Scott Wolchok

Crawling BitTorrent DHTs for Fun and Profit

Torrent Sites Under Attack

The Pirate Bay Guilty; Jail for File-Sharing Foursome

By Wired Staff April 17, 2009 | 2:28 am

Oscar Swartz reports.

ODO

Mininova Goes Legit, Saddens Everyone

wnloading
re still.

Police shut down OiNK BitTorrent site

by **Nicholas Deleon** on October 23, 2007 **30 Comments** 0 tweet Share

TorrentBits.org and SuprNova.org Go Dark

Posted by michael on Sun Dec 19, 2004 12:49 PM
from the **last-one-out-turn-off-the-lights** dept.

earlier today after a
community.
as arrested. Good luck,

TorrentFreak Home

isoHunt Loses US Lawsuit Against Movie Studios

Large-Scale BitTorrent Surveillance

"Spying the World from Your Laptop" @ LEET

Crawl Pirate Bay, scrape trackers

Tracked downloads for millions of IPs

"Worlds most resiliant tracking"

TPB added magnet links last year

No more .torrent files; get data from DHT

"no central tracker that can be down"

"don't need to rely on a single server"

http://thepiratebay.org/blog/175

In This Talk…

DHT crawling presents challenges and opportunities for torrent downloaders

Uses for crawling:

*AA can track users & torrents

Pirates can build search engines overnight!

Background

What's a BitTorrent DHT?

Remember your last torrent?

