Effective Flow Filtering for Botnet Search Space Reduction

Walsh, Lapsley, Strayer
BBN Technologies

CATCH 2009

March 3, 2009

Presented by Robert Walsh

This work was funded by DHS S&T

'Cyberwar' Emerges Amid Russia-Georgia Conflict

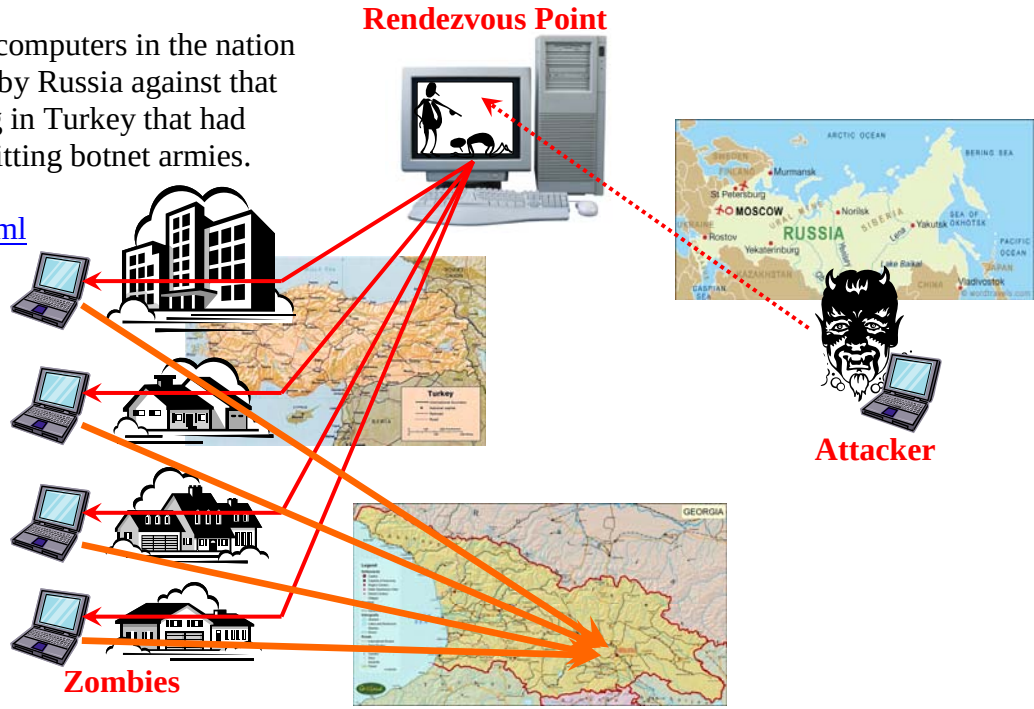
Georgia's recent conflict with Russia over the fate of two separatist provinces brought with it a first in international cyber-warfare, as Georgia faced a slew of Internet attacks.
…
Georgia's Internet system was crippled, as hackers manipulated computers to flood government, news, and information Web sites in a way that renders them useless.

http://www.pbs.org/newshour/bb/europe/july-dec08/cyberwar_08-13.html

Russian "hacktivists" used Turkish botnets to attack Georgia

Many of the cyber attacks that were launched against government computers in the nation of Georgia -- which coincided with the real-world military attacks by Russia against that country last month – were actually carried out by computers sitting in Turkey that had been captured by Russian "hacktivists" and drafted into huge, unwitting botnet armies.
…
http://www.gsnmagazine.com/cms/features/news-analysis/1042.html

A Botnet is a collection of infested PCs under the control of someone else.



Overview

- **Describe IRC-based Botnet structure**
  - **Actors, roles, and communications**
- **Describe our Botnet Identification Approach**
  - **Proactively focus on command and control communications (mission orders)**
  - **Analyze network traffic to identify the (multicast) C2 comms.**
- **Drill in on data reduction filtering portion of approach**
  - **Quickly reduce size of haystack without losing the needle**
- **Describe interesting behavior**
  - **Very persistent and well-formed comms that correspond to C2**

Why are Botnets Important?

BBN TECHNOLOGIES

- **In addition to spam, and economic losses, botnets can impair life-and-limb environments.**

3 accused of inducing ill effects on computers at local hospital

By Maureen O'Hagan

Seattle Times staff reporter

One day last year, things started going haywire at Northwest Hospital and Medical Center.

Key cards would no longer open the operating-room doors; computers in the intensive-care unit shut down; doctors' pagers wouldn't work.

This might have been just another computer-virus attack, a common and malicious scheme that sometimes is done for little more than bragging rights. But federal officials say it was something far more insidious.

It turns out the Seattle hospital's computers — along with up to 50,000 others across the country — had been turned into an army of robots controlled by 20-year-old Christopher Maxwell of Vacaville, Calif., according to a federal indictment issued Thursday. And Maxwell, along with two juveniles, earned about $100,000 in the process, court documents state.

