Albert-Ludwigs-
Universität Freiburg

Rechnernetze und Telematik
Seminar Rechnernetze

e-mail

Peter Rabiański

Betreuer: Prof.Dr. Christian Schindelhauer

Electronic Mail

- MUA - one of the most available application service
- Must provide, when remote destination temporary unreachable
- Uses independent addresses

```
local-part @ domain-name
```

Email Client

- Document Editor
- Address Book
- Permanent Storage
- Communications Module

Electronic Mail System

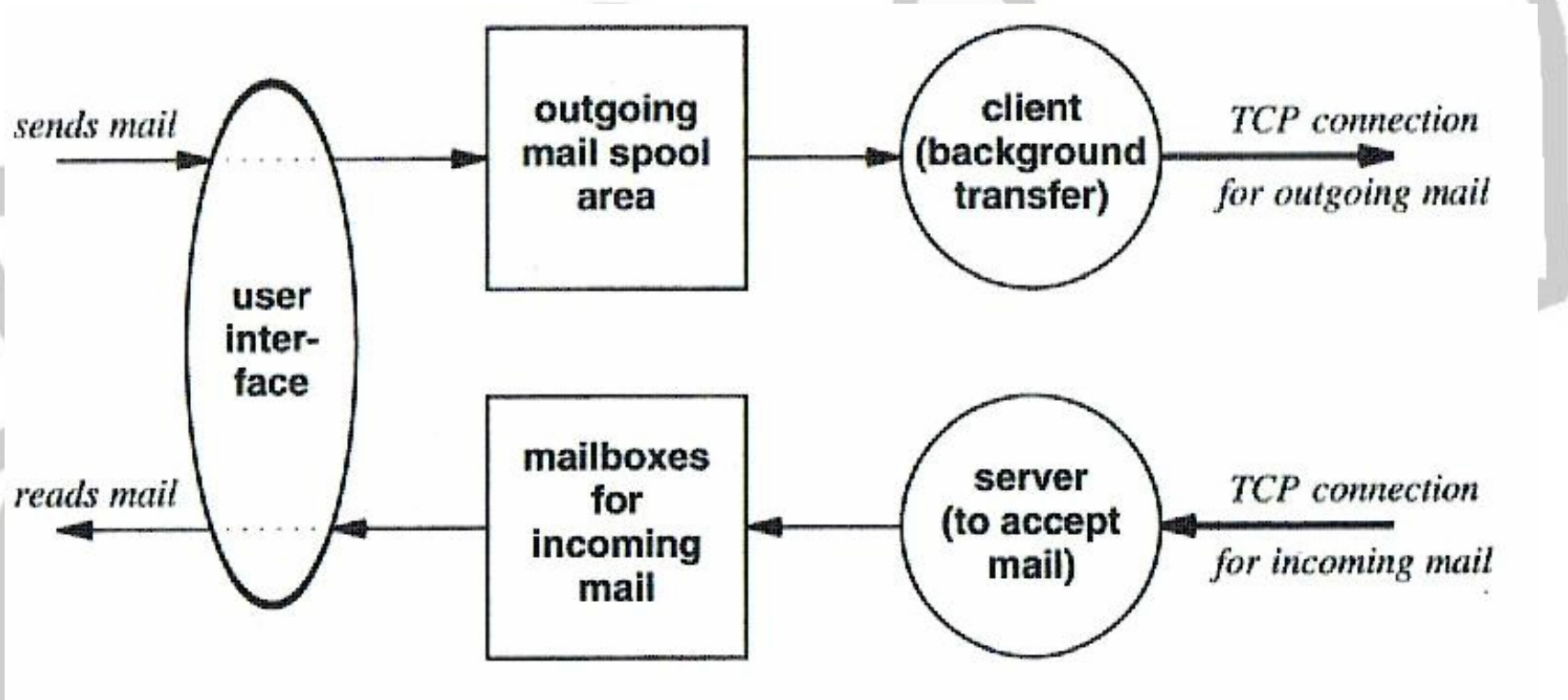


Figure 26.1 Conceptual components of an electronic mail system, Corner

EMS with Mail Forwarding

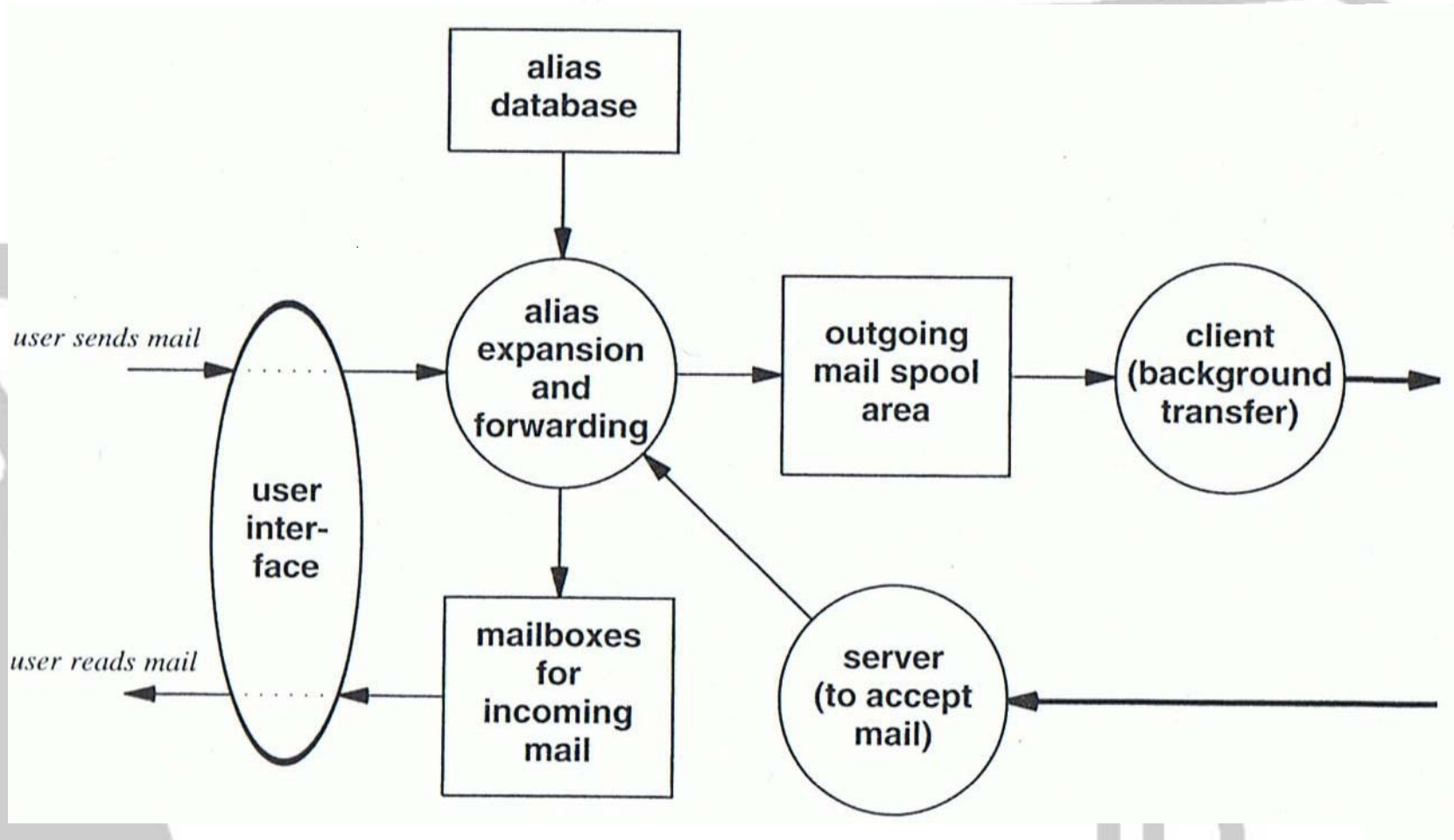
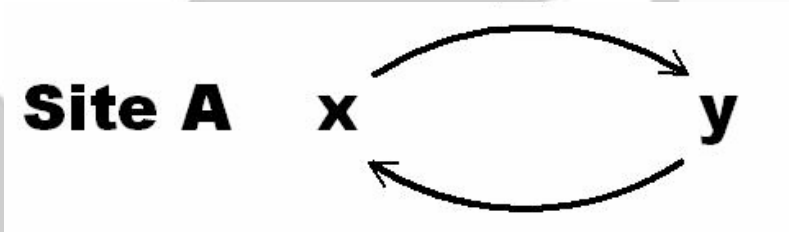


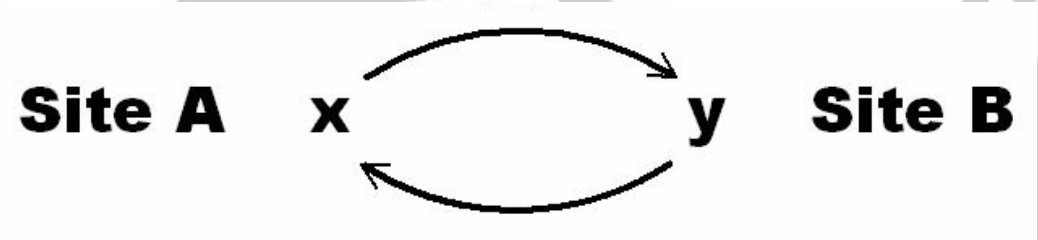
Figure 26.2 An extension of the mail system, Corner