Search Torrents | Browse Torrents | Recent Torrents | TV shows | Music | Top 100

Preferences
Languages

☑ All ☐ Audio ☐ Video ☐ Applications ☐ Games ☐ Other

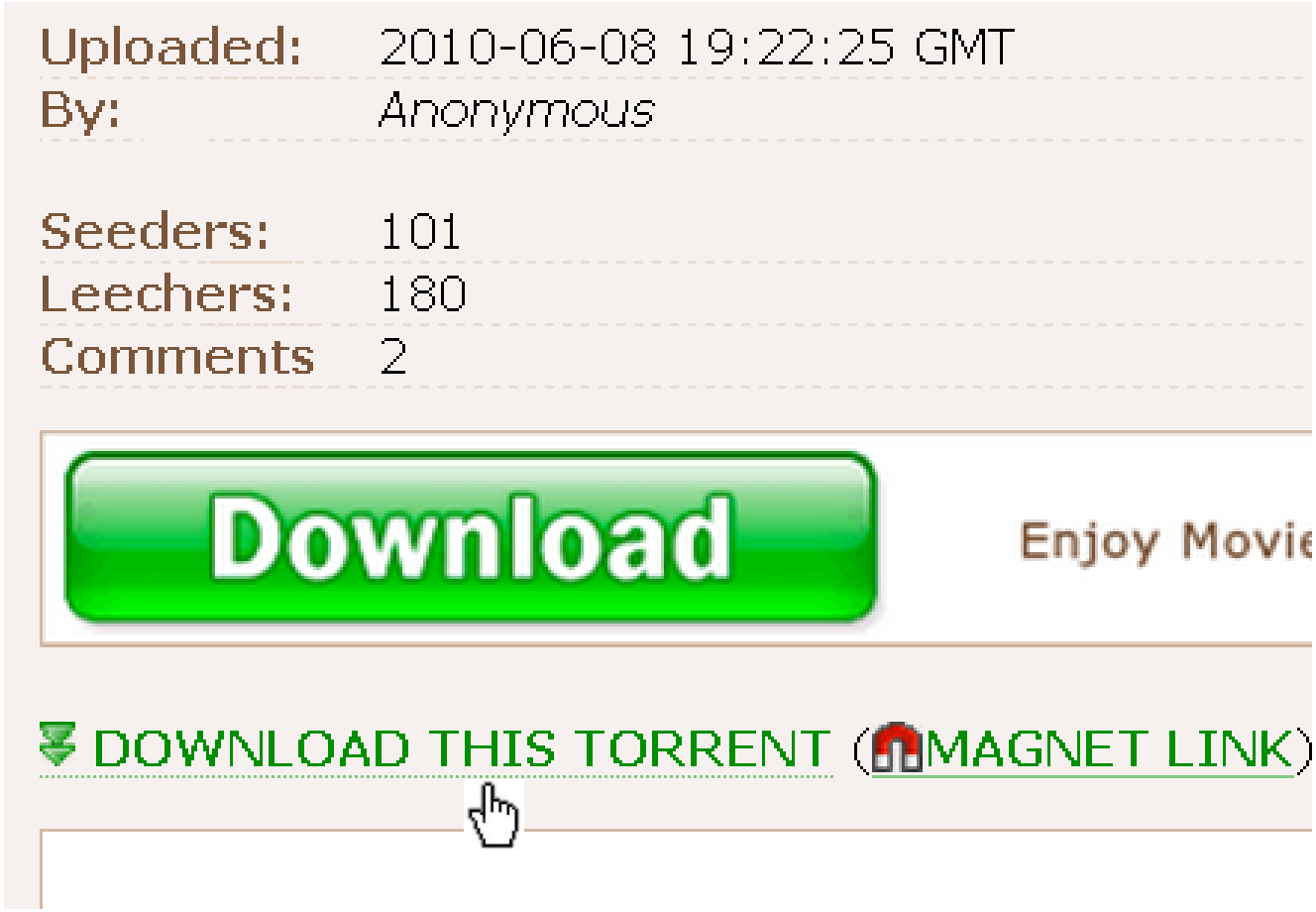
Pirate Search | I'm Feeling Lucky

How do I download?

Login | Register | Language / Select language | About | Legal threats | Blog
Contact us | Usage policy | Downloads | Doodles | Search Cloud | Tag Cloud | Forum | **TPB T-shirts**
SlopsBox | BayWords | BayImg | PasteBay | Pirate Shops | **IPREDator** | Pirate Chat

4.461.142 registered users. Last updated 07:36:04.
23.611.389 peers (14.251.431 seeders + 9.359.958 leechers) in 2.868.492 torrents.

How do you find peers?



The "old" way: trackers

List of trackers (servers) in the .torrent file

Torrent client sends "announce" to tracker

Tracker notes you're there & sends back peers

Example: tracker.openbittorrent.com

Distributed tracking

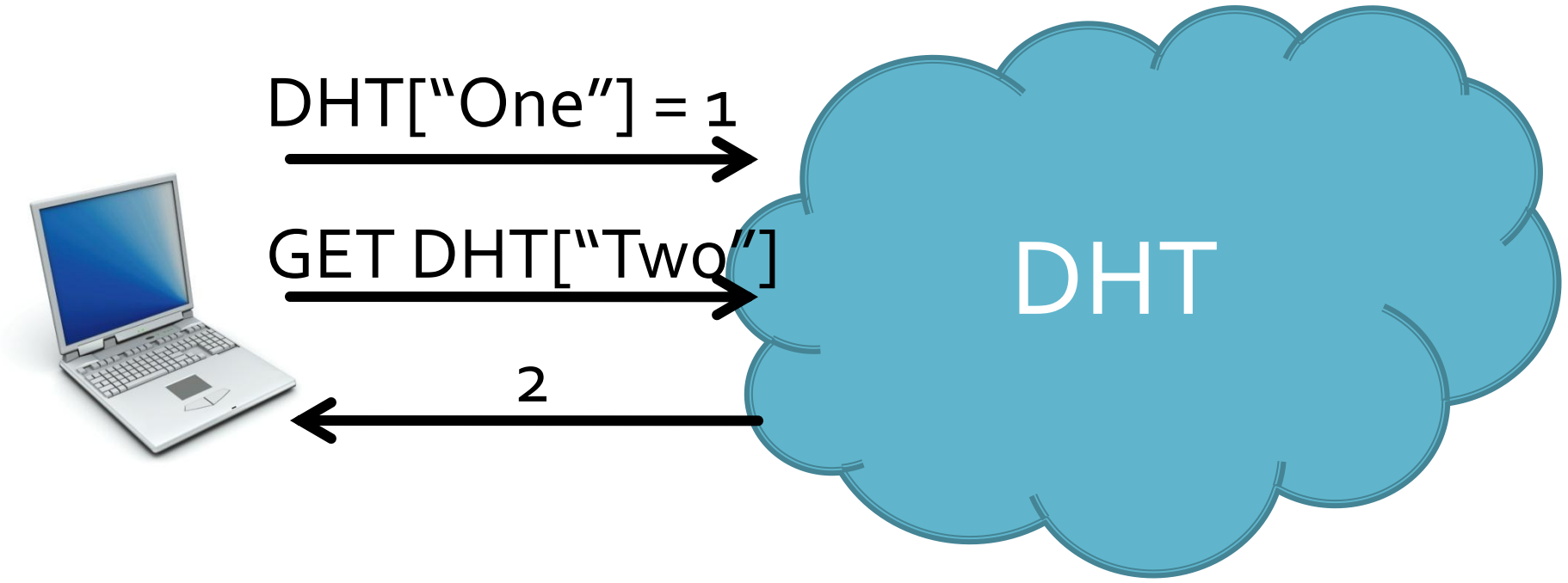
Trackers tend to go down (read: get sued)