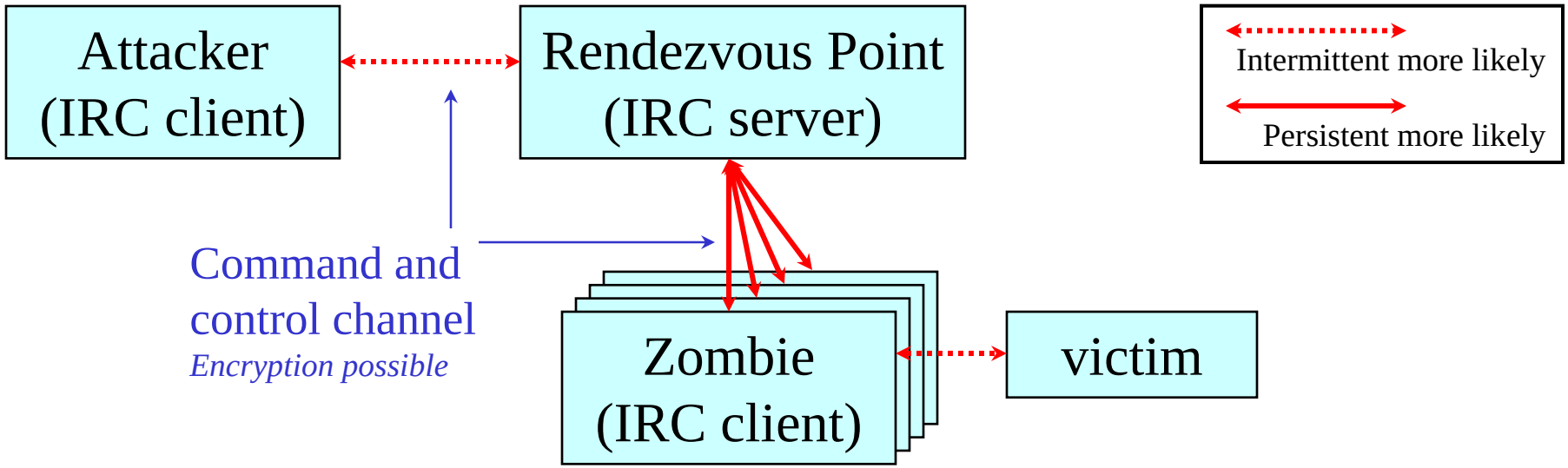
The trio had created a "botnet," a phenomenon that is on the cutting edge of computer crime, federal officials say.

http://seattletimes.nwsource.com/html/localnews/2002798414_botnet11m.html

How to Catch a Botnet

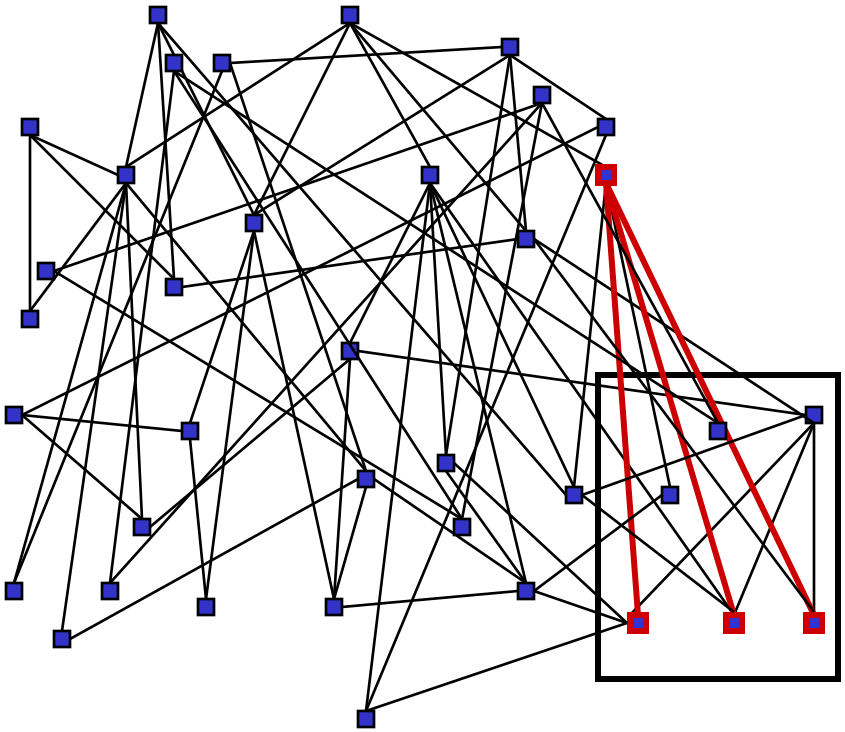
- **Variety of methods used to detect botnets**
  - Host-based
    - Use host-based scanning software to examine hosts for rootkits, trojans, and other malware
    - Construct Honeynets to surreptitiously join a botnet
  - Network-based
    - Use snort to examine payloads for IRC commands
    - Monitor free DNS hosting services for rendezvous
    - Analyze traffic for patterns and correlations
- **Each method has strengths and weaknesses**
- **Our work concentrates on traffic analysis**

Command and Control



- **IRC is still the dominant C2 technique**
  - For scope of this work, exploit IRC characteristics to exclude "likely-safe" traffic and to discover botnet clusters
- **As botnet C2 infrastructures change, we must continue to discover fundamental characteristics**

