

Spam, Spam, Spam

Mr. Stephen Ulmer <ulmer@ufl.edu>

Peer2Peer VIII
University of Florida
March 30, 2005

Gainesville, FL

What is Spam?

- Spam is a colloquialism used to refer to Unsolicited Commercial Email (UCE)
- Other monikers include Unsolicited Bulk Email (UBE), and simply "junk mail"
- Named after a sketch on Monty Python's Flying Circus ("Spam")
- Originated by a couple of lawyers who wanted to "Make Money Fast!" on Usenet

Variations on a Theme

Different types of unsolicited messages are emerging as other technologies become prevalent:

- SPam on Instant Messaging (SPIM)
- Unsolicited Short Messages sent to your mobile phone (SMS Spam)
- Automated commercial replied to postings in discussion forums, specifically, those implemented with *Movable Type*.

Cost of Employee Effort (Conservative)

- An IDC study concluded that the average corporation with 5,000 loses \$4.1M/year of productive work time to spam
- UF has about 15,000 employees
- We may be losing \$12.3M/year in productivity to dealing with spam

Cost of Employee Effort (Radical)

- Another study indicated that employees spend about 30 minutes/day dealing with spam
- At UF, this would be about \$49.6M/year
- It's hard to get a good accounting of "time lost", but whatever estimate is used the costs are staggering

Non-Tangible Risks

- Accidental erasure of legitimate messages
- Disclosure of personal or financial information to phishers
- Change in personal attitudes towards email
- Email delivery to outside entities becomes less certain

Spam Detection Technologies

Spam Detection Technologies

Bayesian Statistical Analysis

- Algorithm for inductive statistical analysis
- The frequency of words or strings in combination is used to calculate the probability that this message is spam
- Requires a large corpus of example ham and spam
- Non-statistical methods can be used to identify ham and spam, thus creating a self-teaching system

Real-Time Black Lists

- Collection of hostnames and IP addresses thought to produce spam
- Queries are posed with the Domain Name Service (DNS) protocol
- There are many RBL services: MAPS, ORBS, SORBS, Spamhaus, NJBL, DS-BL,
- Our favorite RBL is Spamhaus

Content Based Scoring

- Specific *features* thought to be unique to spam messages are identified. Some examples:
 - No *Date:* header
 - No real name in the *From:* header
 - Message contains "Click Here!"
- Each of these *features* is assigned a score
- A test is performed to detect each *feature*
- No single test should cause a message to be considered spam

End-User Feedback

- Users forward messages that they think are spam to a central authority
- The reported messages are used to feed various other identification systems
- Sometimes this feedback-loop is completely automated
- At UF:
 - report-ham@ufl.edu
 - report-spam@ufl.edu

Central-Site Measures

Central-Site Measures

13 / 27

So What's Central?

- Implemented on `smtp.ufl.edu`, which handles incoming and outgoing Gatorlink email
- Affects these `centralufl.edu`, `nersp.nerdc.ufl.edu`, `cns.ufl.edu`, `lists.ufl.edu`,
- Also affects mail sent to these domains that is forwarded to other systems

Central-Site Measures

14 / 27

Scoring System for Gatorlink Mail

- Implemented with SpamAssassin open source software
- Bayesian analysis contributes between -4.9 and 4.9 to the total score
- Currently, messages with a score of 10 or more are rejected

Central-Site Measures

15 / 27

RBL Employment

- The Spamhaus RBL is employed to reject messages from *known* or *admitted* spam sites
- The rejection is accomplished by introducing a permanent error into the conversation between mail servers
- The remote mail server is still responsible for the message; it is required to notify the sender

Central-Site Measures

16 / 27

Additional Local Anti-Spam Measures

- Messages with known, detectable viruses are rejected
- Messages having attachments with filenames most email clients open automatically are rejected
- Messages of type "message/partial" are rejected because they can't be checked for the above

Additional Message Headers

- Several headers are added to each message:
 - *X-Scanned-By*: contains a URL for more information

```
X-Scanned-By: CNS Open Systems Group \  
(http://open-systems.ufl.edu/services/smtp-relay/)
```
 - *X-Spam-Level*: is a graphical representation of the score.

```
X-Spam-Level: *****
```
 - *X-Spam-Status*: is a report containing the tests that this message "failed"

```
X-Spam-Status: hits=8.472, required=5, \  
tests=BAYES_90, FORGED_YAHOO_RECVD, FREE_CONSULTATION, FROM_HAS_MIXED_N
```

UF's Spam Volume

- smtp.ufl.edu processes about 1.2M messages per day
- 50% of that gets delivered
- 15% of the mail presented is for addresses that don't exist
- 20% of the messages are rejected for their spam score (greater than 10)
- 2.5% of the messages are rejected because of features indicating that they are unsafe
- 1.5% of the messages are rejected as viruses
- 11% of the messages experience some other kind of error (usually the remote host is unavailable)

Outside Unit Participation

Outside Unit Participation

Domain Level Participation

Units outside of CNS can consume UF' sspam scoring service for their own domain.

1. Contact the CNS Open Systems Group. Include the domain(s) you want to participate and the DNS name of the server to which the mail should be delivered.
2. OSG will implement private mail routing for those domains and return the contact.
3. Change the MX records for those domains in the following way:
 - Remove all but one MX record
 - Assign that MX record to `smtp.uf1.edu`
 - Reject SMTP connections from everyone but the local users and `smtp.uf1.edu`
 - Configure the departmental mail server to use `smtp.uf1.edu` as the *smart host*

Instructions are available at <http://open-systems.ufl.edu/service/smtp-relay/>

End-User Participation

End-users can participate in the system individually:

1. Advertise your Gatorlink email address (`username@ufl.edu`).
2. Either use your Gatorlink INBOX, or forward that address to your departmental mail server.
3. Use `smtp.uf1.edu` as your *outgoing* mail server.

Instructions for configuring most popular clients are available at <http://www.cns.ufl.edu/spam/>

University Relations

University Relations

Another Brick In the Wall

- Outside entities sometimes refuse mail from UF mail servers
- When this happens, end-users must attempt to resend the affected message at a later time
- The actual reason for the failure is often not obvious to inexperienced users

The AOL Problem

- There are primarily two "blocks" that affect UF' s mail flow
- One is related to the number of messages per unit time that a particular server sends to AOL
- The other is based directly on the number of AOL customer complaints
- No distinction is made between messages originating at the University and messages forwarded for a third-party
- The AOL interface does not adequately indicate the gravity of marking a message as "Junk"

Current Solutions (Policy)

- *Students* will no longer be able to forward their Gatorlink mail to addresses outside of the `ufl.edu` domain
- Will ensure that students receive official email communications from the University
- Will be completely implemented on October 5, 2005

Additional Information

- Instructions for configuring several email clients are available at:
<http://www.cns.ufl.edu/spam/>
- If your department replaces *X-Spam-Level:*, you can use *X-UFL-Spam-Level:* instead
- More information about the mail servers is available at:
<http://open-systems.ufl.edu/services/smtp-relay/>