SWIN BUR NE

CENTRE FOR ADVANCED INTERNET ARCHITECTURES

End-users' resource consumption of spam and a 3D anti-spam evaluation framework

Minh Tran
Grenville Armitage

Talk outline

- Introduction
- Quantifying end-user cost of spam
  - Overview of methodology
  - Comparison of three basic email retrieving models
  - Comparison of ten email retrieving mechanisms
  - Financial cost for residential and business uses
- Three-dimensional evaluation framework
  - An overview of anti-spam methods
  - A cost-based three-dimensional evaluation framework
- Conclusion

Introduction

- There is a weak binding between service consumption and the cost to use. This happens in the current Internet email system -> Spam
- Spam
  - "Mass unsolicited electronic mail"
    (RFC 2505, "Anti-Spam Recommendations for SMTP MTAs ")
- This paper:
  - Quantify the end-user cost of spam
  - Introduce our comparison model with three cost-based factors

Talk outline

- Introduction
- Quantifying end-user cost of spam
  - Overview of methodology
  - Compare three basic email retrieving models
  - Compare ten email retrieving mechanisms
  - Financial cost for residential and business uses
- Three-dimensional evaluation framework
  - An overview of anti-spam methods
  - A cost-based three-dimensional evaluation framework
- Conclusion

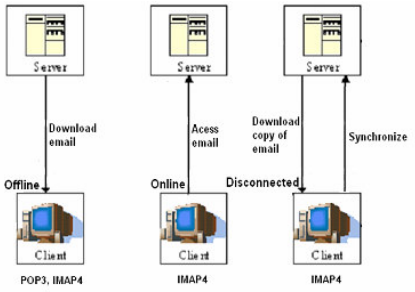
Overview of methodology

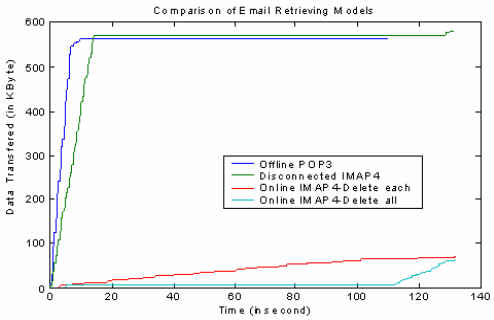
- Quantifying the network consumption cost of spam at end-users using different common email retrieval mechanisms
  - Retrieved and deleted 100 spam emails for each trial and repeated 100 trials for every email retrieving method
  - Asymmetric link of 1.5Mbps downstream and 256Kbps upstream
  - Tcpdump and Ethereal to capture the traffic and provide traffic statistics
  - Matlab 6.5 is used with our own software to plot data
- Sample spam emails
  - 6955 spam emails over a 24 hour period in early 2004
  - Sizes range from 1 KByte to 11 Kbytes, mean of 4.6 KBytes

Talk outline

- Introduction
- Quantifying end-user cost of spam
  - Overview of methodology
  - Compare three basic email retrieving models
  - Compare ten email retrieving mechanisms
  - Financial cost for residential and business uses
- Three-dimensional evaluation framework
  - An overview of anti-spam methods
  - A cost-based three-dimensional evaluation framework
- Conclusion

Basic POP3 and IMAP4 retrieval models





- Offline: Download email and process locally
- Online: Access email stored at mail server
- Disconnected: Download a copy of email. Synchronize with server

- Online model accounts for only 10% of network traffic consumed by Offline or Disconnected model
- Disconnected model takes longer than Offline model to download and delete emails

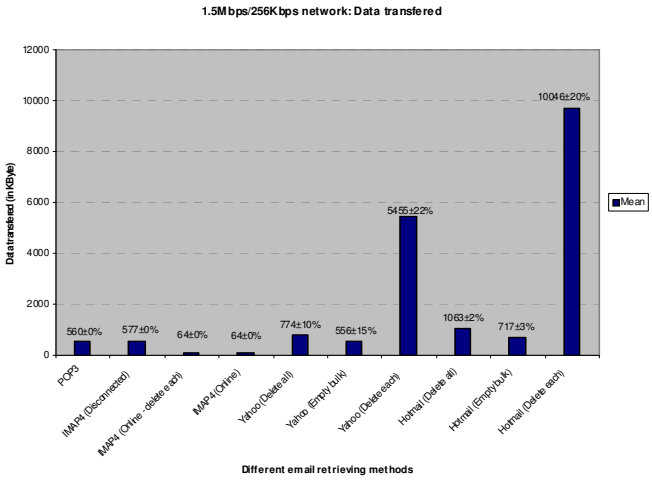
Talk outline

- Introduction
- Quantifying end-user cost of spam
  - Overview of methodology
  - Compare three basic email retrieving models
  - Compare ten email retrieving mechanisms
  - Financial cost for residential and business uses
- Three-dimensional evaluation framework
  - An overview of anti-spam methods
  - A cost-based three-dimensional evaluation framework
- Conclusion