Want something more reliable

Solution: distributed hash tables (DHTs)

DHTs in 2 minutes

P2P network that stores key-value pairs



More DHT Details

Peers & data have 160-bit IDs

Peer ID: “random”

Data ID: SHA-1 hash

Peers store data with similar IDs

DHT operations

PING

STORE(key, value)

FIND_NODE(id) – returns k closest peers (apply repeatedly)

FIND_VALUE(key) – like FIND_NODE, but returns value if known

DHT tracking

Replace the tracker with a DHT

Magnet URLs

magnet:?xt=urn:btih:cfa86e0e8f3831c24120b7f
ee7413b4da31ee748&dn=Linux+Mint+9.0+x8

Link straight to files, no .torrent (btih=infohash)

Find peers from DHT, fetch .torrent from them

Why? Legal shenanigans…

Mainline and Vuze

Two DHTs; 1 for Vuze, 1 for everyone else

Only cover Vuze in this talk to keep it simple

Should be possible to crawl Mainline as well

Our DHT Crawler

Reimplemented the Vuze protocol in C

Sybil attack: simulate 1000+ clients at once

Just sit and wait for values to come in

Cheaply captures 90%-99% of the DHT

"Defeating Vanish with Low-Cost Sybil Attacks Against Large DHTs"

Building a Search Engine

Design Overview

Crawl: download torrent data from DHT (filenames, sizes, peers)

Index and search: import into PostgreSQL, use its keyword search against filenames

Rank results by popularity – we've got lists of downloaders, so count them!

Torrent Descriptions

Problem: DHT has (SHA-1(infohash), peers) but we want the infohash!

Solution: torrent descriptions

Leaked into the DHT by Vuze client

d1:d35:**Fast & Furious[2009]DvDrip[Eng]-FXG1**:h20**:\xaf\x19x.a\xf2\xab\xc9;(Ib\xac\xoc\x1a\xa8\xc8\xob\x1dO**1:ri**646**e1:si**733484531**ee

Indexing

One pass over logs

Import into PostgreSQL

```
CREATE INDEX idx_name_gin_new ON
   torrent_descs_new USING
    gin(to_tsvector(\'english\', name))
```

Searching

Just a big SQL query & simple Web page

```
SELECT * FROM (
 SELECT DISTINCT ON (hash) name, hash, size, seeders, leechers,
              ts_rank_cd(to_tsvector('english', name), query, 0) AS rank,
              COALESCE(seeders, 0) +
              COALESCE(leechers, 0) as myrank
              FROM torrent_descs,
                       plainto_tsquery('english', %s) AS query
              WHERE to_tsvector('english', name) @@ query AND
                        hash is not NULL AND
                        COALESCE(seeders, 1) <> 0) AS results
              ORDER BY results.myrank DESC NULLS LAST LIMIT 100
```

Monitoring BitTorrent Users

Overview

Same crawl, just repeat it over time

Import makes 2 tables: peers and torrents

To map IPs <-> content, join on infohash!

SELECT … FROM peer_lists P,
torrent_descriptions D WHERE
SHA-1(D.infohash) == P.dhtkey …

How Well Does It Work?

