Understanding the Network-Level Behavior of Spammers

Anirudh Ramachandran

Nick Feamster

Georgia Tech

Spam

- Unsolicited commercial email
- As of about February 2005, estimates indicate that about 90% of all email is spam
- Common spam filtering techniques
  - Content-based filters
  - DNS Blacklist (DNSBL) lookups: Significant fraction of today’s DNS traffic!

State-of-the-art: Content-based filtering

Problems with Content-based Filtering

- Content-based properties are *malleable*
  - **Low cost to evasion:** Spammers can easily alter features of an email's content can be easily adjusted and changed
  - **Customized emails are easy to generate:** Content-based filters need fuzzy hashes over content, etc.
  - **High cost to filter maintainers:** Filters must be continually updated as content-changing techniques become more sophistocated

- Content-based filters are *applied at the destination*
  - **Too little, too late:** Wasted network bandwidth, storage, etc. Many users receive (and store) the same spam content

Network-level Spam Filtering is Robust

- Network-level properties are more fixed
  - Hosting or upstream ISP (AS number)
  - Botnet membership
  - Location in the network
  - IP address block
  - …

- **Challenge:** Which properties are most useful for distinguising spam traffic from legitimate email?

Very little (if anything) is known about these characteristics!

Studying Sending Patterns

- **Network-level properties of spam arrival**
  - From where?
    - What IP address space?
    - ASes?
    - What OSes?
  - What techniques?
    - Botnets
    - Short-lived route announcements
    - Shady ISPs
  - Capabilities and limitations?
    - Bandwidth
    - Size of botnet army

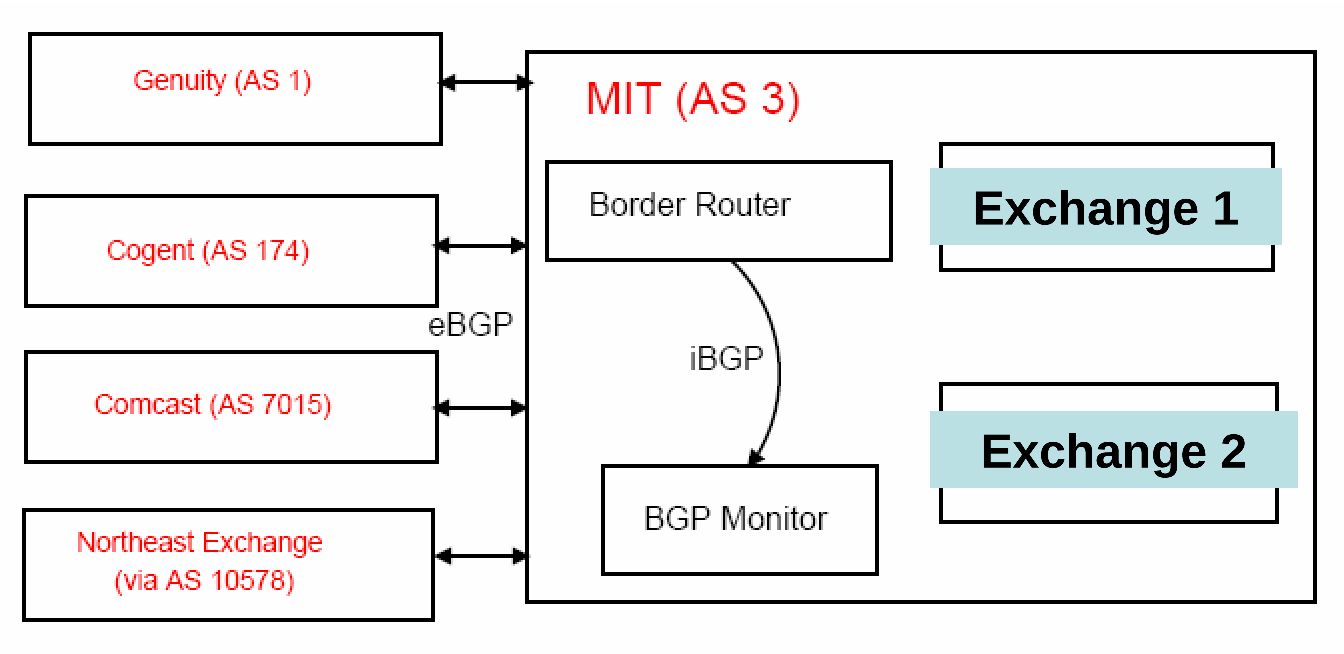
Spamming Techniques

- Mostly botnets, of course
- Other techniques, too…
- We’re trying to quantify this
  - Coordination
  - Characteristics
- How we’re doing this
  - Correlation with Bobax victims
    - from Georgia Tech botnet sinkhole
  - Other possibilities: Heuristics
    - Distance of Client IP from MX record
    - Coordinated, low-bandwidth sending