Alias Expansion

- Replacement within a site

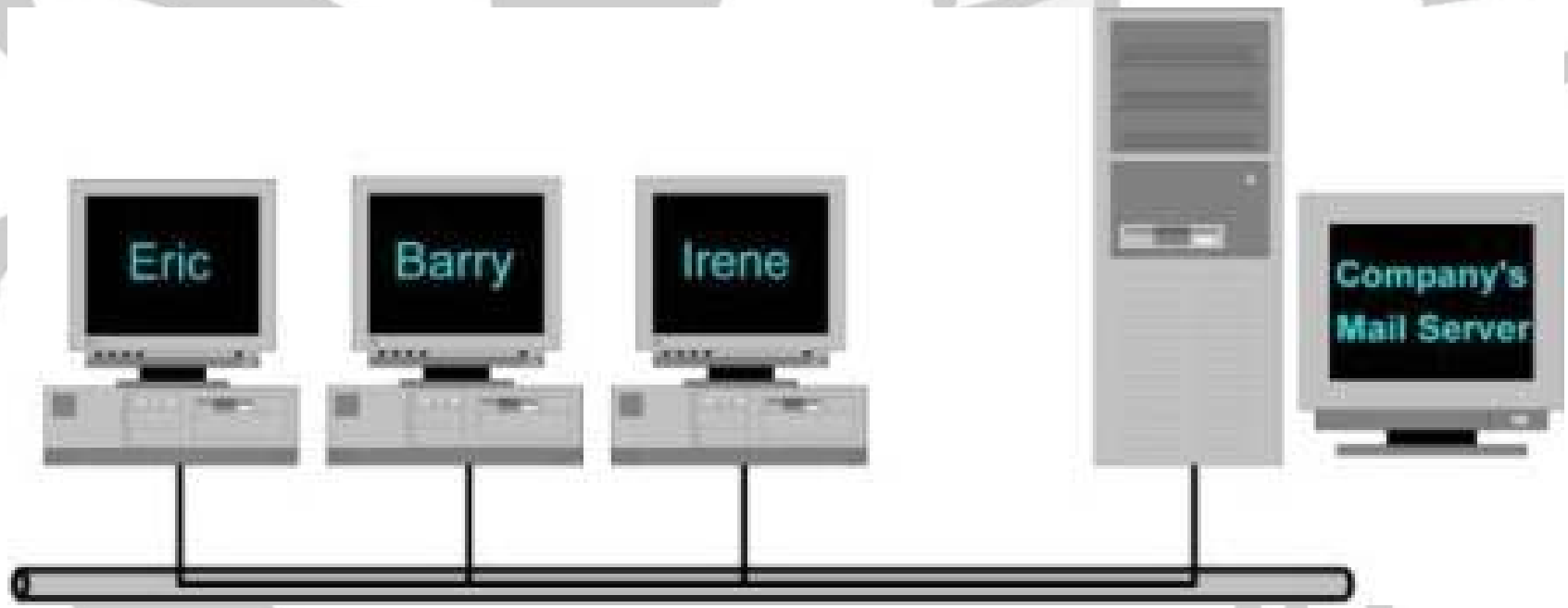


- Conflicting aliases



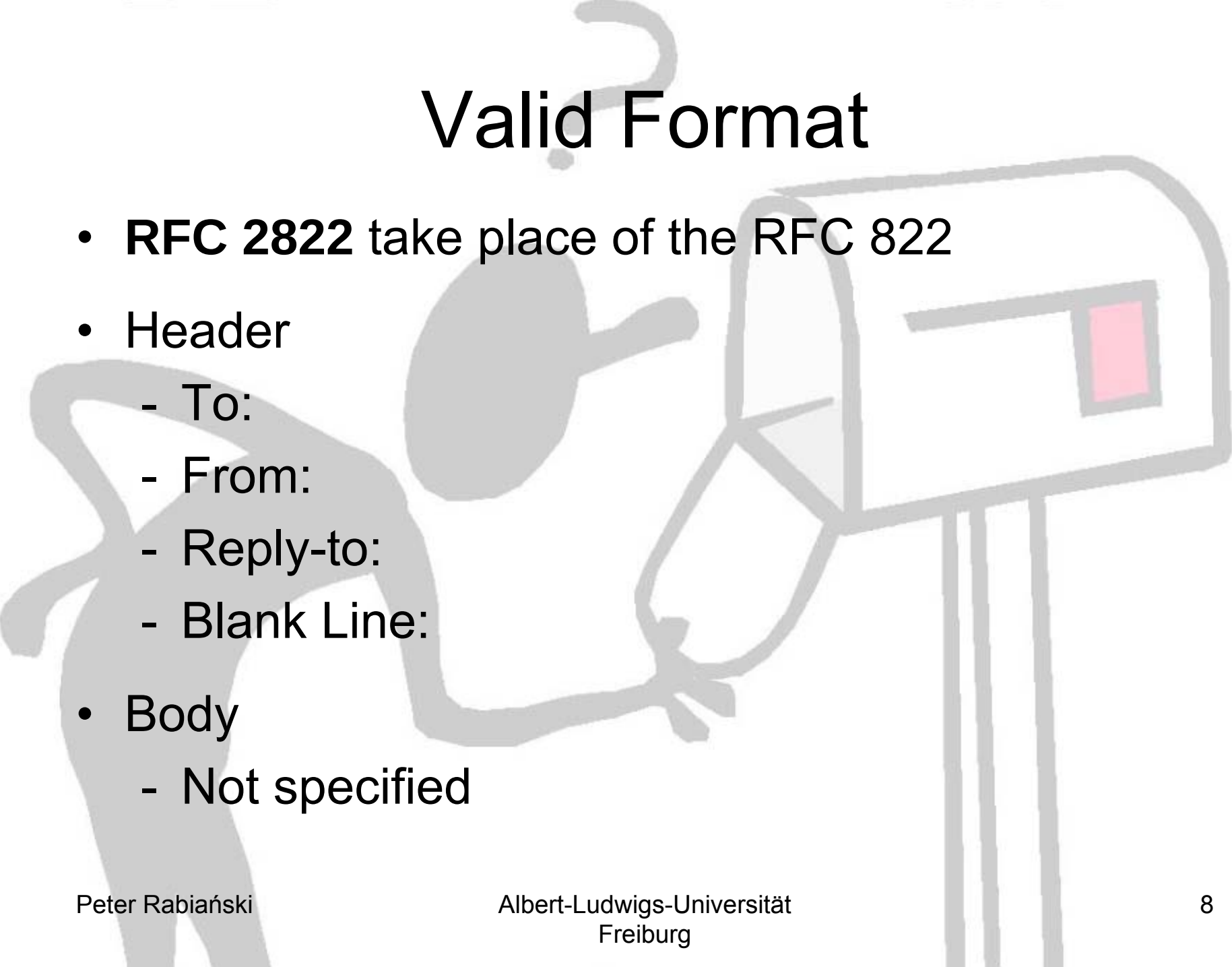
Possible Architecture

- no TCP/IP Connection needed
- Sufficient?



Valid Format

- **RFC 2822** take place of the RFC 822
- Header
  - To:
  - From:
  - Reply-to:
  - Blank Line:
- Body
  - Not specified



Simple Mail Transfer Protocol(SMTP)

- Mail transfer from server to another server
- Commuication in ASCII text

  - Abbriviated commands with 3-digit numbers
- Transport Layer Security (TLS) for encrypted session

SMPT Communication 2

```
S: 220 Beta.gov Simple Mail Transfer Service Ready
C: HELO Alpha.edu
S: 250 Beta.gov

C: MAIL FROM <Smith@Alpha.edu>
S: 250 OK

C: RCPT TO:<Jones@Beta.gov>
S: 250 OK

C: RCPT TO:<Green@Beta.gov>
S: 550 No such user here
C: RCPT TO:<Brown@Beta.gov>
S: 250 OK

C: DATA
S: 354 Start mal input; end with <CR><LF>.<CR><LF>
C: ...sends body of mail message...
C: ...continues for as many lines as message contains
C: <CR><LF>.<CR><LF>
S: 250 OK

C: QUIT
S: 221 Beta.gov Service closing transmission channel
```