Experiments

16 days of crawling, 3 crawls per day

Crawl: 8000 nodes over 2 hours

Should see ~20% of DHT

Search Coverage

Average crawl indexed 1 million torrents

The Pirate Bay: 2.8 million torrents

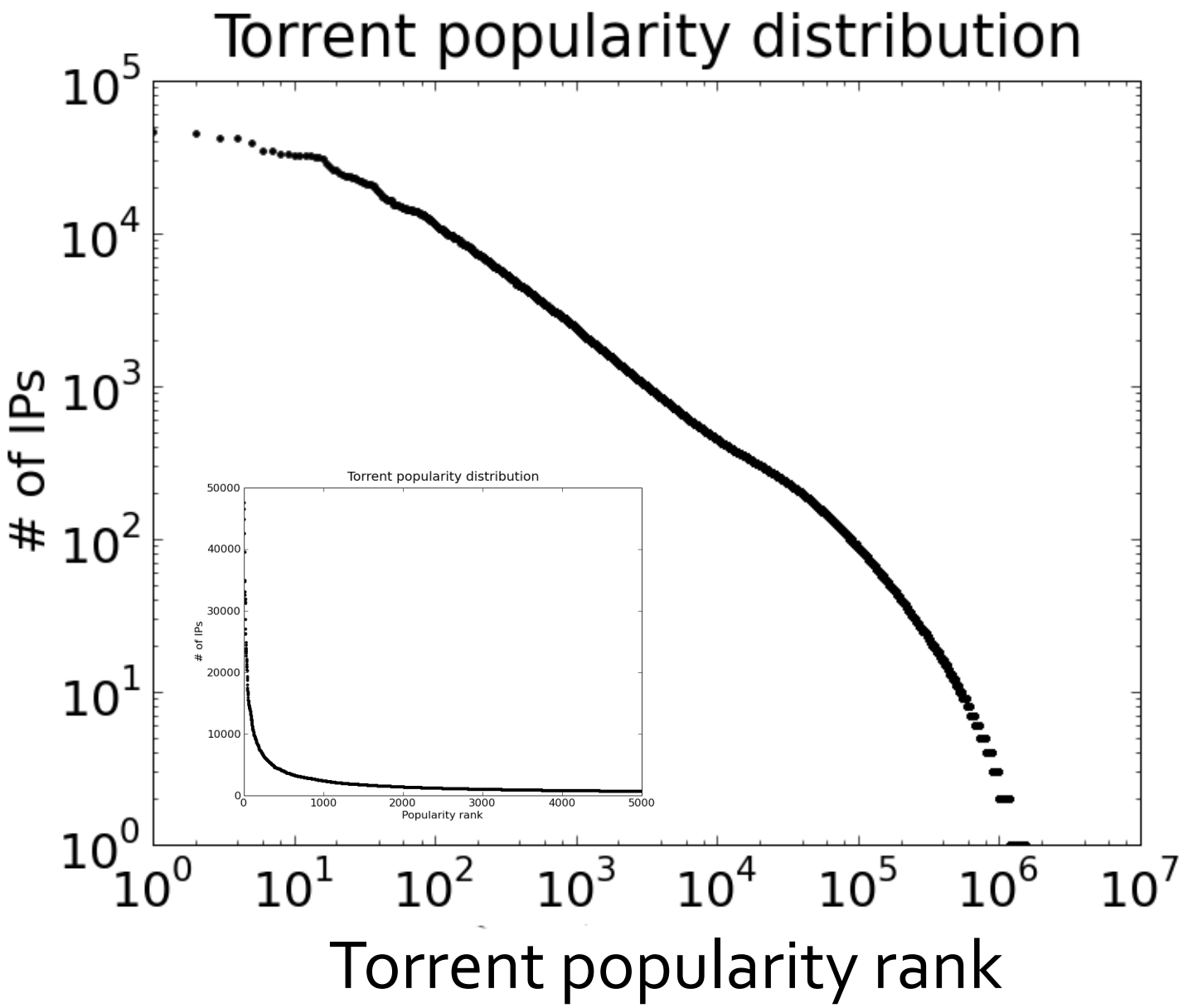
Bootstrap Time

Crawl: 81 minutes
Import crawl results: 13 minutes
Indexing: 6 minutes

Total: 100 minutes

Can go faster with more bandwidth

Trade off time vs. coverage


Torrent popularity distribution
# of IPs
Torrent popularity rank
Torrent popularity distribution
# of IPs
Popularity rank

Monitoring Coverage

15.1 million peer lists

3.6 million torrent descriptions

1.5 million torrents w/both peers and descriptions, mapped to 7.9 million IPs

Top 7 Torrents

Content	Downloads
The Pacific, Part 9 (TV)	47,612
Iron Man (movie)	46,549
Alice in Wonderland (movie)	44,922
Lost, Ep. 16 (TV)	42,571
Dear John (movie)	42,562
The Back-Up Plan (movie)	39,568
Lost, Ep. 13 (TV)	34,979

Top 1000 Torrents?