Collection

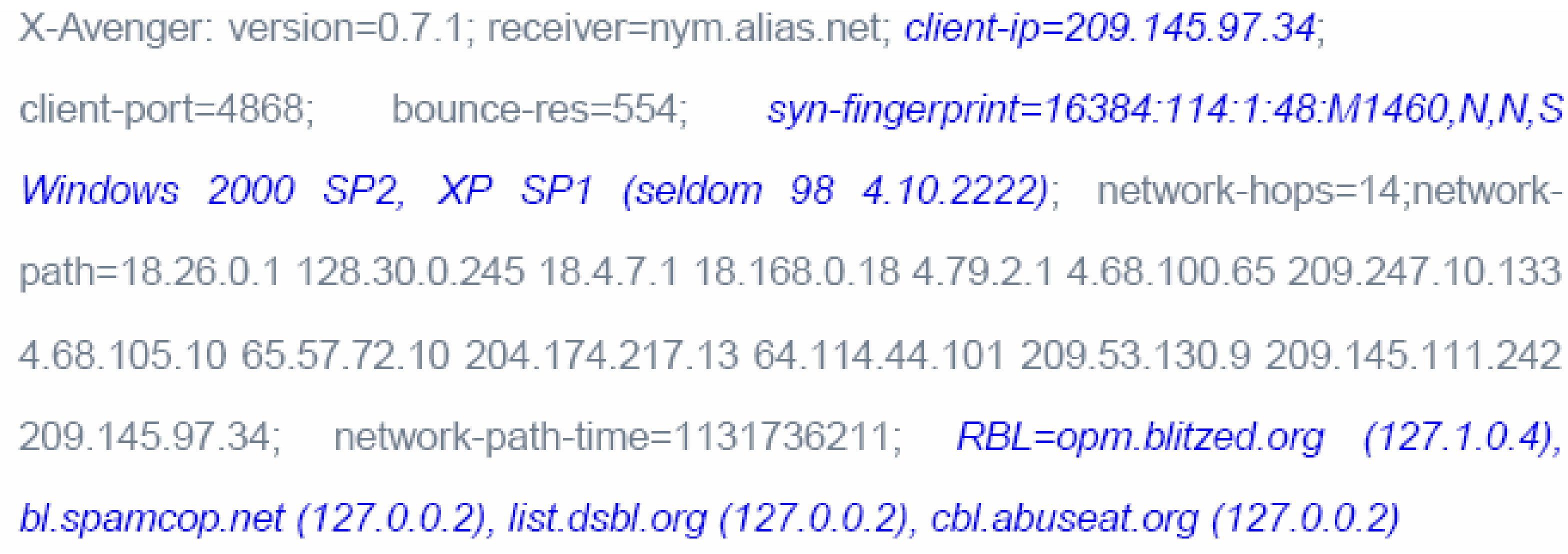
- Two domains instrumented with MailAvenger (both on same network)
  - Sinkhole domain #1
    - Continuous spam collection since Aug 2004
    - No real email addresses---sink everything
    - 10 million+ pieces of spam
  - Sinkhole domain #2
    - Recently registered domain (Nov 2005)
    - “Clean control” – domain posted at a few places
    - Not much spam yet…perhaps we are being too conservative
- Monitoring BGP route advertisements from same network
- Also capturing traceroutes, DNSBL results, passive TCP host fingerprinting *simultaneous with spam arrival*
  (results in this talk focus on BGP+spam only)