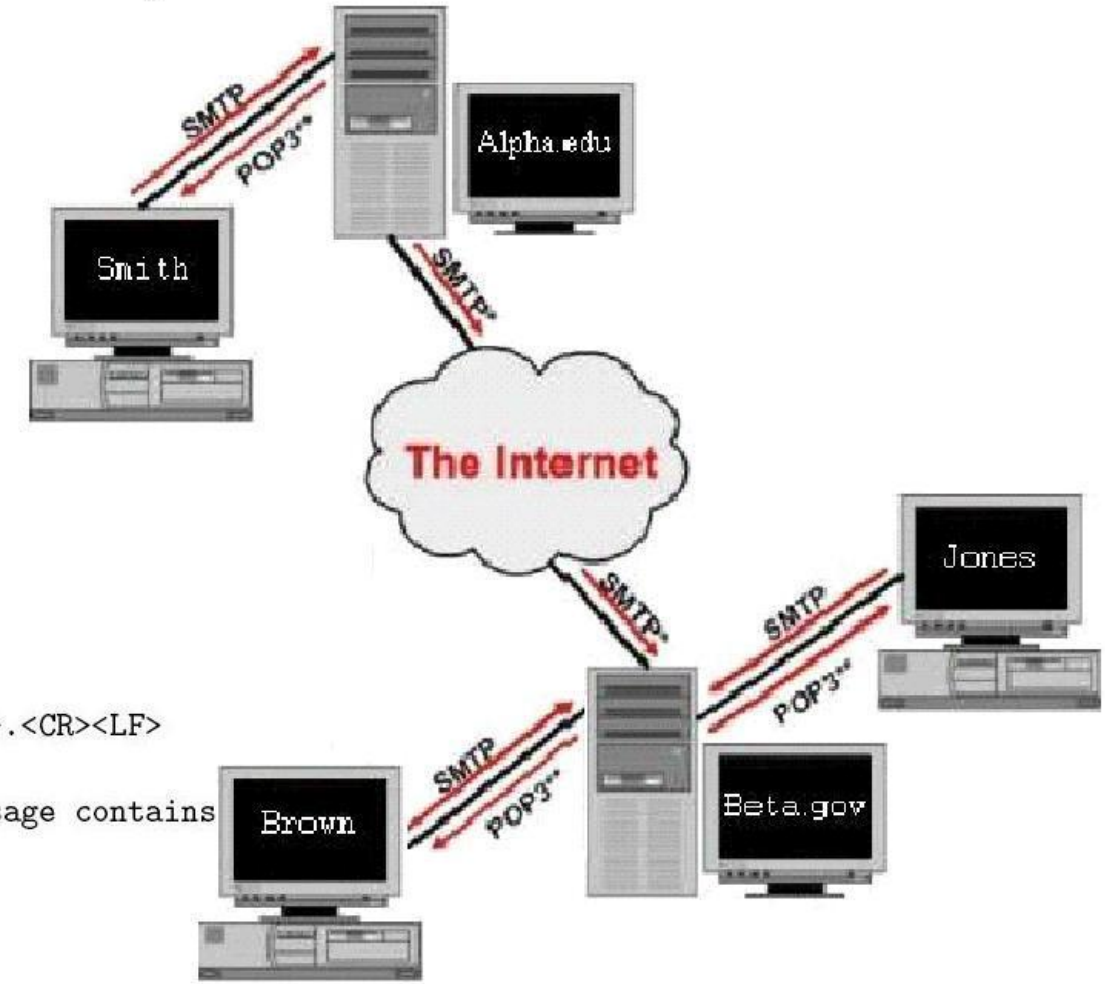
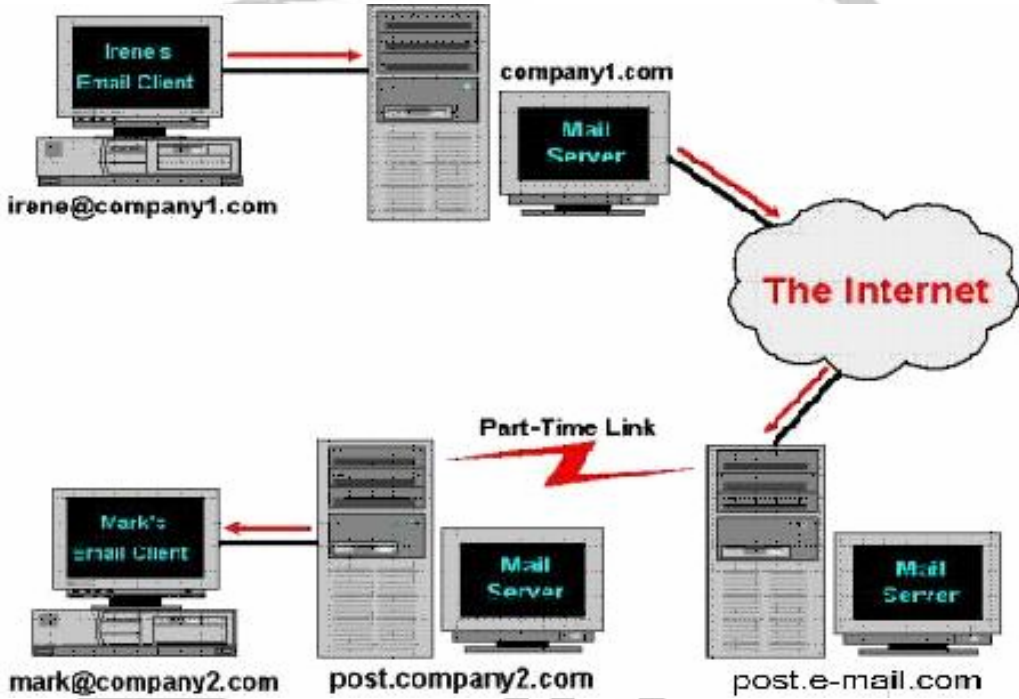


Figure 26.3 [4, chap.26]