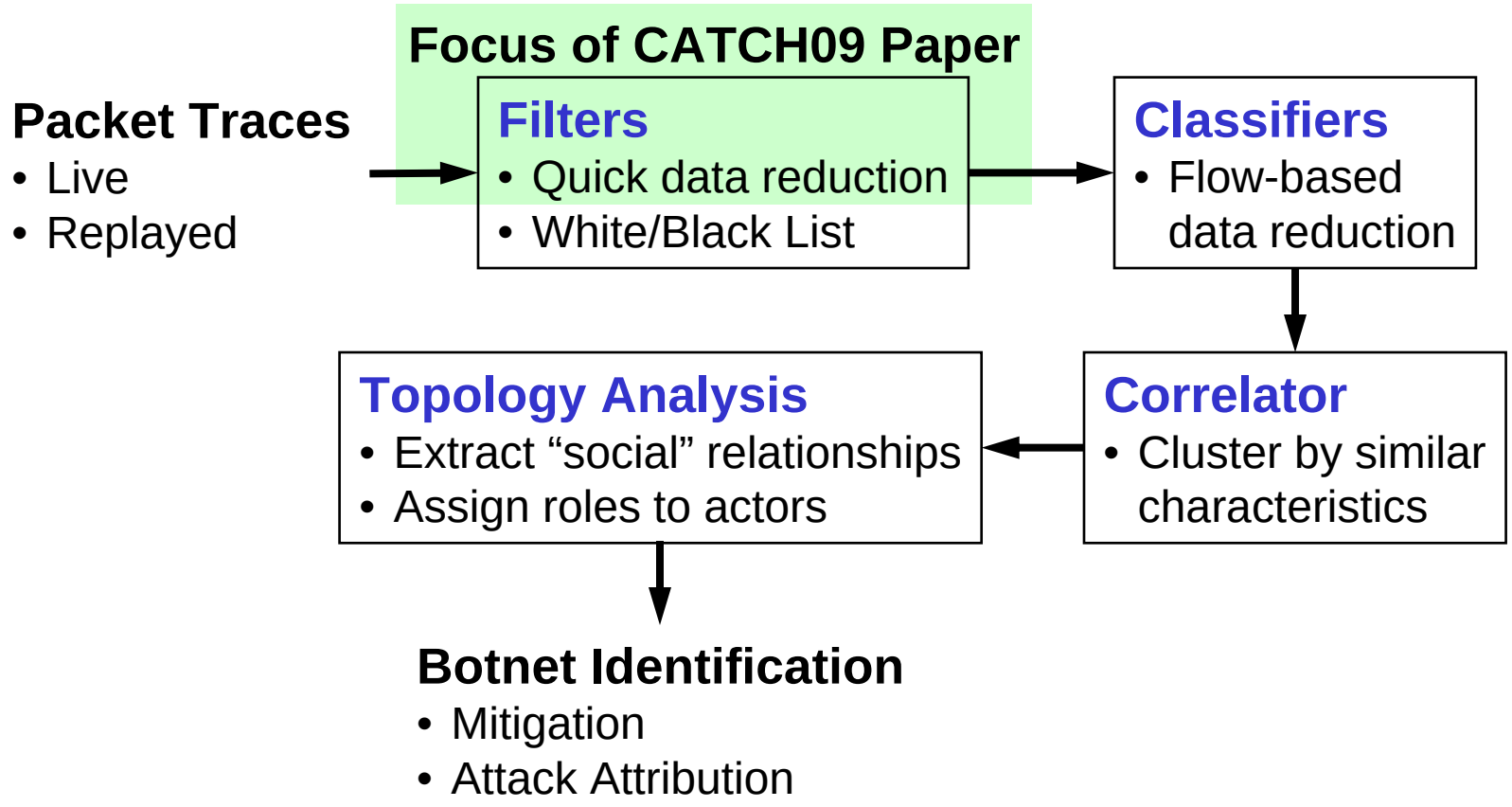
Traffic Analysis Botnet Detection



1. Monitor traffic within a region
2. Classify and filter out unlikely flows
3. Correlate flows to form a cluster
4. (Exchange with other monitors to widen the cluster)
5. Analyze the social network to piece together the botnet structure

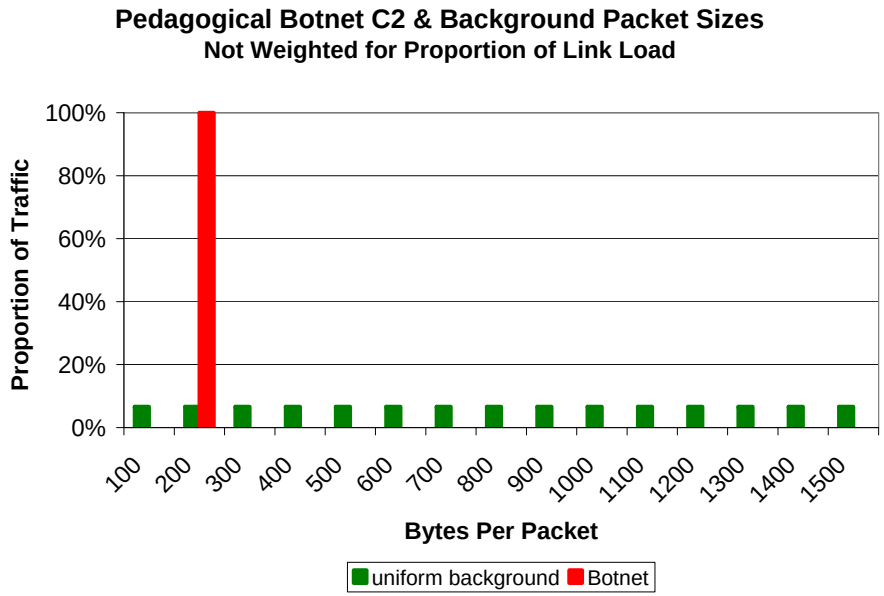
Today We Discuss Filtering -- One Step in a Larger System