Data Collection Setup



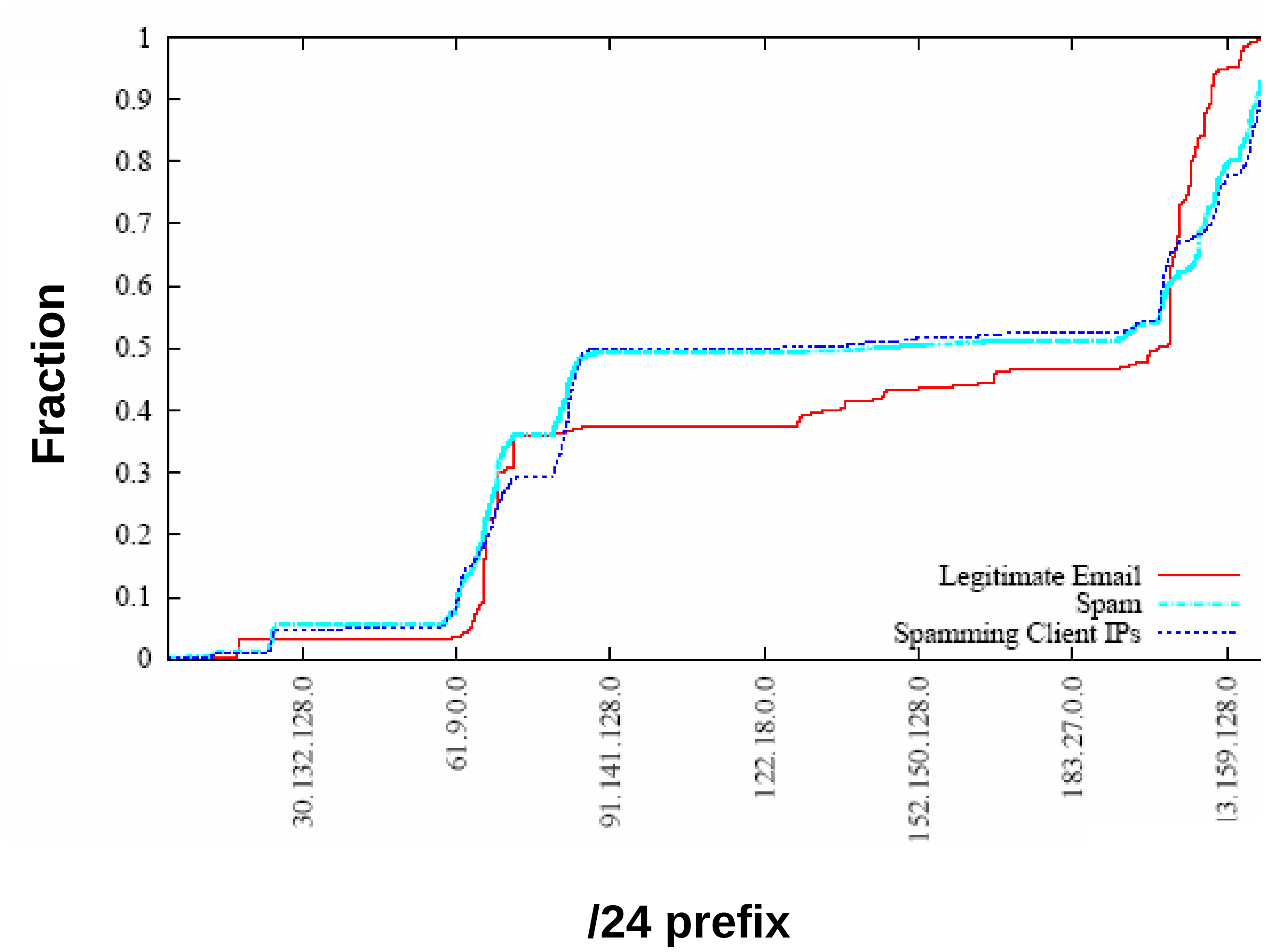
Mail Collection: MailAvenger

- Highly configurable SMTP server that collects many useful statistics



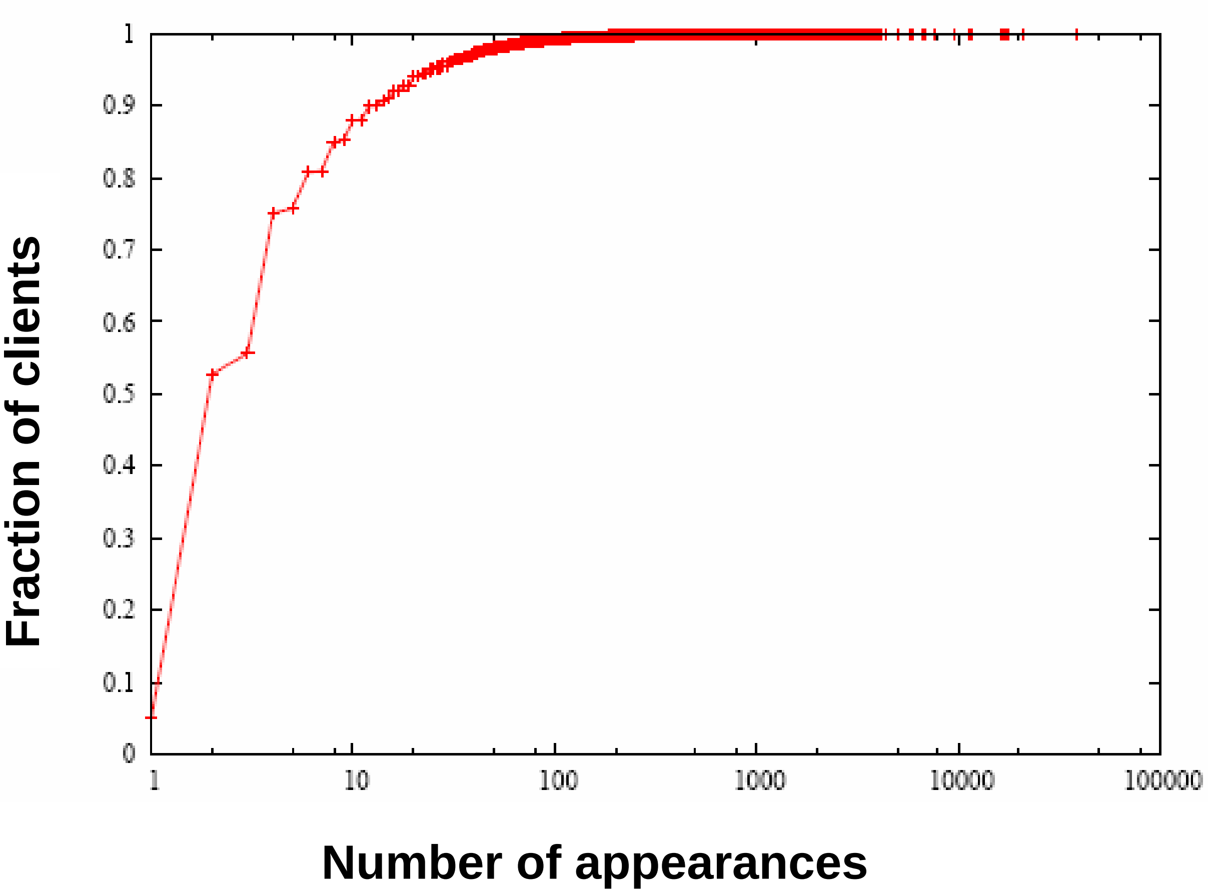
X-Avenger: version=0.7.1; receiver=nym.alias.net; *client-ip=209.145.97.34*; client-port=4868; bounce-res=554; *syn-fingerprint=16384:114:1:48:M1460,N,N,S Windows 2000 SP2, XP SP1 (seldom 98 4.10.2222)*; network-hops=14;network-path=18.26.0.1 128.30.0.245 18.4.7.1 18.168.0.18 4.79.2.1 4.68.100.65 209.247.10.133 4.68.105.10 65.57.72.10 204.174.217.13 64.114.44.101 209.53.130.9 209.145.111.242 209.145.97.34; network-path-time=1131736211; *RBL=opm.blitzed.org (127.1.0.4), bl.spamcop.net (127.0.0.2), list.dsbl.org (127.0.0.2), cbl.abuseat.org (127.0.0.2)*

Distribution across IP Space



Is IP-based Blacklisting Enough?