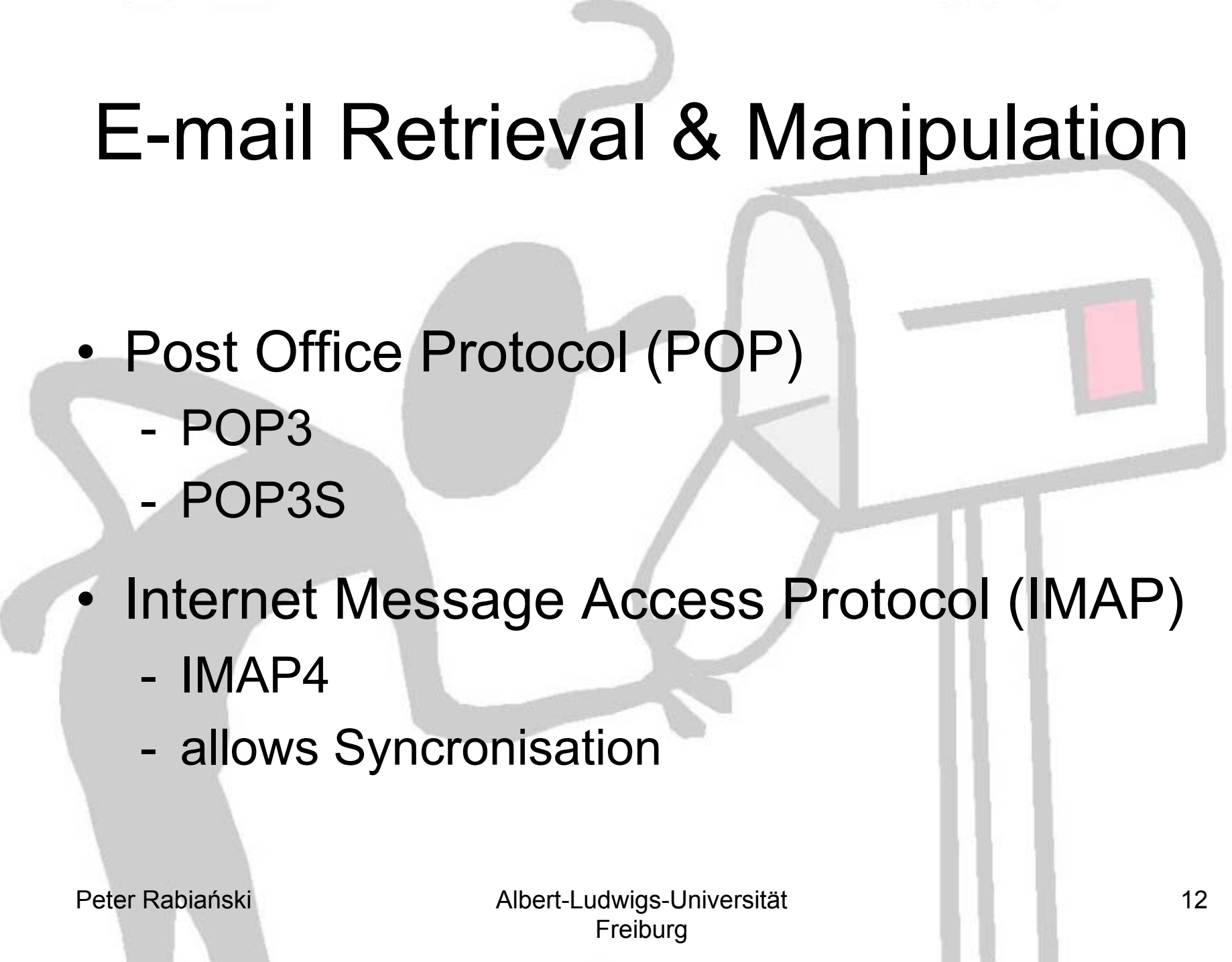
MX Record Email eXchanger

- DNS to decouple mail destination from the domain name assigned to machine than a ping request
- MX Record (Mail eXchanger)



$$\underbrace{e-mail.com}_{DomainName}\ IN\ MX\ \underbrace{10}_{Priority}\ \underbrace{post.e-mail.com}_{MailServer}$$

E-mail Retrieval & Manipulation



- Post Office Protocol (POP)
  - POP3
  - POP3S
- Internet Message Access Protocol (IMAP)
  - IMAP4
  - allows Syncronisation

Multipurpose Internet Mail Extensions (MIME)

- Transmission of non-ASCII data through email
- 7-bit ASCII coding
- RFC 2822 Format
- MIME-Version
- Content-Type
- Content-Transfer-Encoding
- Base64 for sixty-four ASCII characters

Content Types

- Text
  - Textual document
- Image
  - Photograph or computer generated image
- Audio
  - Sound Recording
- Video
  - Video Recording with motion
- Application
  - Raw data for a program
- Multipart
  - Messages with separate content type and encoding
- Message
  - Forwarded an entire e mail

MIME Multipart Messages

- Mixed
  - one message contain multiple, independent submessages with independent type and encoding
- Alternative
  - multiple representation of the same data
- Parallel
  - single message includes subparts to be viewed together (audio and video)
- Digest
  - single message contain a set of other messages

MIME Multipart Messages

```
From: bill@acollage.edu
To: john@example.com
MIME-Version: 1.0
Content-Type: Multipart/Mixed: Boundary=StartOfNextPart

--StartOfNextPart
Content-Type: text/plain
Content-Transfer-Encoding: 7bit
John,
 Here is the photo of our research lab that I promised
 to send you. You can see the equipment you donated.

Thanks again,
Bill

--StartOfNextPart
Content-Type: image/gif
Content-Transfer-Encoding: base64
...data for the image...
```

