About Me
Past
Ph.D., UMass Amherst, 2008-ish

- cheat-proof network gaming
- opportunistic networking with Bluetooth
- weaknesses in Tor (an anonymous communication system)
Past

Visiting Assistant Professor, Wesleyan, 2008–2009

- courses on computer and network security
- research on denial-of-service attacks in Tor
Present
Research Scientist (here) 2009–present

I work on forensics and privacy: - build tools to forensically investigate p2p networks - analyze logged investigation data - create models of types of offenders - identify mobile phone location leaks across the Internet
Present
Occasional Teaching

- Computer Crime Law (*CS391LI*), Fall 2011 (with Brian Levine); maybe again soon?
- Computer Networking (*CS453*), Spring 2014
The Problem
First, a Warning

Child sexual abuse and exploitation is an atypical topic for Computer Science.

I will try to keep all discussions at a relatively high level, and you should do the same.
What is the Problem?
Child Sexual Abuse Imagery
+ The Internet
= The Problem
Child Sexual Abuse Imagery is Contraband
What is Contraband?

Any item which is illegal to possess, or to sell, trade, or otherwise distribute.

Examples in the USA:

- child sexual abuse imagery (aka “child porn”)
  
Also:

- certain classes of drugs
- certain types of weapons
The Internet: Great Technology?
Or The Greatest?
Maybe Just Neutral

which really means: for good or bad

- Other technologies that are arguably neutral:
  - metalworking
  - the telephone
  - electricity
  - the automobile

Today, we’re talking about a specific bad use of the Internet.
Crime on the Internet

The Internet serves as conduit for “traditional” crimes:

- bullying / threats / harassment
- fraud
Scalable Crime on the Internet

and greatly increases the scope/scale of certain other types of crime:

- copyright violations
- trafficking in contraband (drugs, etc.; but also CP)
CP Trafficking in Prehistory

Before 1999 or so, CP was considered a contained problem. Then came widespread Internet access and p2p filesharing.
CP on the Internet

Suddenly it became trivial to duplicate and distribute images, seemingly without being seen.
Law Enforcement on the Internet
A Simple Maxim

Where go the crimes, so too go the investigators.
Investigators are Constrained

(Jokes about the NSA aside)

US law enforcement operates under many constraints:

- Constitutional
- Statutory
- Judicial
The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.
A Reasonable Expectation of Privacy

secure in their persons, houses, papers, and effects, against unreasonable searches and seizures,

A two-part legal test, only law since 1967 (Katz v. US):

1. a subjective (personal) expectation of privacy
2. that is objectively (societally) reasonable
   For (1): is something in “plain view,” or hidden?
Warrants

No Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.

Search warrants allow investigators to enter a home and search and seize computers. Without a warrant such actions are not permitted to government agents. We’ll talk more about probable cause later.
Investigating Crimes on the Internet
The Key Steps
(for CP trafficking, anyway):

1. Observe a crime *in plain view* on the Internet
2. Determine a network-level identifier
3. Match that identifier to a physical address or person
4. Obtain a warrant
5. Search and seize evidence
6. Arrest a suspect

Our lab builds tools to help with steps 1 and 2.
(Steps 3–6 are outside our ambit.)
Observing Crime on the Internet

How can investigators:

- observe the Internet,
- see things in plain view,
- know that they’re crimes,
- and link them to an identifier?

Let’s dive into p2p networks to see one class of answers.
P2P Filesharing Networks
Many P2P Protocols (and applications)

- Gnutella 1&2 (Limewire, Frostwire, Shareaza)
- eDonkey (eMule, aMule)
- BitTorrent (BitTorrent, µTorrent, Vuze)
- Ares (Ares Galaxy, gIFT)
- etc. etc. etc.

Let’s look at **Gnutella**
(one of the first fully p2p systems for filesharing) in more detail.
Gnutella Overview

- Gnutella is a protocol; an implementation can run on almost any computer.
- There are no powerful servers, unlike Yahoo or Google.
- All Gnutella peers act as a server and as a client.
- Anyone can write a program that speaks the Gnutella protocol.
Wherefore Gnutella?

- Gnutella exists to let users share and download files to and from one another. Two basic pieces of functionality are required:
  - a network to share *queries* for files (and responses), and
  - a way to transfer files once they’re found
- Let’s talk about each.
Forming the Network 1

No central server, peers must:

- Use a stored list of previously seen peers
- Or use a stored list of peers stored in software
- Or connect to a Gnutella WebCache and retrieve a list of peers (it’s just a web server)
Forming the Network 2

TCP links are established with a few other peers. A haphazard network topology is the result.
Equality?
Not all peers are equal:

- Every node starts as a *leaf*.
- Particularly fast or stable nodes become *ultrapeers*.
  
  In theory, only fast nodes should become ultrapeers. In practice, any node can declare itself an ultrapeer.
Leaves and Ultrapeers

- Leaf nodes each connect to a few ultrapeers (by convention, three)
- Ultrapeers connect to 32 (or more) other ultrapeers
- Ultrapeers answer requests on behalf of leaves (goal: improve overall performance)
The Network

The Gnutella network is between ultrapeers and their associated leaves.
The Network’s Purpose

The network allows peers to efficiently propagate queries and replies to queries.