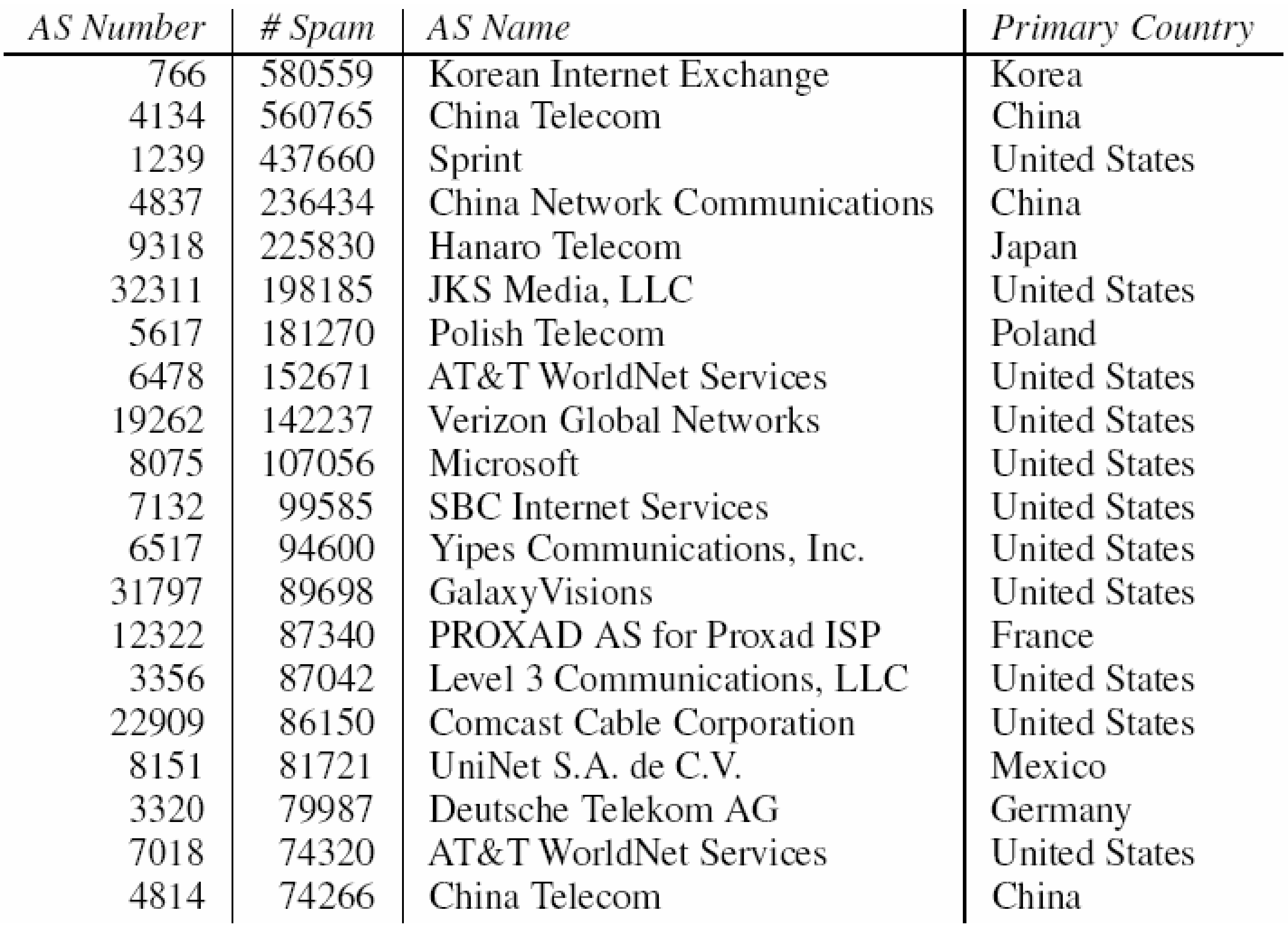
- Probably not: more than half of client IPs appear less than twice



Number of appearances

Distribution across ASes

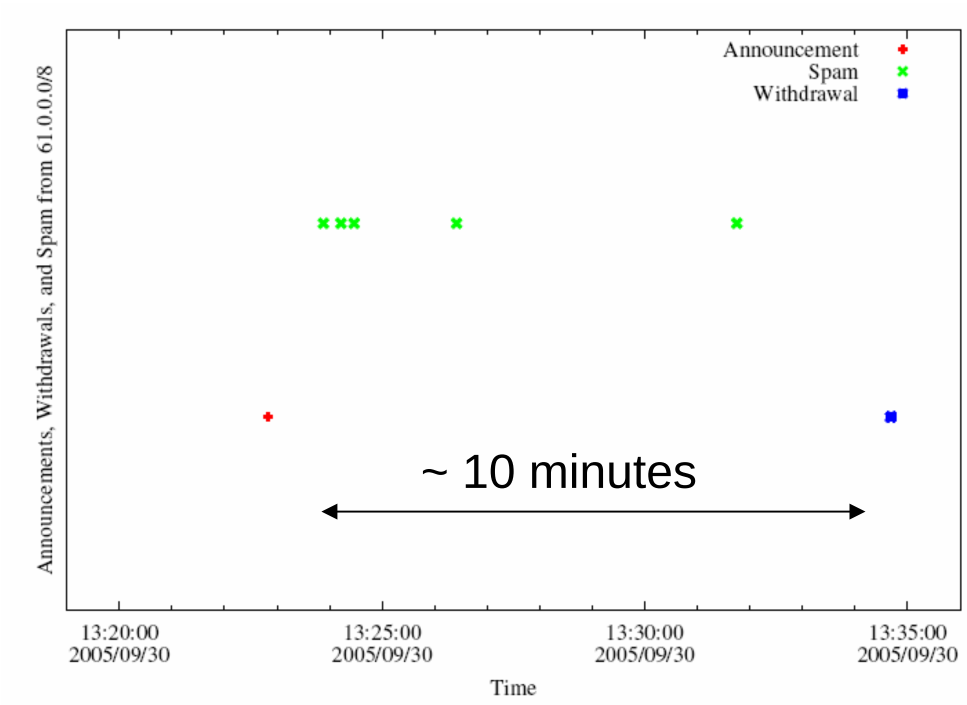
Still about 40% of spam coming from the U.S.



AS Number	*# Spam*	*AS Name*	*Primary Country*
766	580559	Korean Internet Exchange	Korea
4134	560765	China Telecom	China
1239	437660	Sprint	United States
4837	236434	China Network Communications	China
9318	225830	Hanaro Telecom	Japan
32311	198185	JKS Media, LLC	United States
5617	181270	Polish Telecom	Poland
6478	152671	AT&T WorldNet Services	United States
19262	142237	Verizon Global Networks	United States
8075	107056	Microsoft	United States
7132	99585	SBC Internet Services	United States
6517	94600	Yipes Communications, Inc.	United States
31797	89698	GalaxyVisions	United States
12322	87340	PROXAD AS for Proxad ISP	France
3356	87042	Level 3 Communications, LLC	United States
22909	86150	Comcast Cable Corporation	United States
8151	81721	UniNet S.A. de C.V.	Mexico
3320	79987	Deutsche Telekom AG	Germany
7018	74320	AT&T WorldNet Services	United States
4814	74266	China Telecom	China

BGP Spectrum Agility

- Log IP addresses of SMTP relays
- Join with BGP route advertisements seen at network where spam trap is co-located.



