Spam Resources

How can I help you?

William Stearns
wstearns@pobox.com
http://www.stearns.org
http://www.surbl.org

Sa-blacklist Overview

- Blacklist of spam hyperlink **domains**, not sender IP's
  - Follow the money
- Multiple volunteer contributors
  - Likely more than 20K person-hours from core
- Published in multiple formats
- Used in spam filters and MTA's
- Significantly helps identification

Statistics (3/16/2006)

- 193,728 blacklisted domains
  - Some are raw IP addresses when used in link
- 80,829 whitelisted domains
- ~40 contributors
- Updated hourly
- More than 1TB/month text file downloads
  - Hosted in Greece and the Netherlands, upcoming Germany
  - No easy way to identify amount of DNS lookup traffic (43 DNS servers)

Definition

- Unsolicited Bulk Email
  - “Commercial” misses some
  - CAN-SPAM's definition too narrow
- Includes phish
- Excludes bounces, Joe Jobs
- Not generally focused on viruses

Criteria

- Be conservative
- Generally avoid Joe Jobs, stock domains, 419's, foreign language
- No ISP domains, redirectors, Free hosting sites, “~user” accounts, dynamic DNS
- Whois data
  - Recent domains more likely spam
- Whitelist large corporations

14 Text File Formats

- Spamassassin multiple formats
- Sendmail, Postfix, Qmail
- Privoxy, Squid
- Bind, rbldnsd
- Raw domains
- Others on request

One of 6 DNS RBLs

- Published as DNS-based RBL ws.surbl.org
- Others: Spamcop, Outblaze, Abusebutler, Phishing, jwSpamSpy
- Usable by
  - sendmail, qmail, qmail-ldap, Exim, Exchange, Policy Patrol, MailScanner, Declude Junkmail, Merak Mail Server, MDaemon, NetIQ MailMarshal, ALOAHA, ASTPS, NEMX Power Tools, XWall, ORFilter, GEE Whiz, SpamBouncer, STAT AntiSpam, MTS Professional, Trash Finder, SpamPal, Secure Mail Suite

Raw spam data

- 2002-2006, multiple accounts
- 5.1GB anonymized spams
  - http:// ford.stearns.org/ spamevidence/{spamvertised_domain}.bin
  - Password provided on request
- 840,000 (5.8GB) unanonymized hand and machine checked spam
  - 3.95G is categorized

Categories

- adult, cable, casino, childporn, contactinfo, diploma, emaillist, faraway, miscspam, mixed, mortgage, nigeria, oem, phishing, pill, rolex, spyware, tobacco, toner, virus
- Stock spams, broken out in ~400 specific ticker symbols

Available Formats

- Anonymized (unanonymized with special request)
- Broken out into individual files or grouped into files according to:
  - hyperlinked domains in the body
  - month sent
  - last smtp relay
- Custom searches for specific text

Additional data

- ~220K whois records
- NS and A records for most
  - Worst offending class C and B stats
- ~220K Home pages
- ~78K raw home page downloads

Attachments

- 48K (573MB) categorized attachments
  - Largely images, some viruses, some others
- md5 and sha1 checksums at http://www.stearns.org/spamattach/
  - Also directly queriable at attachsum.com:
  - ping 00006e57b2bfdb4141e8402aa7b267b3
    - .md5sum.attachsum.com
      - If it returns 127.0.0.2 as an IP address, that checksum is a spam attachment

Automatic Spam Reporting

- Spamfeed - feed spam into bayesian learning system
  - My spam trains your spam filter
  - Currently Spamassassin, easily do others
- Checksum services
  - Razor, pyzor, DCC
- Categories fed to Agencies
  - FTC, Cable Assoc, FBI, Interpol, Treasury, Software Industry Assoc, APWG, FDA, USPIS, Rolex, SEC, ATF, Antivirus companies
- Custom feeds for ongoing investigations

Custom processing of your messages

- Extracting last smtp relay, spamvertised domains, images
- Breaking up messages by date, month, hyperlinked domains
- Anonymizing messages

Resources

- http://www.sa-blacklist.stearns.org /sa-blacklist/
- http://www.surbl.org/
- http://www.stearns.org/spamattach/
- http://www.stearns.org/spamfeed/
- http://ford.stearns.org/spamevidence/
- wstearns@pobox.com
- http://www.ists.dartmouth.edu