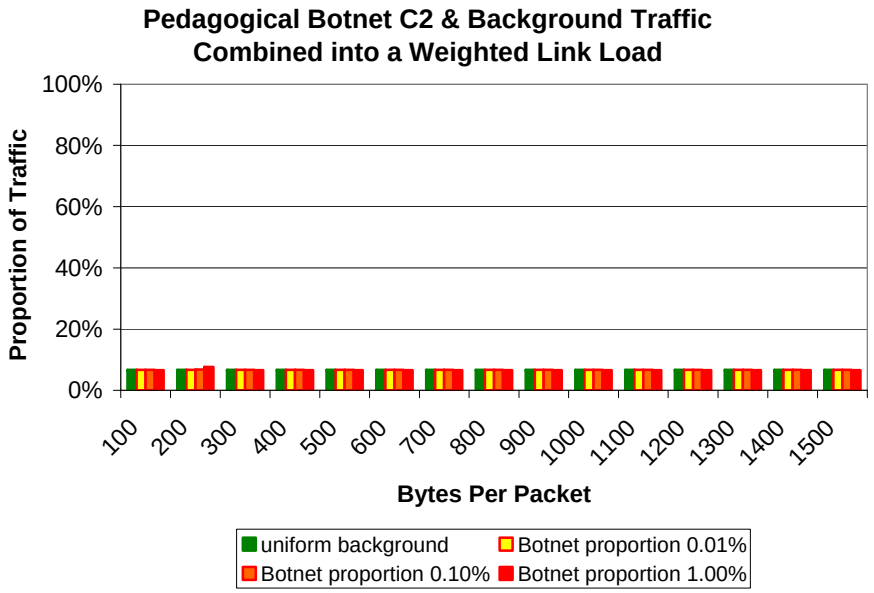
Processing Pipeline Overview



Filtering is One Tool for Addressing System Scalability

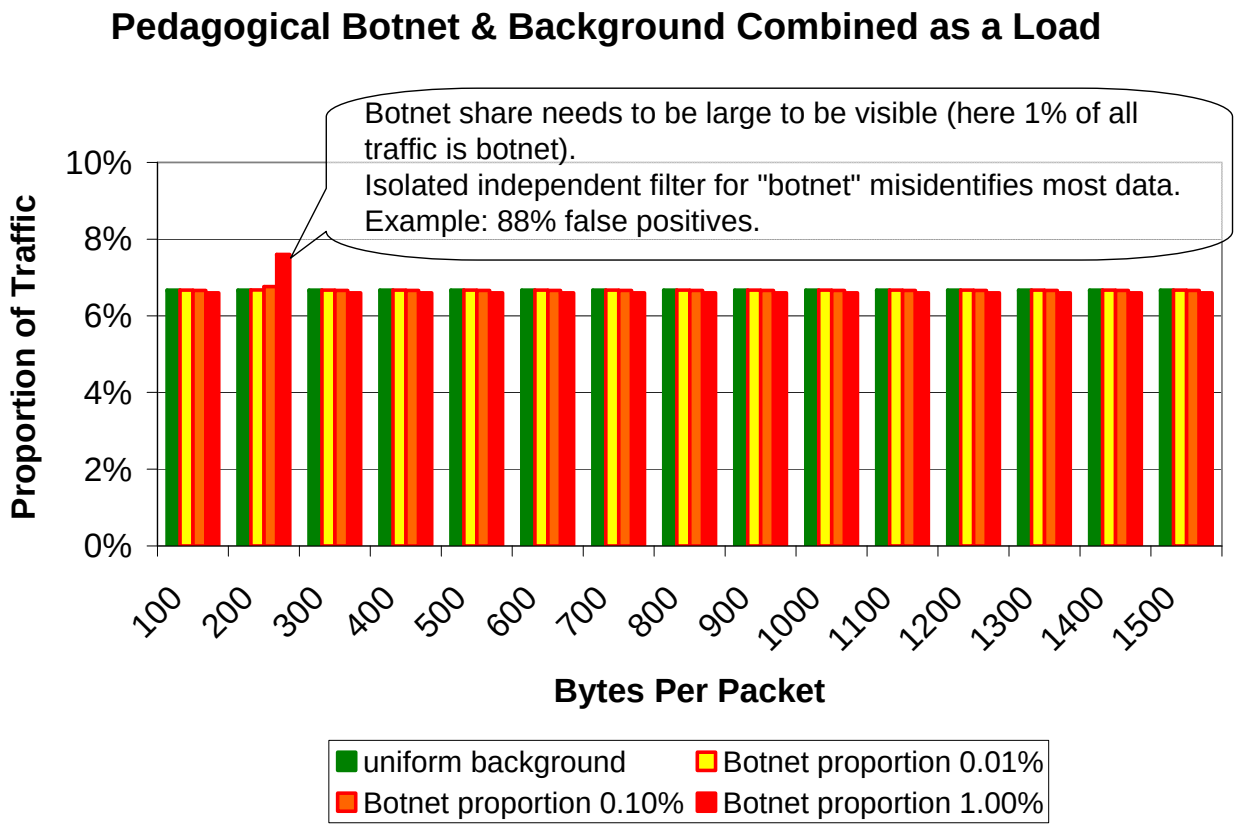
Botnet C2 Identification is Hard





- **Botnet & Background depictions artificial for illustration**
- **Botnet clearly distinguished in flow characteristics (left)**
- **Botnet C2 hidden in link load even when it is large relative to adversary's surreptitious goals (right)**
  - Hard to know if anything is going on. Hardly noticeable.

Botnet C2 Identification is Hard…



- **Takes high botnet C2 load to discern effect on link load**
- **A lot of other traffic gets swept in as false positives**