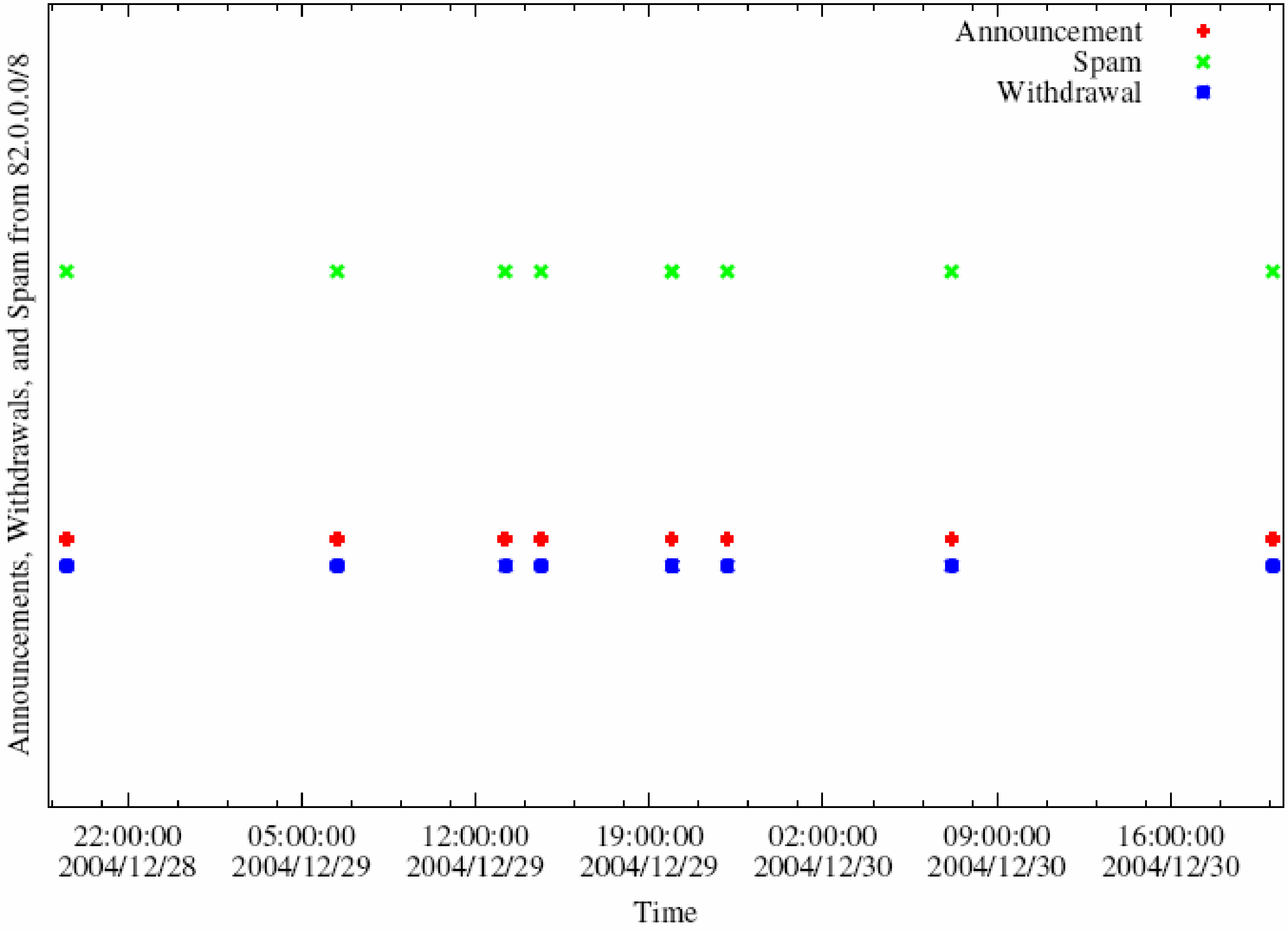
A small club of persistent players appears to be using this technique.

Common short-lived prefixes and ASes

61.0.0.0/8 4678
66.0.0.0/8 21562
82.0.0.0/8 8717

Somewhere between 1-10% of all spam (some clearly intentional, others might be flapping)

A Slightly Different Pattern



Why Such Big Prefixes?