Compare ten email retrieving mechanisms

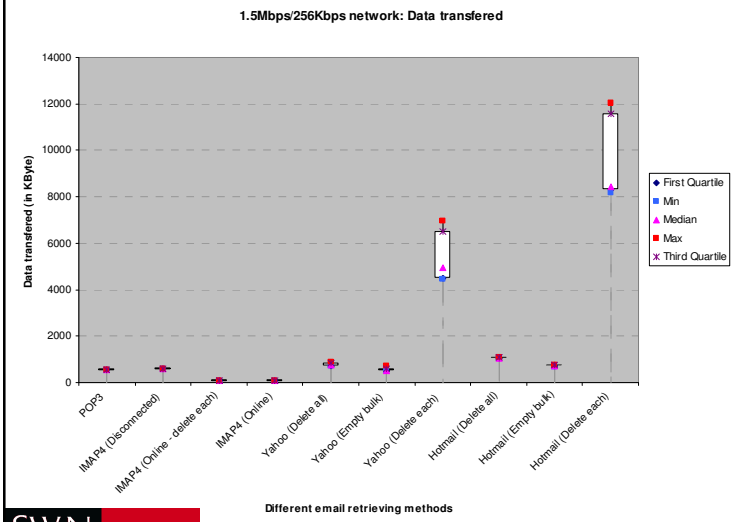
- We implemented ten methods of accessing emails to represent end-users' behaviors
  - Previous POP3 and IMAP4 scenarios
    - POP3, IMAP(O), IMAP(D)
  - Log into the Yahoo or Hotmail mailbox and empty spam email bulk folder
    - Yahoo(EB)
    - Hotmail(EB)
  - Log into the Yahoo mailbox, read all spam email headers on one page and delete all of them at the same time
    - Yahoo(DA)
    - Hotmail(DA)
  - Log into the Yahoo mailbox, read each spam email header and delete each of them
    - Yahoo(DE)
    - Hotmail(DE)

Compare ten email retrieving mechanisms



- Hotmail-delete-each causes the most data transferred, 2 times the second highest (Yahoo-delete-each), 20 times POP3/Disconnected-IMAP4 and 200 times Online-IMAP4
- Yahoo has a higher relative standard deviation (in both time and data transferred) than Hotmail

Compare ten email retrieving mechanisms

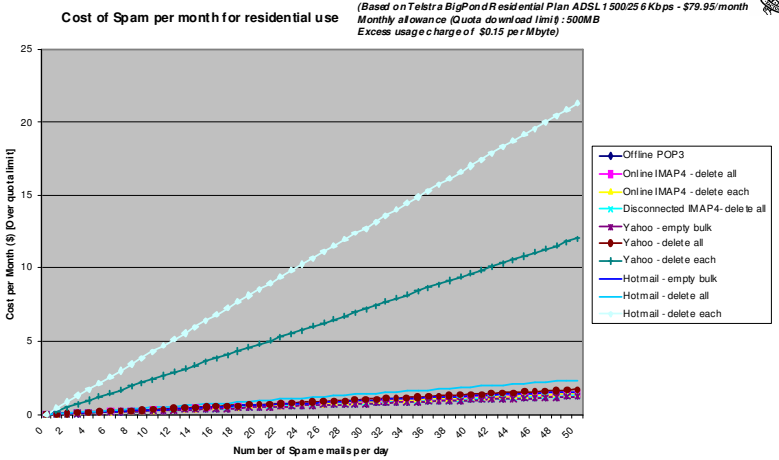


Graphs of Yahoo-delete-each and Hotmail-delete-each method suggest there might be large web objects and advertising banners

Talk outline

- Introduction
- Quantifying end-user cost of spam
  - Overview of methodology
  - Compare three basic email retrieving models
  - Compare ten email retrieving mechanisms
  - Financial cost for residential and business uses
- Three-dimensional evaluation framework
  - An overview of anti-spam methods
  - A cost-based three-dimensional evaluation framework
- Conclusion

Financial cost: a residential scenario



When over quota limit, spam can be quite costly for people using Yahoo(DE) or Hotmail(DE) as their retrieval method

Financial cost: business use

Cost of Spam per month for medium/large business*

Based on Optus Business Plan ADSL 1500/256 Kbps
$160/month
Monthly allowance (Quota download limit): 9GB
Excess usage charge of $0.07 per MByte

Different methods of retrieving and deleting spam emails	Within quota limit — Data consumed toward quota limit		Over quota limit
	(to nearest Mbyte)	(in percentage of quota limit)	Cost per month (to nearest dollar)
Offline POP3	82	1%	6
Online IMAP4 - delete all	9	2%	1
Online IMAP4 - delete each	9	2%	1
Disconnected IMAP4- delete all	85	17%	6
Yahoo - empty bulk	81	16%	6
Yahoo - delete all	113	23%	8
Yahoo - delete each	799	160%	56
Hotmail - empty bulk	105	21%	7
Hotmail - delete all	156	31%	11
Hotmail - delete each	1420	284%	99