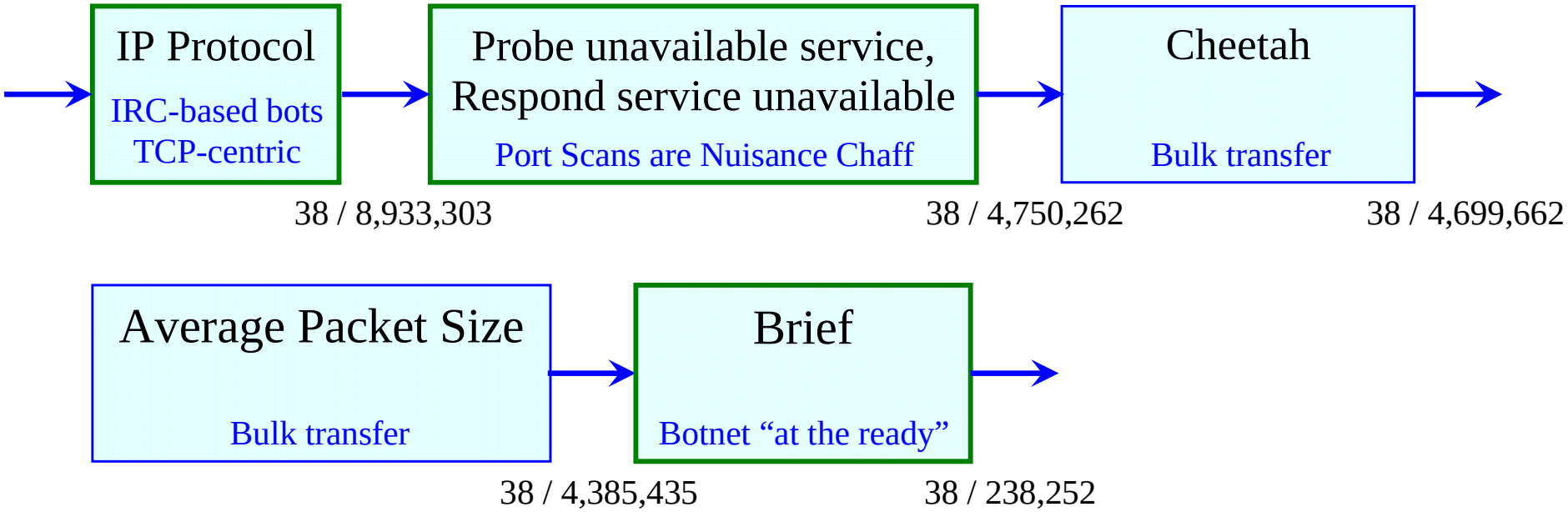
Raw Packet Traces (Haystack)

- **TCP/IP traces from Dartmouth Campus Wireless**
  - A “CRAWDAD” data set
  - Variety of locations (dorm, library, academic buildings)
    - We used all, and did not prejudge likelihood of activity
      - privacy of dorm; anonymity of library; *etc*.
  - Gathered Nov 1, 2003 through Feb 28, 2004
    - Packet headers: ~164 Gbytes compressed (~3.8x larger uncompressed)
      - Note that converting packets to flows reduces data set
      - Filters for data reduction discussed here are flow-based
    - About 228M total half-duplex flows in 4 months
  - All IP addresses were obfuscated, no payloads

Botnet Traffic Traces (Needle)

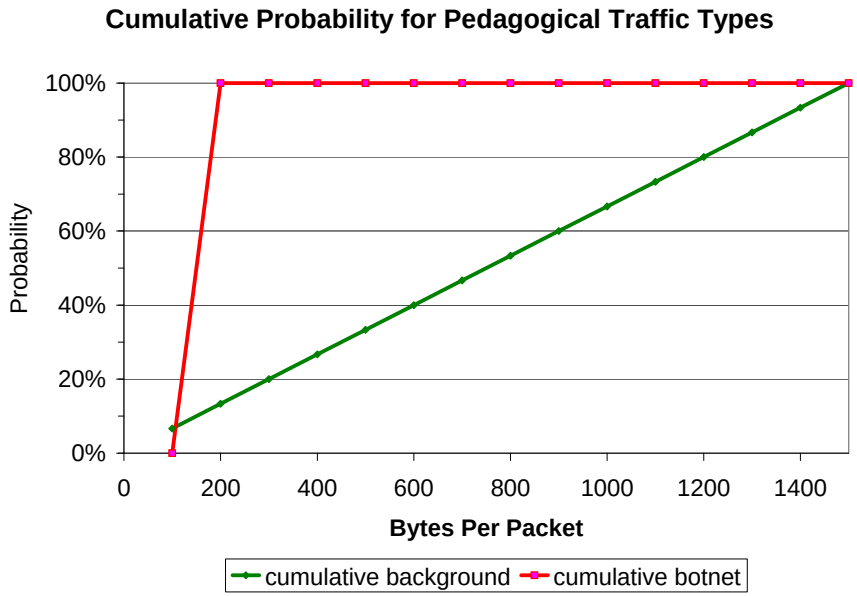
- **Built a botnet testbed**
  - Need to have “ground truth” traffic traces
  - Allow us, not outsider, to control the internally-deployed malware
  - Reimplemented “Kaiten” bot client
    - Bound risk taken within our enterprise (c.f. code/binary from Internet)
  - 13 zombies, 1 attacker, 1 IRC server, 1 update server, 1 victim
- **The botnet traces were overlaid with Dartmouth traces**
  - 74 half-duplex flows appropriately translated to the tenth day of Dartmouth data at one of the wiretaps
    - 8.93M half-duplex flows at that wiretap
    - 1.34M half-duplex flows in first 10 days
  - A subset of the 74 were botnet C2
    - attacker – IRC server flows ; zombie – IRC server flows

