



Internet Measurement Data Catalog

Colleen Shannon

cshannon@caida.org

<http://imdc.caida.org>

WIDE – Nov 3, 2006



Cooperative Association for Internet Data Analysis



DatCat Goals (1)

- to facilitate searching for and sharing of data among researchers
 - Index as much as possible, including datasets not publicly available
 - DatCat doesn't store any network data itself





DatCat Goals (2)

- to enhance documentation of datasets via a public annotation system
 - Easy place for anyone (not just the dataset creator) to provide additional information
 - Persistent reference that stays with the dataset (not a footnote in a paper)





DatCat Goals (3)

- to advance network science by promoting reproducible research
 - Paper X ran their detection algorithm on dataset X and had a false positive rate of 0.2. Using our algorithm on dataset Y, we get a false positive rate of 0.1. Therefore our algorithm is better. ...
 - Persistent handles to allow for consistent citing and comparison:

<http://imdc.datcat.org/collection/1-003M-5=AOL+500k+User+Session+Collection>





DatCat lets you...

- Find data for your research
- Annotate datasets to note features, background information, or bugs
- Cite data
- Contribute data (coming soon!)





DatCat Tour

DatCat: Browse - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

DatCat: Browse

Internet Measurement Data Catalog

Home | Browse | Search | Help | You are not logged in. | [Log in](#) | [Create an Account](#)

Path to data: [Browse](#) Select Data Select Packages Select Locations [Contact us](#)

Browse the Catalog

[Browse all 11 collections](#)

Featured Data Collections

[CAIDA skitter AS Links Topology](#) - 2346 files, starting 2000-01-02
Autonomous system (AS) topology derived from skitter traces (an AS is roughly an ISP). Possible uses include studying statistical and topological properties of AS graphs, constructing realistic Internet topologies for modeling and simulation, and studying AS relationships. Data collection has been continuing for over 6 years as of 2006.

[CAIDA Witty Worm Data, public access](#) - 7 files, 2004-03-20 to 2004-03-25
Information useful for studying the spread of the Witty worm, as observed by the UCSD Network Telescope over a 5-day period in Mar 2004. This dataset consists of public-access files that do not individually identify infected computers. Data available include time, duration, country, and connection speed distributions of infected hosts. This public-access dataset does not include packet traces of traffic generated by infected hosts. Possible uses include modeling worm propagation. Statistics: 55,909 infected IP addresses.

[CAIDA OC48 Traces 2003-04-24](#) - 26 files, 2003-04-24 to 2003-04-24
Anonymized packet header traces (but no packet payload) collected in both directions of an OC48 link at AMES Internet Exchange (AIX) on Apr 24, 2003 (1 hour). This link is a west coast peering link for a large ISP. Possible uses include research on the characteristics of traffic, including application breakdown, security events, geographic and topological distribution, and flow volume and duration. Statistics (both directions): 13GB of traces, 203 million packets, and 96GB of observed IP traffic.

[CAIDA Backscatter-2004-2005](#) - 63 files, 2004-05-26 to 2005-12-01
Information useful for longitudinal study of denial-of-service (DoS) attacks. This dataset consists of 5.5 billion IPv4 packets sent by DoS attack victims in response to spoofed attack traffic. This backscatter from victims was collected by the UCSD Network Telescope, one week of data per quarter, between May 2004 and November 2005. Possible uses include modeling DoS attacks, understanding victim populations, and using real packet traces to validate algorithms for detecting or classifying malicious traffic. This set is particularly valuable because it is extremely challenging to artificially generate the kind of real-world noise present on the Internet.

Other Recently Contributed Collections

[AOL 500k User Session Collection](#) - 10 files, 2006-03-01 to 2006-05-31
Web queries to AOL search engine

[CAIDA Code-Red Worm Dataset](#) - 14 files, 2001-07-19 to 2001-08-19
non-sensitive summaries on worm spread

[CAIDA Backscatter-TOCS](#) - 231 files, 2001-02-01 to 2004-03-06
denial-of-service backscatter 2001-2004

[CAIDA Witty Worm Data, restricted access](#) - 132 files, 2004-03-20 to 2004-03-25
raw packet traces and summaries

[CAIDA OC48 Traces 2003-01-15](#) - 28 files, 2003-01-15 to 2003-01-15

Browse Collections by Keyword

- [active](#)
- [anonymized](#)
- [AOL](#)
- [ARTS](#)
- [AS](#)
- [AS links](#)
- [background radiation](#)
- [backscatter](#)
- [Backscatter-2004-2005](#)
- [Backscatter-TOCS](#)
- [BGP](#)
- [blackhole address space](#)
- [CAIDA](#)
- [Code-Red](#)
- [Code-Redv2](#)
- [CodeRed](#)
- [CodeRedII](#)
- [CodeRedv2](#)
- [Crypto-PAN](#)
- [DAG](#)

Done 1.402s





Collaboration

- **Current:**
 - CRAWDAD: Community Resource for Archiving Wireless Data at Dartmouth
 - UCSD-CSE