- POP3 or IMAP4 – if the corporate mail server resides on other side of the ADSL connection
- Yahoo or Hotmail - when employees use personal emails at work

** Based on assumption of 10 emails per day per employee and 50 employees per business*

Talk outline

- Introduction
- Quantifying end-user cost of spam
  - Overview of methodology
  - Compare three basic email retrieving models
  - Compare ten email retrieving mechanisms
  - Financial cost for residential and business uses
- Three-dimensional evaluation framework
  - An overview of anti-spam methods
  - A cost-based three-dimensional evaluation framework
- Conclusion

Current anti-spam methods (1/2)

- Black-listing and white-listing
  - DNS Black-lists: SpamHaus,Open Relay Database (ORDB)
  - Challenge-Response
- Content Filtering
  - Text characterisation
  - Statistical methods
    - Machine learning algorithms, training stage, classification stage
    - Naive Bayes, k-Nearest Neigbour (k-NN) and Neural Networks
- Economics-based schemes
  - Bankable postage, e-postage, Microsoft's 'stamp of approval'

Current anti-spam methods (2/2)

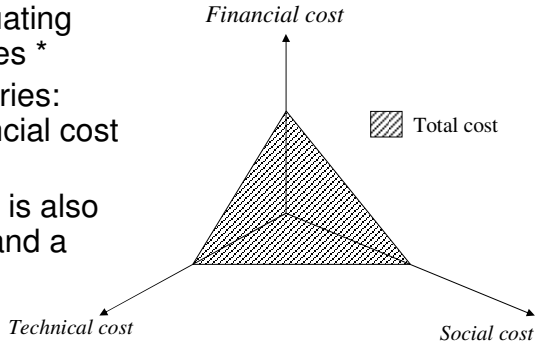
- Rate-limiting
  - Anti-Spam Router (ASR) of TurnTide, our own work MT Proxy
- Authentication against spam forgery and phishing
  - Sender ID Framework' of Microsoft and Pobox, DomainKeys Identified Mail (DKIM) of Cisco's and Yahoo
- Other techniques
  - 'Collaborative Filtering', 'Disposable Email Address' with Rolling Email Address Protocol (REAP)

Talk outline

- Introduction
- Quantifying end-user cost of spam
  - Overview of methodology
  - Compare three basic email retrieving models
  - Compare ten email retrieving mechanisms
  - Financial cost for residential and business uses
- Three-dimensional evaluation framework
  - An overview of anti-spam methods
  - A cost-based three-dimensional evaluation framework
- Conclusion

A three-dimensional valuation framework

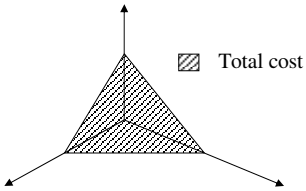
- Propose a visual 3D framework for evaluating anti-spam techniques *
- Three 'cost' categories: technical cost, financial cost and social cost
- Each measurement is also assigned a weight and a numerical score



* Inspired by our colleges' work: *T.T.T. Nguyen, G. Armitage, "Evaluating Internet Pricing Schemes - A Three Dimensional Visual Model", ETRI Journal, vol.27, no.1, Feb. 2005*

A three-dimensional valuation framework

Social Cost (Z)				
Authentication And identity (spam phishing (Z1)	Commercial Spam (Z2)	Spam in combined with virus/worms/trojans (Z3)	Spam in combined with eDOS, DHA (Z4)	Porn and other social spam (Z5)
Numerical Score (0-10) Or NA	Numerical Score (0-10) Or NA	Numerical Score (0-10) Or NA	Numerical Score (0-10) Or NA	Numerical Score (0-10) Or NA



Technical Cost (X)			
Types of spam trick (X1)	Architecture viability/ (X2)	Monitoring & updating (X3)	Maintaining & updating (X4)
Numerical Score (0-10) Or NA	Numerical Score (0-10) Or NA	Numerical Score (0-10) Or NA	Numerical Score (0-10) Or NA

Financial Cost (Y)			
Deployment/license (hardware & Software Cost) (Y1)	Spam Reduction (Y2)	False Positive (Y3)	False Negative (Y4)
Numerical Score (0-10) Or NA	Numerical Score (0-10) Or NA	Numerical Score (0-10) Or NA	Numerical Score (0-10) Or NA

Talk outline

- Introduction
- Quantifying end-user cost of spam
  - Overview of methodology
  - Compare three basic email retrieving models
  - Compare ten email retrieving mechanisms
  - Financial cost for residential and business uses
- Three-dimensional evaluation framework
  - An overview of anti-spam methods
  - A cost-based three-dimensional evaluation framework
- Conclusion

Conclusion

- Quantitative minimum network consumption of spam imposed on end-users
- Three orthogonal/independent cost factors and several cost criteria in our evaluation framework
- Future work:
  - Further investigate the relationship between service consumption and service control techniques across different areas of the Internet
  - Denial-of-Service (DOS) attacks
  - User motivation and malicious nodes in Adhoc networks