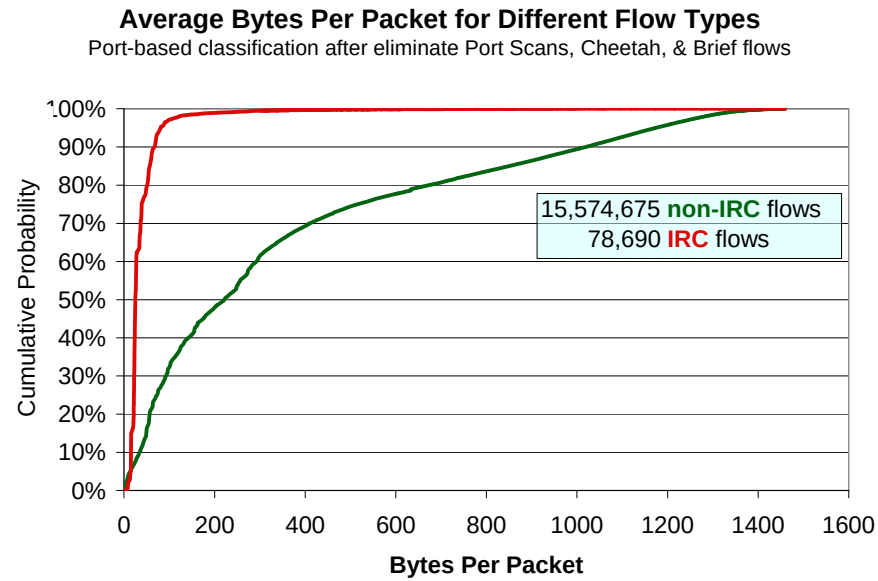
Filters for Data Reduction



- **Quickly reduce data, making later (expensive) steps feasible**
  - 37-fold reduction in data, in addition to packet-to-flow reduction
  - Some steps provide most of flow reduction; others help and do no harm
    - Bulk Transfer filters reduce size of packet-level forensic archive
  - Port scanning is something to pursue, but is not specific to botnet C2
- **All ground-truth botnet C2 flows retained**

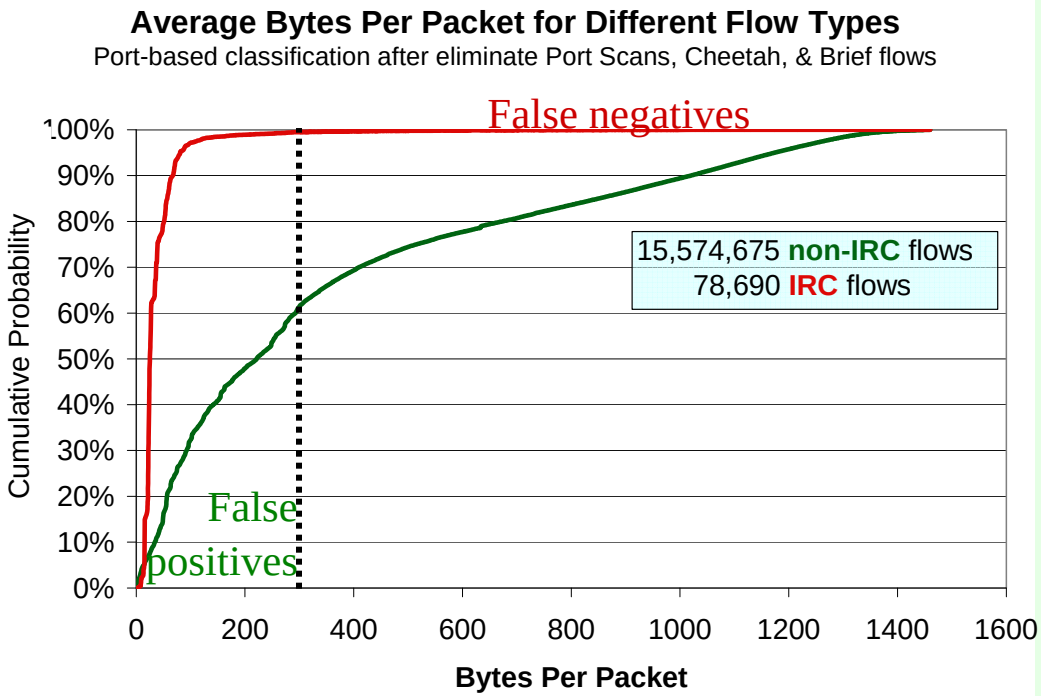
Filter Creation





- **Curve separation indicates a good classification metric**
- **Differences in shape (slope) locate possible discriminatory values of the metric.**
- **Easier to see in CDFs when PDFs overlay one another**
- **Packet size is a good classification metric**

Bias to Avoid False Negatives



- **Always false positives and false negatives**
- **Accept less data reduction, to reduce risk miss botnet flow**
- **Other filters assist with data reduction**
- **300 bytes yields ~0.5% false negatives**

Average bytes per packet	Cumulative Probability IRC	Cumulative Probability non-IRC	False Positives, Botnet Load 0.01%	False Positives, Botnet Load 0.10%	False Positives, Botnet Load 1.00%	False Negatives	Data Reduction
68	90.2%	23.9%	99.96%	99.62%	96.33%	9.8%	76.1%
81	95.2%	27.2%	99.96%	99.65%	96.58%	4.8%	72.8%
215	99.0%	49.3%	99.98%	99.80%	98.01%	1.0%	50.7%
311	99.5%	62.7%	99.98%	99.84%	98.42%	0.5%	37.3%
731	99.9%	81.6%	99.99%	99.88%	98.78%	0.1%	18.4%

