resources on smaller groups of aggressive peers, e.g. Tor peers.
Future work: How can we link network-observable characteristics with real-world behavior?
Thank you!

We would like to thank our law enforcement partners.

Acknowledgements. This work is based in part upon the following awards. UMass personnel were supported in part by CNS-1018615 (re: CP p2p trafficking), CNS-0905349 (re: geographic/mobile analysis in techreport), or 2008-CE-CX- K005 (re: tool development). Wolak was supported in part by CNS-1016788 (re: CP p2p trafficking). Albrecht was supported in part by CAREER award CNS-0845349.
Summary

• Quantified the alarming extent of CP trafficking on p2p networks and highlighted several difficult problems for law enforcement.

• Investigated strategies for removing CP content from the network and found that international cooperation is needed.

• Characterized several subgroups of aggressive peers to help investigators better allocate their limited resources.

http://forensics.umass.edu/rjwalls/