Example Corner

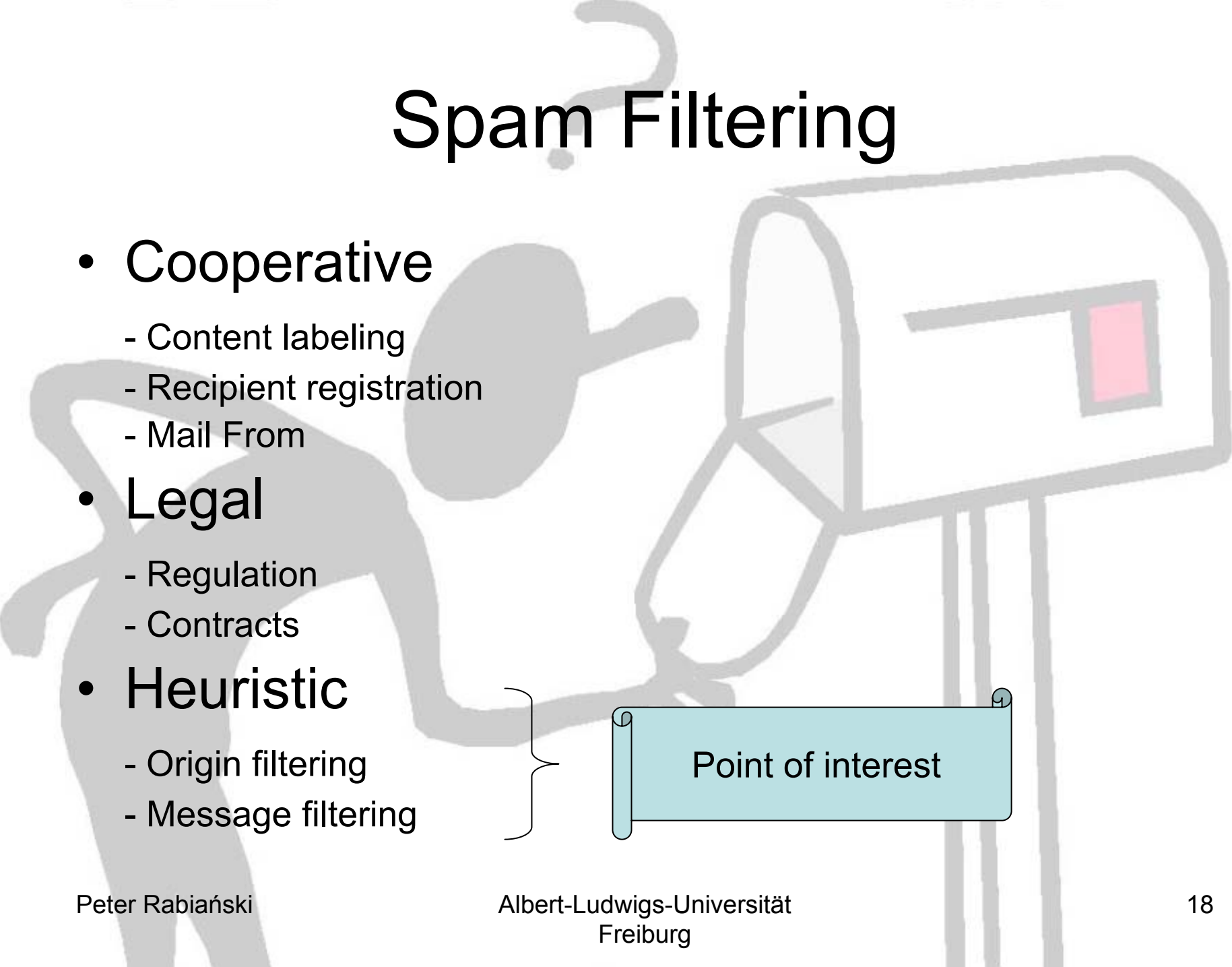
Figure 26.6

Spam

- Unsolicited Bulk Mail (UBE)
- 94% of the e-mail
- 10 Billion € connetion cost per year
- Waste of time and resources
- Decrease of trust in email communication
- 0.16% for lose of legitim e-mail

Spam Filtering

- Cooperative
  - Content labeling
  - Recipient registration
  - Mail From
- Legal
  - Regulation
  - Contracts
- Heuristic
  - Origin filtering
  - Message filtering



Origin filtering

- Refusing IP connection from known UBE originators
- Refusing TCP connections from known UBE originators in the SMTP server
- Refusing SMTP messages from known UBE originators at the MAIL FROM command
- Refusing SMTP messages from originators whose domain name doesn't match their IP address

Message filtering

- In the message store
- At the mail client
  - filter for particular key words, pattern, bag of words
  - with classifier
  - heuristic rules
  - risk of losing or mislabeling ham mail
  - delay

Classification

- tp - ham correctly predicted
- tn - spam correctly predicted
- fp - spam misclassified as ham
- fn - ham misclassified as spam

Category

Prediction	ham(+)	spam(-)
ham(+)	tp	fp
spam(-)	fn	tn

[5] Table 2 SpamAssassin

Statistical Approach: Naive Bayes

- Divide emails in spam and ham
- Assign probalitities to frequently occuring good and bad words
- Use words with pob. far from 0.5
- Calculate probabilities of 15 interesting bad or good words
- New word with 0.4 of spam probability as neutral

Prior doesn't have to be the same for different users

$$\Pr(\text{spam}|\text{words}) = \frac{\Pr(\text{words}|\text{spam})\Pr(\text{spam})}{\Pr(\text{words})}$$