- **Future:**
 - Abilene Observatory
 - ICSI
 - RouteViews





Next Steps

- Currently testing programmatic contribution interface
- Add support for Papers (specialized collection)
- Add support for tools
- GUI contribution interface





For more information

- DatCat: <http://imdc.datcat.org/>
- General questions and comments
 - info@datcat.org
- Announcements
 - user-announce@datcat.org
- Contribution beta-test
 - contribute@datcat.org



Security Research Overview



Cooperative Association for Internet Data Analysis

Current Security Research

- Nyxem/Blackworm/KamaSutra/MyWife
 - <http://www.caida.org/analysis/security/blackworm/>
- Spamscatter
- Botnet Economics
- Worm Risk Analysis



Nyxem Virus Spread

- Nyxem email virus begins to spread January 15, 2006
- Infected hosts automatically generate an http request for an online counter
- Counter then displays number of infectees...
- ...until word about the counter spreads



Analysis Complications

- Folks viewing the counter to see how many machines were infected (raise estimate)
- NATs (lower estimate)
- Web proxies (lower estimate)
- DDoS attacks on the counter (raise estimate)



Strategy

- Remove referer/browser strings set by common DDoS tools (91.1% of all hits)
- Remove requests for pages different from the one accessed by the virus (0.2%)
- Remove any request with a referer string (virus did not use one in its probes) (0.8%)
- Remove requests from invulnerable Operating Systems: MacOS, Unix, cell phone, and PDA devices (0.03%)