- *Queries* are short text strings ("daft punk mp3") that are propagated to other peers.
- Leaves relay queries only to ultrapeers.
- Ultrapeers relay queries only to other ultrapeers.
Responses

- Peers on the network respond to queries.
- *Responses* contain:
  - filename
  - hash of file
  - IP address and port of responder’s computer
- Ultrapeers track what files their connected leaves and ultrapeers claim to have and respond for them.
Filenames

Filenames tend to be fairly descriptive, e.g.:

Arcade Fire - Reflektor - Here Comes the Night Time.mp3

or

The Wolverine 2013 - full movie - action - scifi - marvel.avi
Hashes

- A hash (aka hash value, hash sum, checksum, ...) is the result of a function:
  - input: a variable length bit sequence
  - output: a fixed length bit sequence
- Different hash functions have different purposes and properties:
  - table lookups
  - error detection
  - file fingerprinting
  - digital signatures
  - message authentication
Cryptographic Hash Functions

- The ideal cryptographic hash function has several properties:
  - it should be easy to compute
  - given an output, it should be infeasible to find an input that produces it (preimage resistance)
  - given an input, it should be infeasible to find a different input that produces the same output (second preimage resistance)
  - it should be infeasible to find two different inputs that produce the same output (collision resistance)
Hashes in Gnutella

- Gnutella uses SHA-1, a cryptographic hash function, for each file
- SHA-1 is 160 bits long
  - There are $2^{160} \approx 10^{48}$ possible 160-bit hashes
  - For comparison, Earth consists of about the same number ($10^{50}$) atoms
Hashes are Fingerprints

$2^{160} \approx 10^{48}$ possible 160-bit hashes

- For all practical purposes SHA-1 uniquely identifies its input.
- A known collision attack on SHA-1 requires $\approx 2^{63}$ operations.
- It does not matter to us! (Why? Because who’s the attacker?)
Downloads in Gnutella

File transfers in Gnutella are essentially HTTP 1.1 GET requests:

a direct and voluntary TCP connection between two computers on the Internet.
Swarming

Actually, it’s a little more complicated.

- Using the hash to identify the file, many peers can be contacted at once.
- Different pieces can be requested from each using the HTTP Range: header.

Investigators prefer *single-source* downloads to show a remote peer possessed an entire file.
IP Addresses

Computers on the Internet have an identifier: an IP address.

- Investigators ask an ISP who the responsible party for a given IP was at a given time.
- This is considered administrative information, and does not require a search warrant.
- It does require a subpoena, which is like a warrant but with a lesser standard of evidence than probable cause and narrower applicability.
Back to P2P Systems

• We’ve seen (some of) how Gnutella helps users find, share, and download files.

• Other systems work similarly, but differ in the fine details:
  - query propagation (centralized servers, distributed hash tables, ...)
  - file distribution (credit systems, tit-for-tat, ...)
  - deniability and privacy (transferring files for others, transferring encrypted data, ...)

• This is a big topic.
Investigating P2P Filesharing Systems
Putting It Together

By running a Gnutella peer, you are willing to disclose what files you have, and that you’re willing to share them. Many clients notify you of this fact when installed. Also, they’re called file sharing programs for a reason.
Legal Implications

- Courts have repeatedly affirmed that sharing this information is voluntary, akin to leaving your drapes open while committing a crime in your living room, and a clear example of acting “in plain view.” (US v. Borowy).

- Investigators can watch for known files (or suspiciously named ones) and use these observations without a warrant.
Probable Cause

“Probable cause” is the standard for a warrant. In the US, that means “a fair probability” (US v. Sokolow). If you’re a leaf:

1. Is information from an ultrapeer probable cause?
2. Is information about a file obtained directly from you probable cause?
3. What about if you upload the file to the investigator upon request?
   
   Investigators prefer (3), but any can potentially meet the standard. It’s up to a judge.
Search Warrants

Once a warrant has been issued, investigators are no longer constrained by the 4th Amendment.

- private areas can be entered
- computers can be seized
- forensic analysis of seized disks, etc., can be performed, looking for additional evidence related to the crime
(Some of) What Our Tools Do

Mostly, software engineering to greatly investigator burden:

- Visually identify known files (by hash) when searching
- Perform single-source downloads
- Log all relevant information (IPs, hashes, downloads)
- Disable unwanted functionality (e.g. don’t upload!)

Our tools help investigators correctly get the evidence needed for a warrant.
Remember Our Goals?

- observe the Internet (or at least, p2p systems),
- see things in plain view (p2p!),
- recognize evidence of crimes (hashes and single source downloads), and
- link them to an identifier (IP addresses)
Summary
• Investigators follow criminals onto the Internet.
• Law governing investigators still apply in these investigations.
• Tools must respect these constraints.
Freedom of Speech

The Miller test (California v. Miller) established the criteria for obscenity in the US:

1. whether the average person, applying contemporary community standards would find that the work, taken as a whole, appeals to the prurient interest,

2. whether the work depicts or describes, in a patently offensive way, sexual conduct specifically defined by the applicable state law; and

3. whether the work, taken as a whole, lacks serious literary, artistic, political, or scientific value.
Civil or Criminal?

- Civil torts are violations of civil law (usually, contract disputes).
- On the Internet, this usually means copyright violations (though some of those are criminal).
- How are civil torts different from criminal charges?
  - not crimes!
  - penalties are monetary, not imprisonment, etc.
  - legal standards (for evidence, for decision) are lower
Expectations of Privacy

- Garbage is not private.
- Phone numbers dialed are not private (but they do have statutory protection).
- Email is similar, when stored by a third party (though protected by the SCA).
Flood Control

Any network where loops exist has to consider flooding.

- Messages in Gnutella have a unique identifier; some clients use it to identify and drop duplicates.
- Messages also have a hop count that is decremented each time they are forwarded.
Other Things We’ve Learned