better

better

most accepted traffic is nuisance chaff

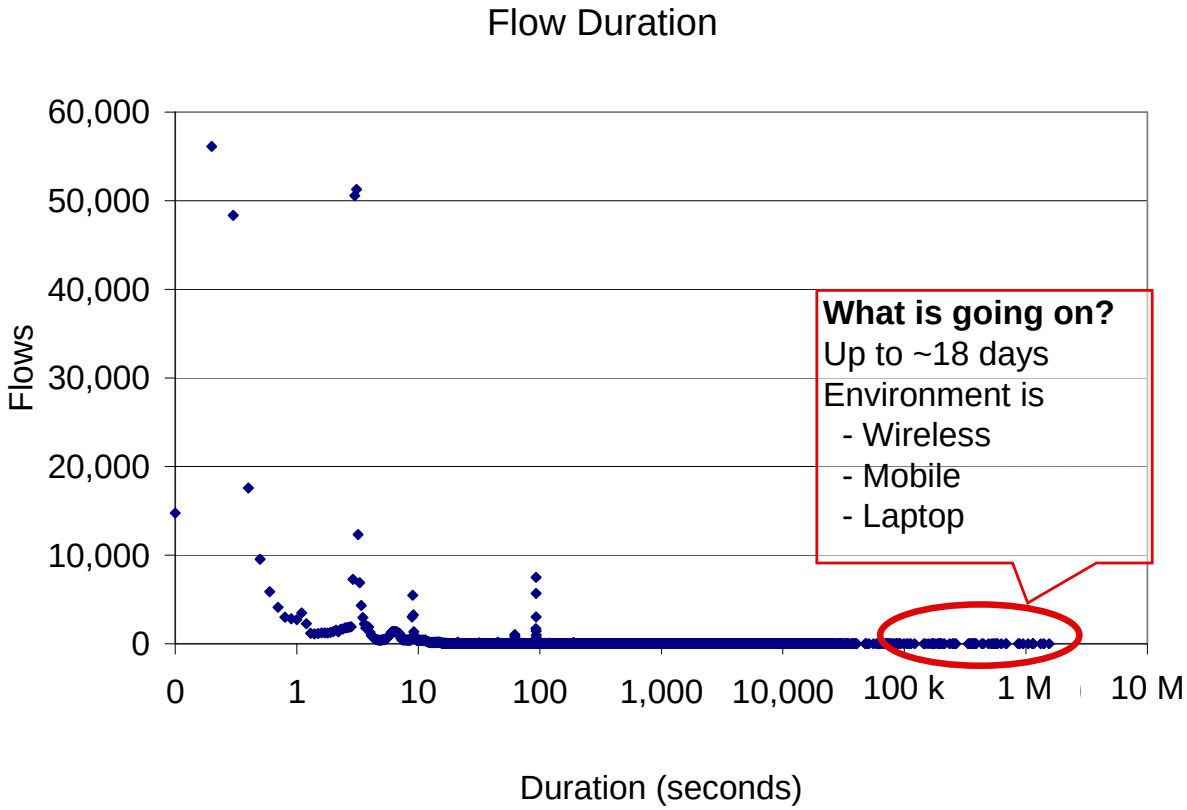
Want few bot flows to be missed

But also need to reduce data set size

Filtering Led to Associations

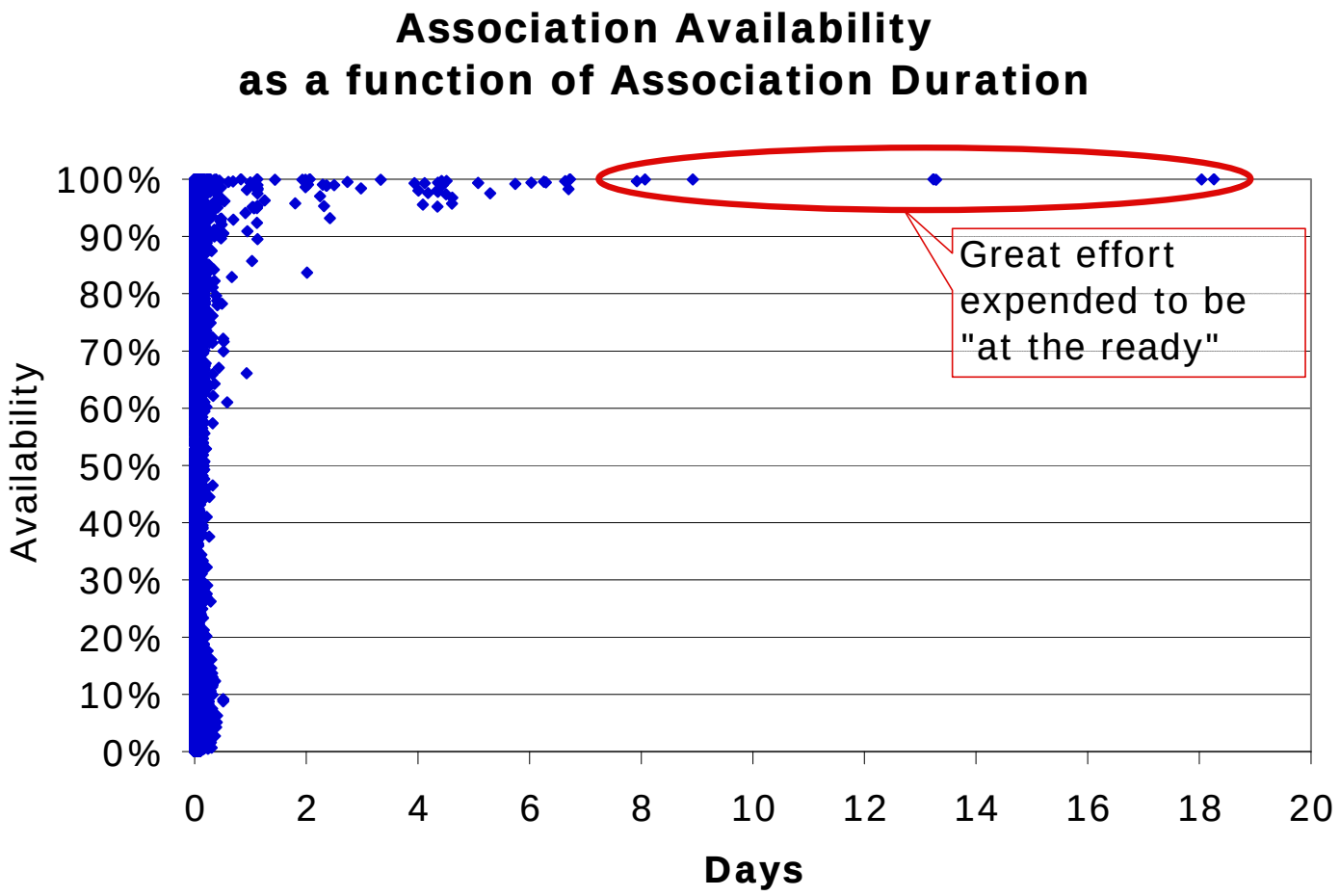
- **The paper contains additional details on filtering**
  - **Average flow data rate was used, to drop bulk transfers**
  - **Flow lifetime was used, to search for "at the ready" C2 networks.**
  - **Flow lifetime dominated data reduction**
  - **Port scan detection 2nd largest contributor to reducing number of flows for later stages to consider**
- **Investigating flow lifetime led to detecting other interesting behavior**
  - **How long-lived were these flows?**
  - **Some of these long-lived flows look similar (from, to, …)**
    - **Is it coincidence?**
  - **How many servers are involved?** **~1% of the servers**