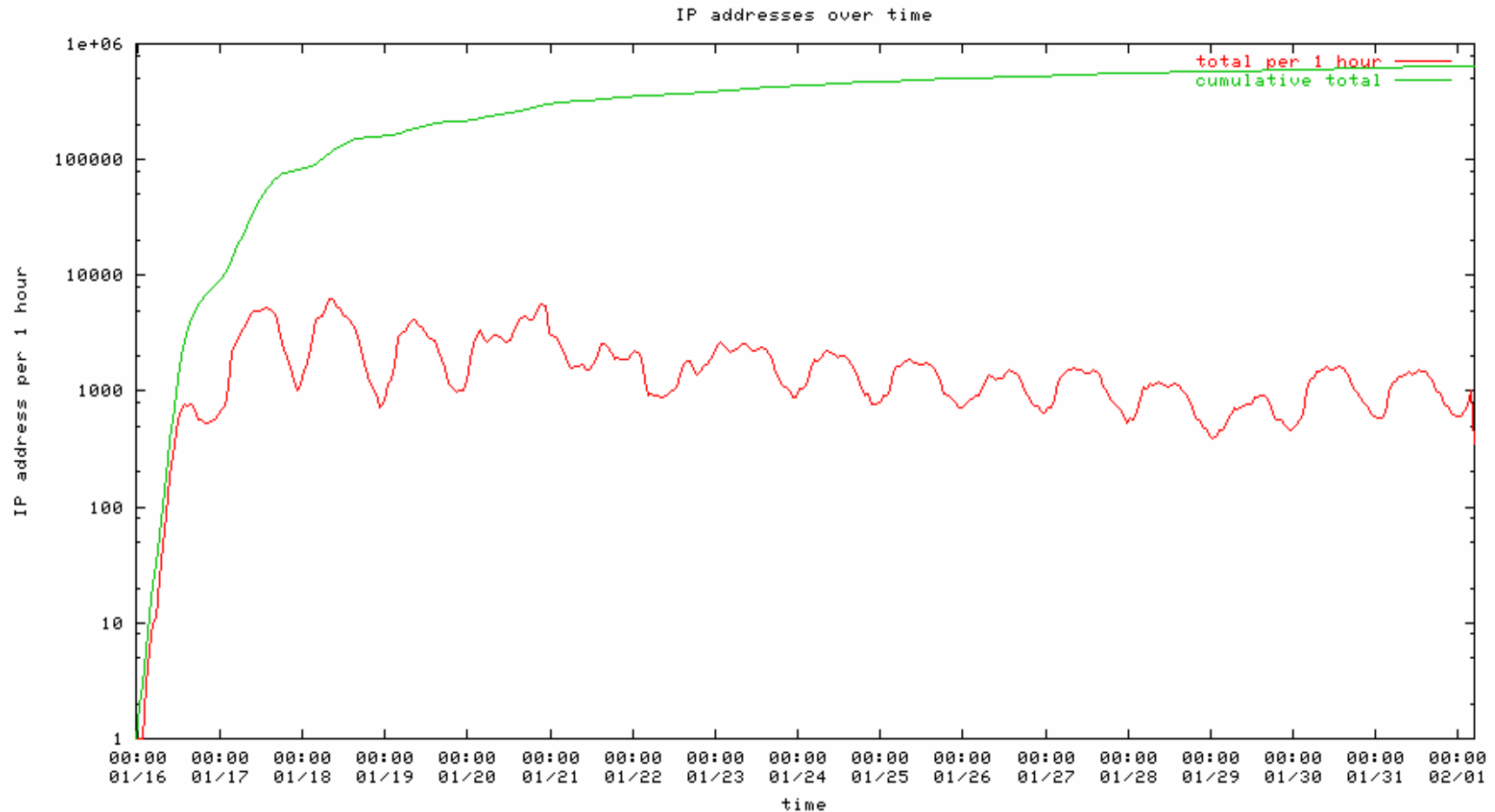


Results

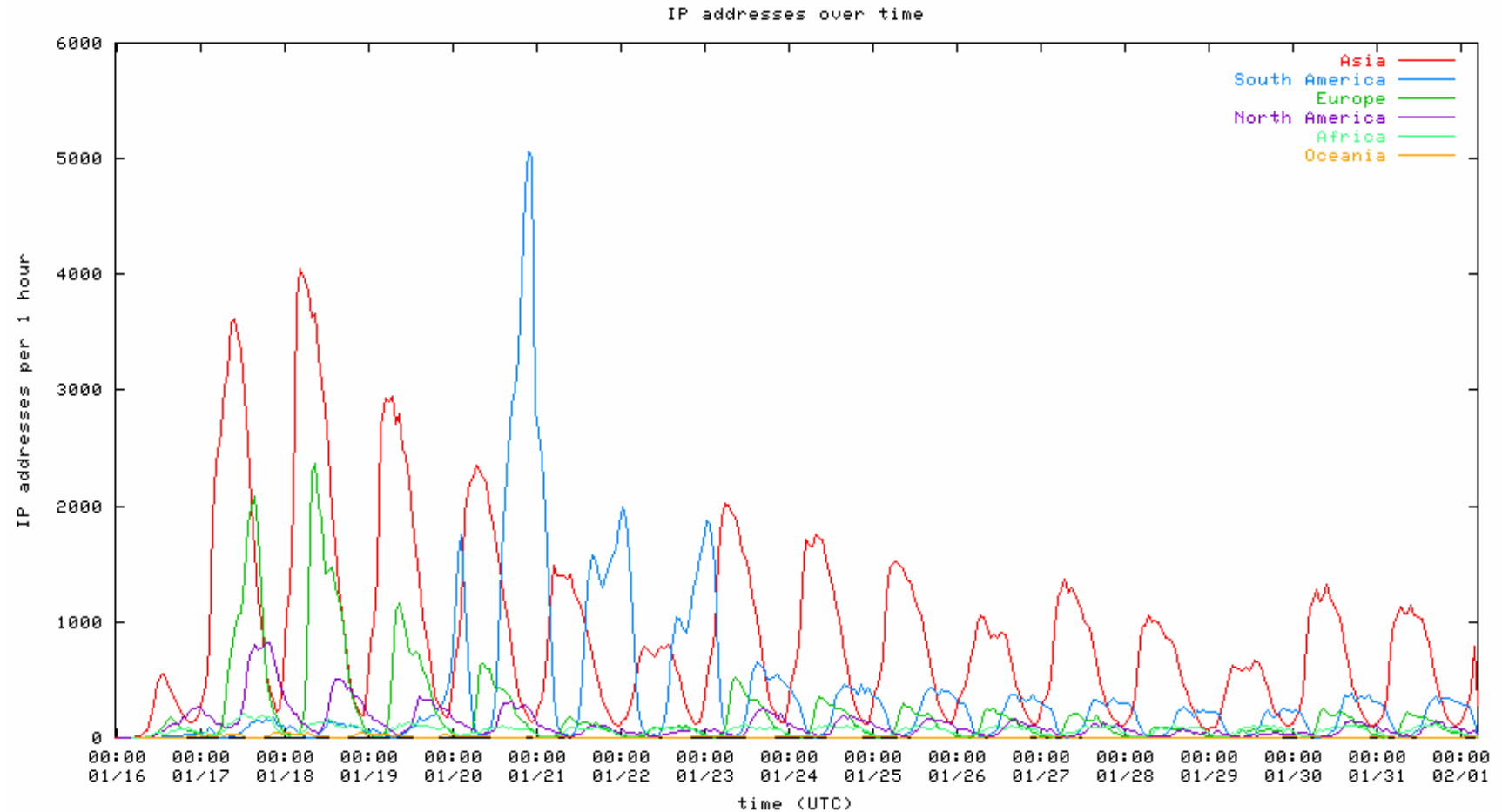
- Nyxem victim estimate: between 469,507 and 946,835 (3.2%-6.4% of original log entries)
- 45,401 Nyxem victims (10%) had concurrent spyware and/or botnet infections advertised in their browser string



Nyxem Spread



Nyxem Spread



Cuttlefish Animation

<http://www.caida.org/analysis/security/blackworm/animations/nyxem-hosts-both-O2.mov>



Cooperative Association for Internet Data Analysis

Acknowledgements

- Thanks to our sponsors:



SDSC



- Thanks also to: Gadi Evron, Paul Vixie, Joe Stewart, Mikko Hypponen, Swa Frantzen, Randy Vaughn, Chris Jackman, Jason Nealis, Rob Thomas, and Lorna Hutcheson for providing us with data and insight into the spread of the virus.