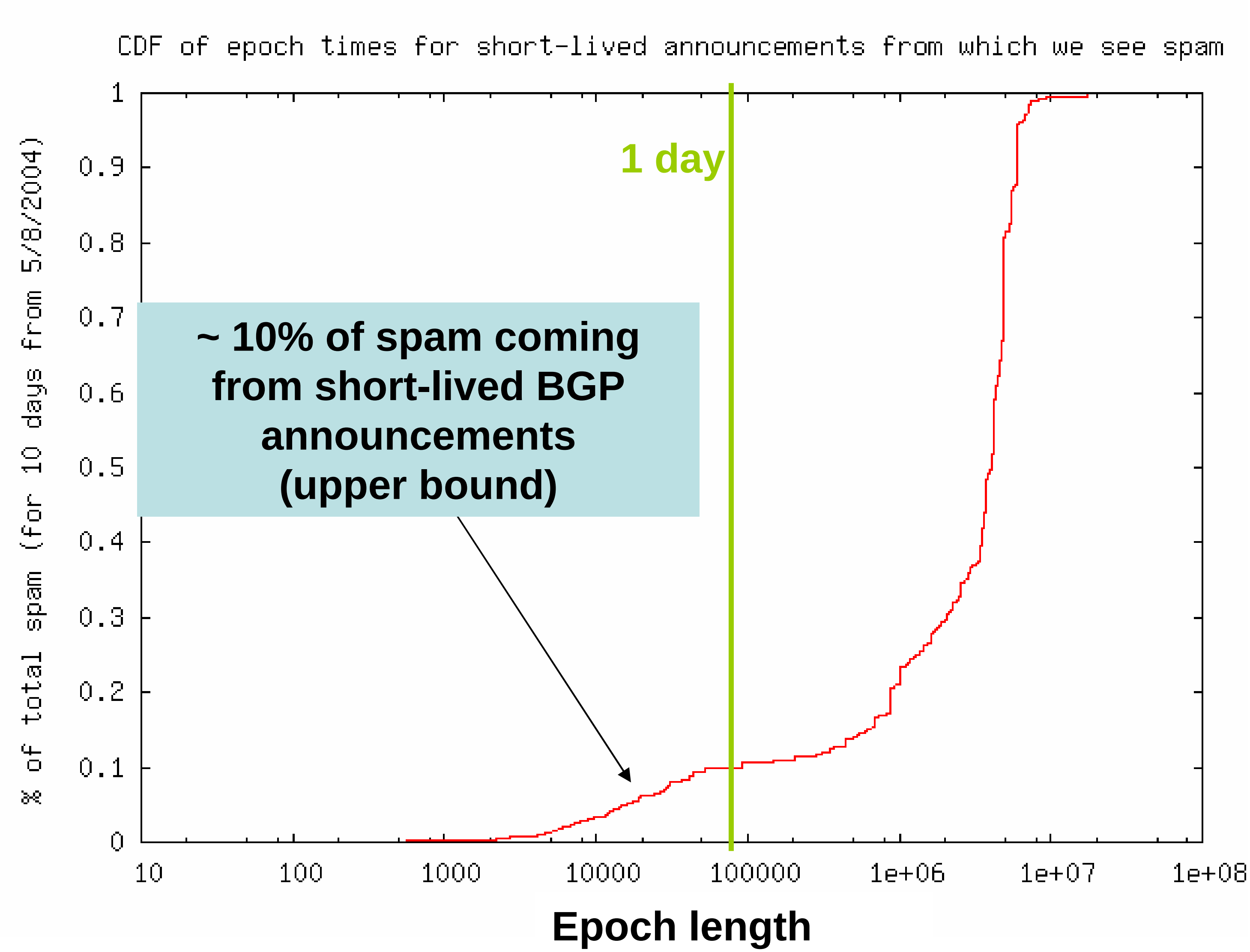
- **Flexibility:** Client IPs can be scattered throughout dark space within a large /8
  - Same sender usually returns with different IP addresses

- **Visibility:** Route typically won’t be filtered (nice and short)

Characteristics of IP-Agile Senders

- IP addresses are widely distributed across the /8 space
- IP addresses typically appear only once at our sinkhole
- Depending on which /8, 60-80% of these IP addresses were not reachable by traceroute when we spot-checked
- Some IP addresses were in *allocated*, albeit unannounced space
- Some AS paths associated with the routes contained reserved AS numbers

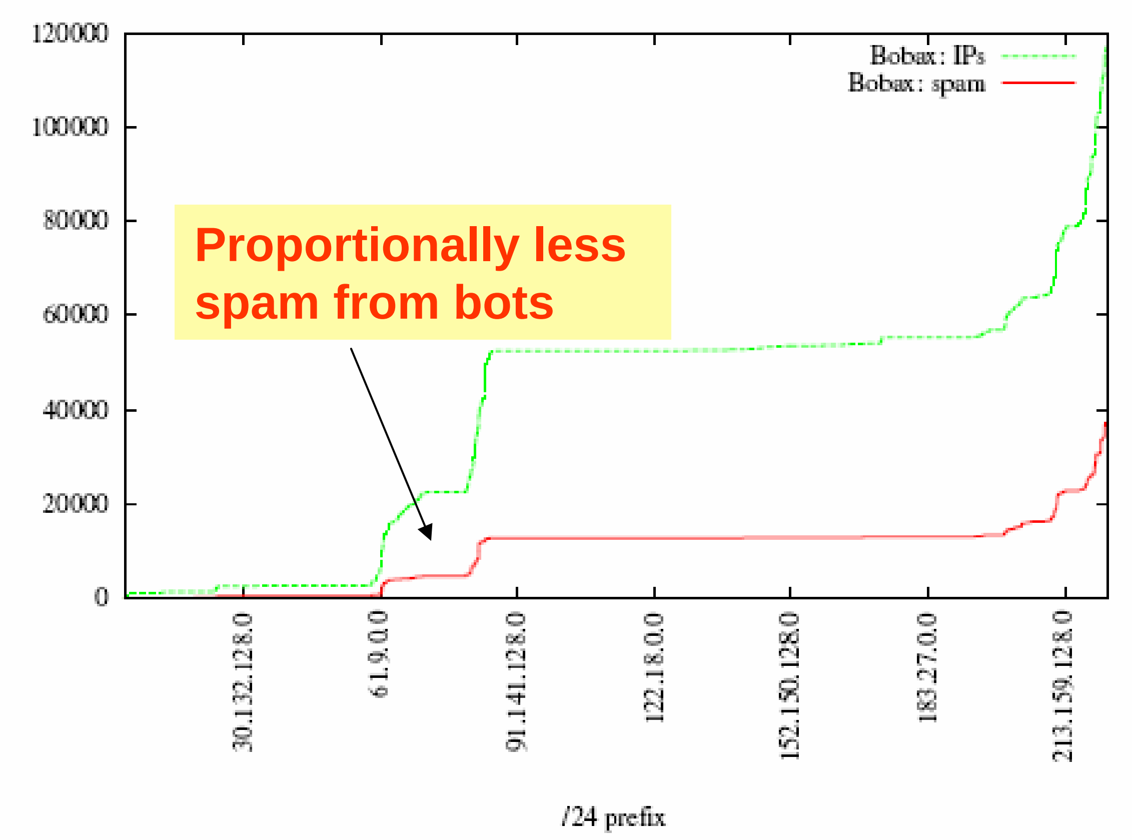
Length of short-lived BGP epochs



~ 10% of spam coming from short-lived BGP announcements (upper bound)