Intrigue – Long Lifetime



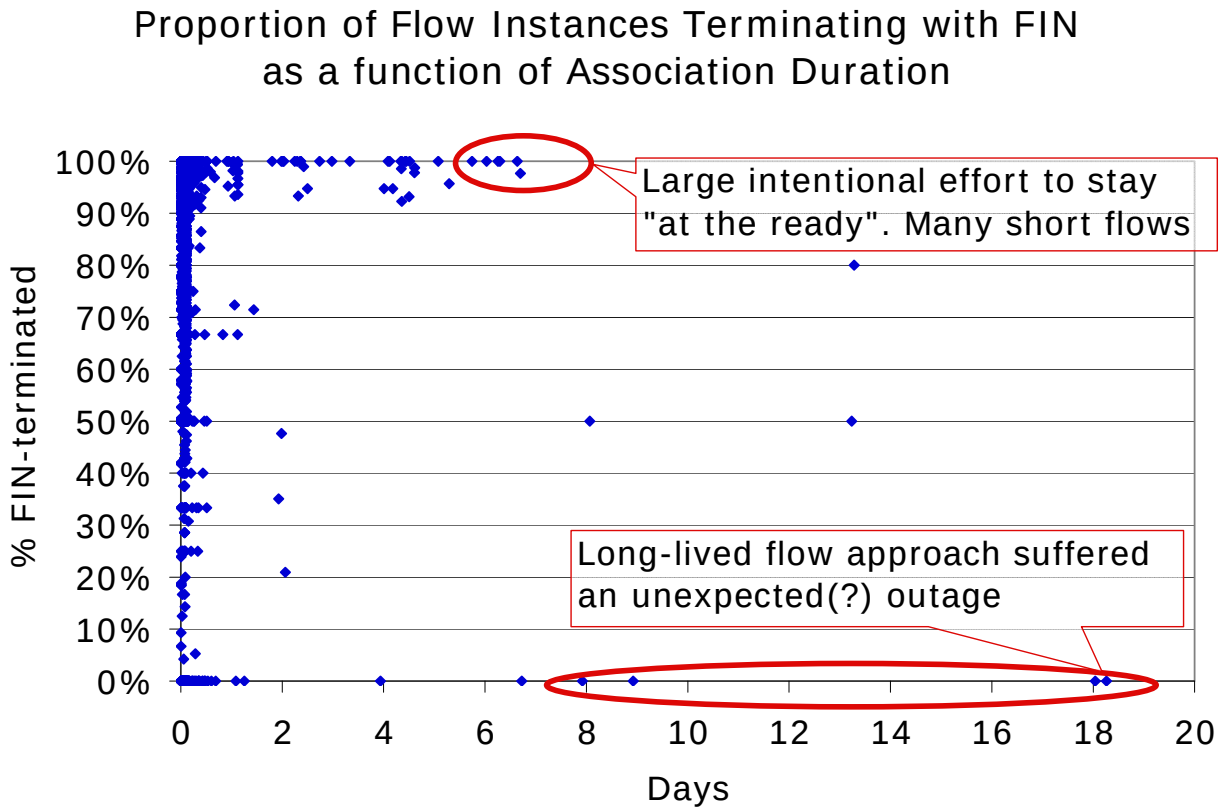
- **There are unexpectedly long-lived flows in an environment where expect devices to occasionally be turned-off or hibernating**

Associations – High Availability



- **High Availability, especially for a Wireless LAN environment**

Associations - Clean Close



- **Clean closes imply intentional use of multiple flows** (espec. W-LAN)
- **There are very persistent associations made up of many (1000's) well-formed (clean close) flows**

Summary

- **Filtering successfully reduces data set size for other computations, while preserving Botnet C2 flows**
  - **Hard problem. Goal is data reduction. Filtering is not silver bullet alone.**
  - **Flows take less space than packets.**
  - **Filters: "At the ready", scans, bulk transfer**
- **"At the ready" association behavior is very suspicious**
  - **Long-lived, Highly available, Persistent (#flows. clean close & re-open.)**
- **Solution involves more components in the system**
  - **Flow correlation to identify multicast participants (Botnet)**
  - **Analyze structure of multicast to assign roles (IRC server)**
- **Proactive Botnet C2 identification not replace the basics**
  - **Port-scanning as a sign of malware participation**