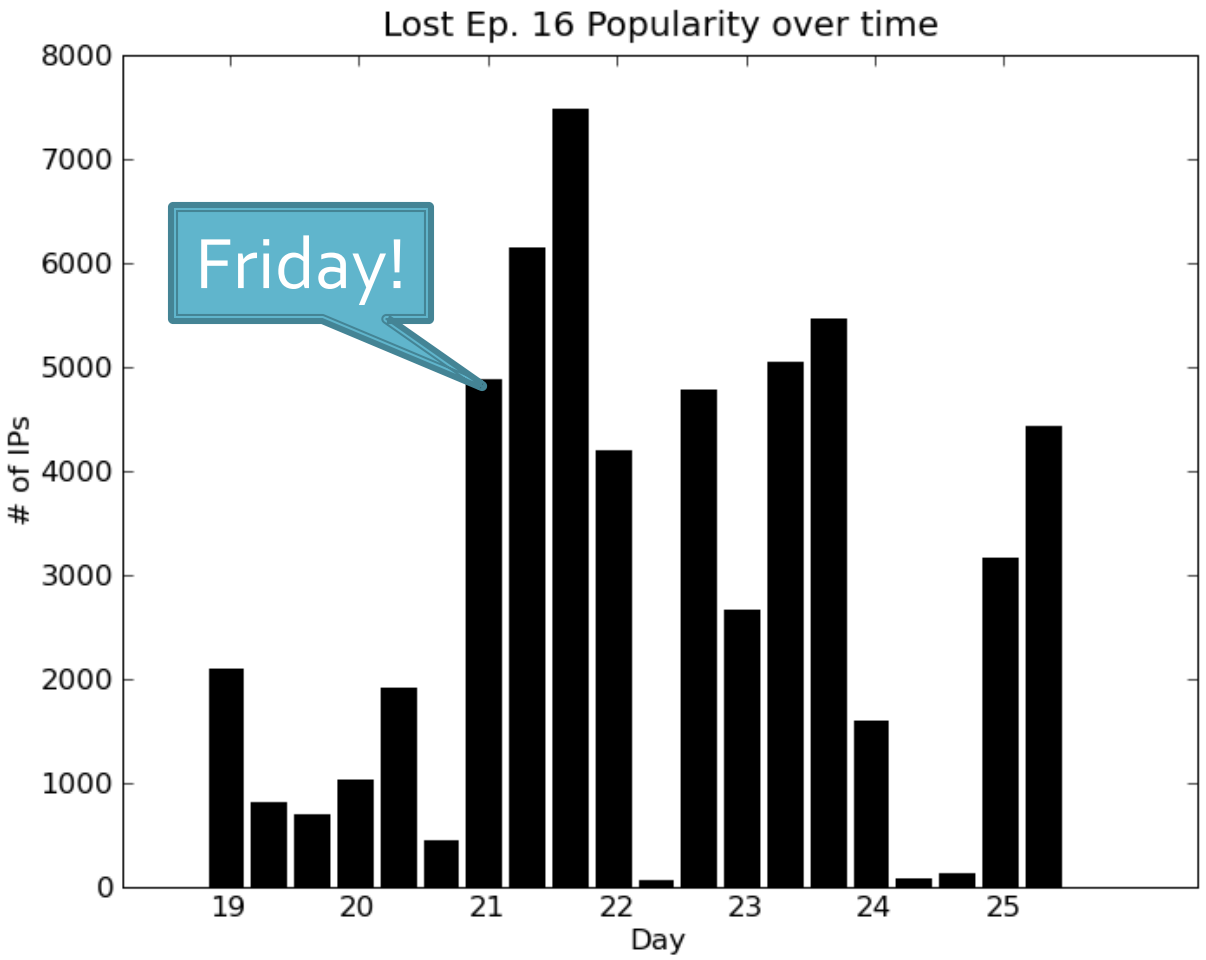
Inspected manually, none obviously not copyright-infringing

…except a subscription to a search for "XXX" on BtJunkie.

If everyone really is "just downloading Linux ISOs," they're not using Vuze to do it.

Popularity of Lost Ep. 16 over time

Air date:
May 18



Specific Users

User #1: all porn

User #2: also Iron Man, The Back-up Plan, Michael Jackson's Greatest Hits, Iron Man 2…

Even caught myself unintentionally seeding a free movie trailer

Conclusions

DHT crawling can **create search engines** & **monitor 8 million users**

Suing torrent sites is a distraction; we can rebuild them **fast**

DHTs won't help users hide

The future: DHT poisoning, more user lawsuits?

Links

http://wiki.vuze.com/w/DHT

http://www.cse.umich.edu/~jhalderm/pub/papers/unvanish-ndss10-web.pdf

http://scott.wolchok.org/dc18/dht/