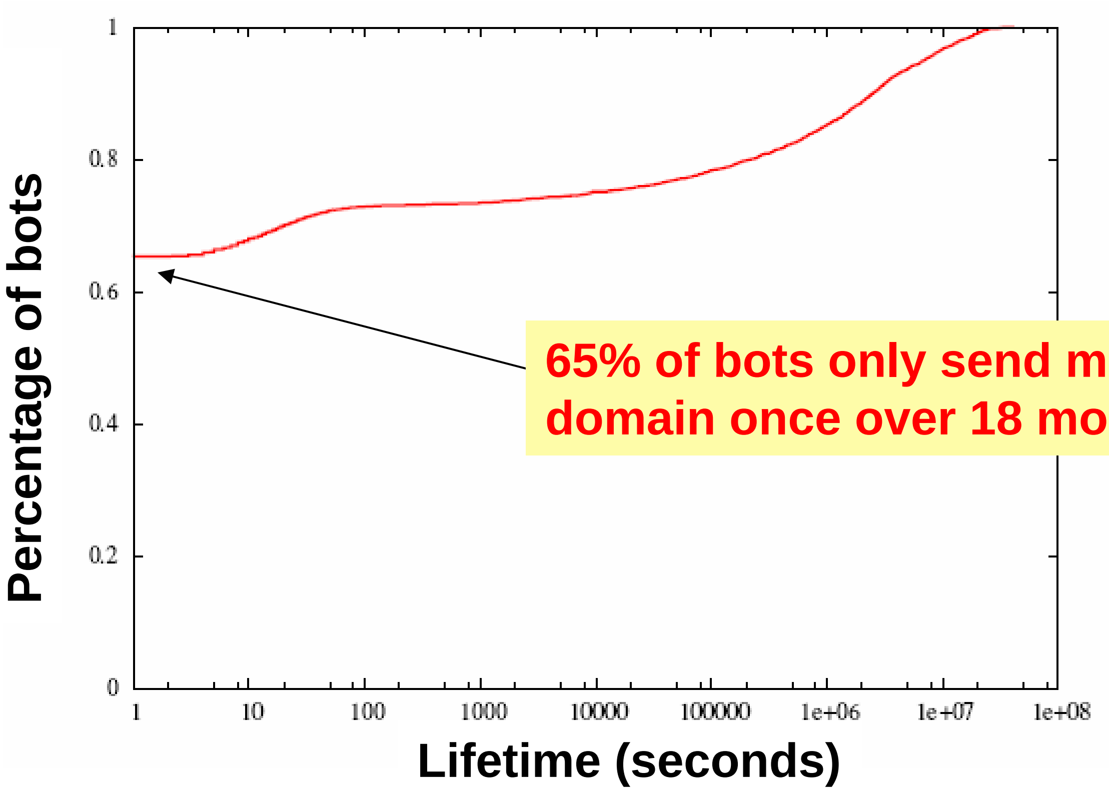
Spam From Botnets

- **Example:** Bobax
  - Approximate size: 100k bots



Proportionally less spam from bots

Most Bot IP addresses do not return

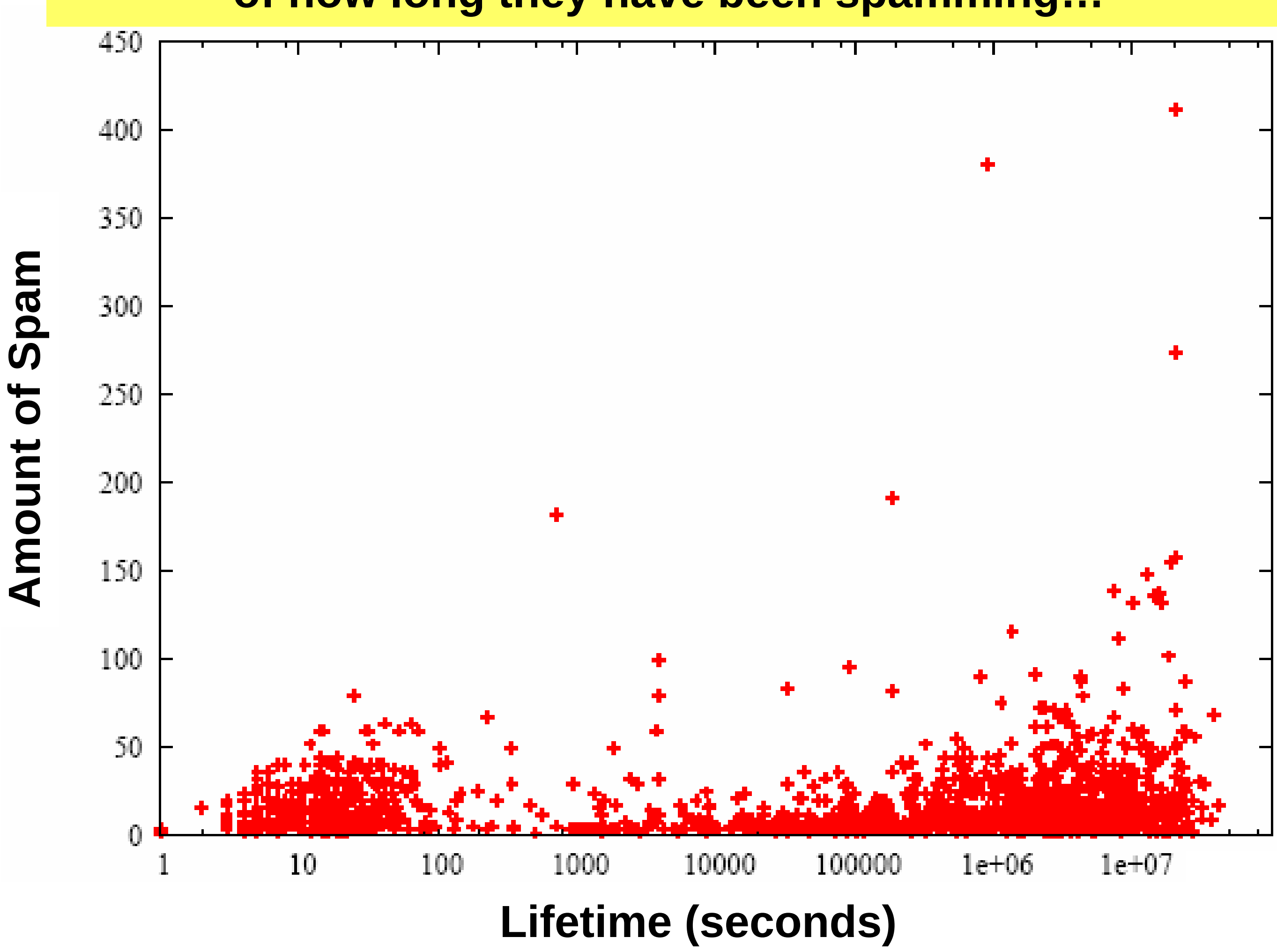


65% of bots only send mail to a domain once over 18 months

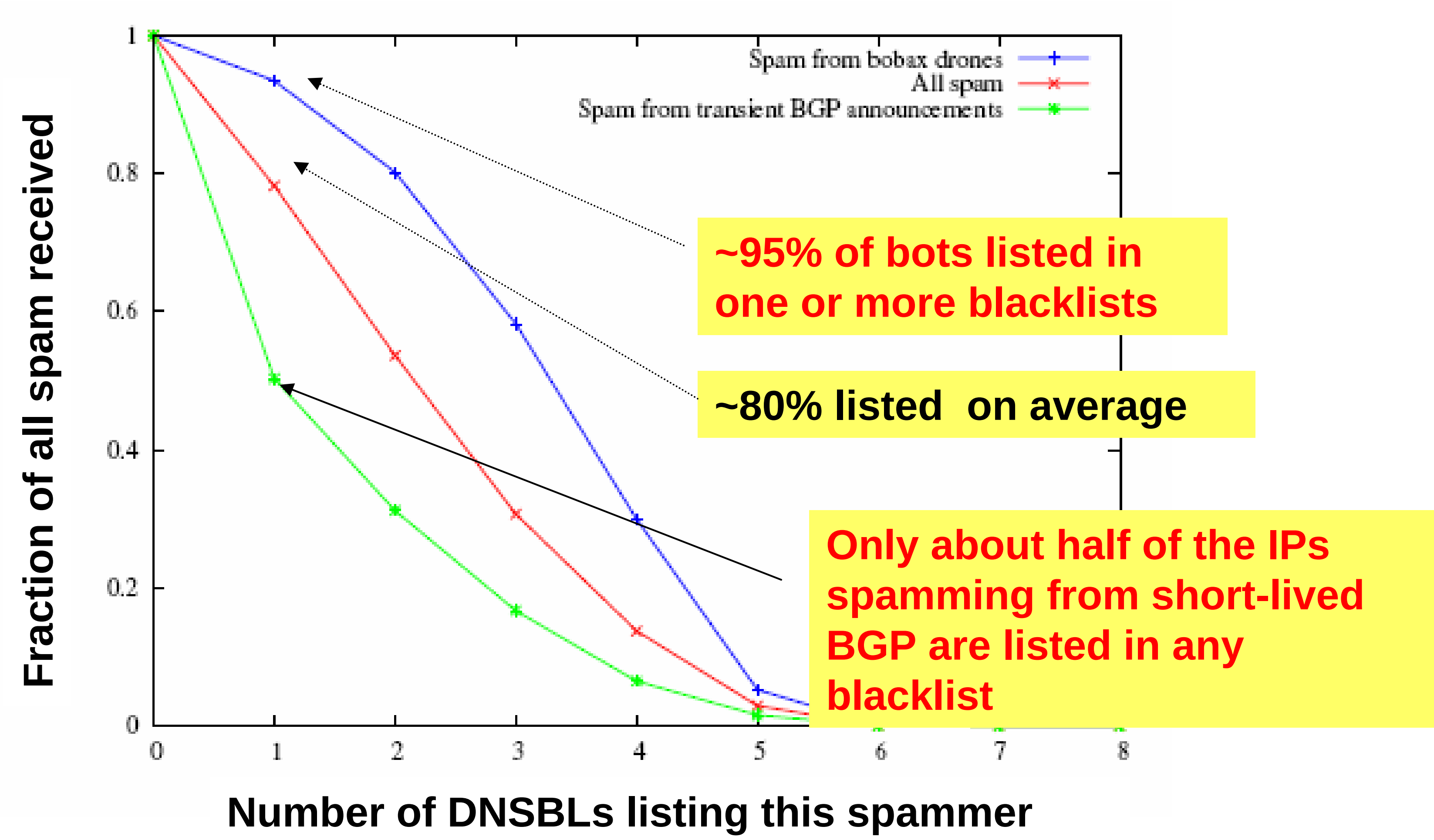
Collaborative spam filtering seems to be helping track bot IP addresses

Most Bots Send Low Volumes of Spam

Most bot IP addresses send very little spam, regardless of how long they have been spamming…



The Effectiveness of Blacklisting



~95% of bots listed in one or more blacklists

~80% listed on average

Only about half of the IPs spamming from short-lived BGP are listed in any blacklist

Spam from IP-agile senders tend to be listed in fewer blacklists

Harvesting

- Tracking Web-based harvesting
  - Register domain, set up MX record
  - Post, link to page with randomly generated email addresses
  - Log requests
  - Wait for spam
- Seed different subdomains in different ways

Preliminary Data: Example Phish

- A flood of email for a phishing attack for paypal.com
- All “To:” addresses harvested in a single crawl on January 16, 2006
- Emails received from two IP addresses, different from the machine that crawled
- Forged X-Mailer headers

Lessons for Better Spam Filters

- Effective spam filtering requires a better notion of end-host identity
- Distribution of spamming IP addresses is highly skewed
- Detection based on network-wide, *aggregate* behavior may be more fruitful than focusing on individual IPs
- Two critical pieces of the puzzle
  - Botnet detection
  - Securing the Internet’s routing infrastructure