Spam vs. Ham

madam	0.99	continuation	0.01
promotion	0.99	describe	0.01
republic	0.99	continuations	0.01
shortest	0.047225013	example	0.033600237
mandatory	0.047225013	programming	0.05214485
standardization	0.07347802	i'm	0.055427782
sorry	0.08221981	examples	0.07972858
supported	0.09019077	color	0.9189189
people's	0.09019077	localhost	0.09883721
enter	0.9075001	hi	0.116539136
quality	0.8921298	california	0.84421706
organization	0.12454646	same	0.15981844
investment	0.8568143	spot	0.1654587
very	0.14758544	us-ascii	0.16804294
valuable	0.82347786	what	0.19212411

[6] www.paulgraham.com/spam.html

SpamAssassin

- Open-source hybrid spam filter of:
  - bayesian learner
  - set of 500+ heuristics
  - each heuristic has a weight score
  - each rule represented as binary attribute

Total Cost Ratio (TCR)

- Sensitive personal messages = 1000
- Business related messages = 500
- E-commerce related message = 100
- Mailing lists / discussion forums = 50
- Promotional offers = 25
- Cost of misclassifying spam = 1

$$TCR = \frac{fp+tn}{\sum_{x \in fn} C(x)+fp} \qquad TCR < 1$$

[5] SpamAssassin

Error Rates

- Estimated probability of misclassification from ham to spam

$$ham_e = \frac{fn}{fn+tp}$$

- Estimated probability of misclassification from spam to ham

$$spam_e = \frac{fp}{fp+tn}$$

Test

- hx – 475 ham mailbox
- ix – 2163 ham mailbox
- ux – 363 ham mailbox
- *- no true ham was misclassified

$$ham_e = \frac{fn}{fn+tp}$$

	as_train	as_test	hx	ix	ux
SA	36.9%	58.5%	0.2%	0.3%	0.8%
SAnb	*7.79%	*8.6%	0.4%	0.1%	1.4%
Log	0.65%	*1.7%	2.7%	1.8%	1.1%
SMO	0.68%	*1.2%	2.7%	0.9%	1.1%
MLR	*0.65%	*1.9%	2.7%	1.3%	1.1%
J48	0.61%	*1.6%	2.9%	0.9%	1.1%

[5] Table 2 SpamAssassin

Test 2

Standard deviation for ham error rates

- V0 to V6- various approches for testing
- Ham split into F1 and F2
- Two fold crosvalidation (CV)

Like human error rate of 0.16%

	hx	ix	ux
V0	2.53%	1.20%	1.10%
V1	4.63±2.38%	0.79±0.07%	1.65±2.34%
V2	2.53±0.60%	0.79±0.07%	0.55±0.78%
V3	5.68±2.08%	1.16±0.33%	2.20±0.78%
V4	0.63±0.30%	0.14±0.07%	0.55±0.78%
V5	0.21±0.30%	0.32±0.07%	0.55±0.78%
V6	0.21±0.30%	0.23±0.07%	1.10±0.00%

[5] Table 3 SpamAssassin

Best Models

Spam error rates

- Good performance of SAnb
- Improvement of factor four is possible
- Spam error rate from as_test

	hx	ix	ux
SA		64.7%	
SAnb		9.5%	
V4	7.3%	45.5%	2.1%
V5	3.2%	4.9%	2.1%
V6	2.1%	2.7%	1.5%

[5] Table 4 SpamAssassin

threshold value of 1.0 from default value 0.5

$$spam_e = \frac{fp}{fp+tn}$$

Total Cost Ratio

- TCR in each case better as manual deletion
- TCR 36.42 vs. 1

$$TCR = \frac{fp+tn}{\sum_{x \in fn} C(x)+fp}$$

	hx	ix	ux
SA	1.53	1.39	1.12
SAnb	9.31	7.29	2.02
V4	10.92	2.07	5.46
V5	26.18	10.50	5.46
V6	36.42	17.25	2.92

[5]Table 5 SpamAssassin

Conclusions

- A well-trained Bayesian model as core of a good spam filter
- heuristics alone in any case insufficient
- No direct content features are available:
  - such as phrase occurrence
  - filtering for word pairs, or even triples words
- Example:
  - word "offers" has a probability of 0.96
  - "special offers" and "valuable offers" o 0.99
  - "approach offers" (as "this approach offers") 0.1 or less

References

1. Comer,Chap.26, Electronic Mail (SMTP, POP, IMAP, MIME)
2. P. Resnick, RFC 2822, 2001
3. Alexander K. Seewald, Combining Bayesian and Rule Score Learning: Automated Tuning for SpamAssassin
4. Graham, Paul. A Plan For Spam, www.paulgraham.com/spam.html, 2002
5. Graham, Paul. Better Bayesian Filter, www.paulgraham.com/spam.html, 2003
6. Paul Hoffman and Dave Crocker, Unsolicited Bulk Email: Mechanisms for Control, Internet Mail Consortium IMCR-005, October 13, 1997
7. http://www.tekguard.com/Content/Software/Tutorials/EMailTutorial_P2.htm
8. Wikipedia