Hit the Ground Spam(fight)ing

v2.0

LISA '06, Washington, D. C.

December, 2006

John "Rowan" Littell

rowan (at) hovenweep (dot) org

Know your users

- Who are they?

  Are they liberal arts professors? Professional geeks? Your family? Stock traders?

- How do they access mail?

  IMAP, POP, web mail? Do their clients work best with quarantine folders, server-side filtering, or client-side filtering?

- How much support can you give?

  Do you have time to debug their procmail scripts?

Where to fight spam

- Mail client (MUA)
  - Pros: It's there and ready to use, can work with enough attention.
  - Cons: Non-technical users may consider it too much hassle, no domain-wide benefits
  - Examples: Thunderbird, Mail.app
- Mail server or mail exchanger (MTA)
  - Pros: Domain-wide benefits, can be quite effective, allows users to "just get their mail."
  - Cons: More work for you **OR** very expensive.
  - Examples: SpamAssassin, dspam, Iron Port, Postini, etc., etc., etc.

Protocol Hacks: DNSBL

- Reject mail from IP addresses presumed to be spammers via DNS lookup
  - Pros: Quick, widely supported
  - Cons: Quality varies, tends toward either false positive or false negative
  - Suggestions: Choose a well-respected one, have a method in place for exceptions, run a dedicated caching name server
  - Reference: http://en.wikipedia.org/wiki/DNSBL

Protocol Hacks: Greylist

- Tempfail the first instance of sender/recipient/IP address triplet, accept when it tries back
  - Pros: Entirely within SMTP, effective against virii
  - Cons: Delays the first message, some server architectures don't play well, the spammers are getting smarter
  - Suggestions: Choose a flexible one, use the well-known whitelist, have a method for exceptions, check against /24 address space instead of /32.
  - Reference: http://en.wikipedia.org/wiki/Greylist

Protocol Hacks: SMTP and TCP Tricks

- Require senders to follow RFCs and basic good behavior
  - Possible Methods: HELO before data, HELO string checking, Sendmail "greet pause", DNS sanity checking, throttling connections, feedback from MTA into firewall
  - Pros: Catches a number of spamware systems
  - Cons: Catches a few legitimate mail server implementations, some methods need maintenance, some methods are only implemented as hacks (milters, etc.).
  - Suggestions: Watch for exceptions, don't use high-maintenance "tricks"

Content Analysis: SpamAssassin

- General clearing house for all kinds of tricks: content matching, DNSBL, fuzzy checksums, auto-whitelisting, image analysis, Bayesian analysis. . .
  - Pros: Strong community support, fairly effective, plugins add accuracy
  - Cons: Requires constant updates, processor intensive
  - Suggestions: Update frequently, look for good plugins, don't waste time writing your own rules (unless you want to)

Content Analysis: Bayesian Classification + Learning

- Calculate probability of spam content based on learned spam words and tokens
  - Pros: Over time can become very accurate, requires little maintenance
  - Cons: Diverse mail content can lower accuracy
  - Suggestions: Allow users to build individual Bayes databases for individual accuracy, combine with site-wide database for shared known spam
  - Reference: http://en.wikipedia.org/wiki/Naive_Bayes_classifier

Content Analysis: Antivirus

- Identify known e-mail viruses and executable content
  - Pros: AV engines are very accurate for viruses, some include phishing matching
  - Cons: Takes resources
  - Suggestions: Dump or quarantine positive matches, **do not** send sender notifications – **this is spam!**

Performance Tuning

- Tune your OS: for network, memory and processor
- Learn your MTA: timeouts, threading, queue structure
- Use your database wisely: cache and share connections, prepare statements
- Cache everything: DNS lookups, user preferences, results of simple checks
- Ananlyze your system: log everything and run log analysis, generate graphs and reports... but don't chase every detail