Collaborative E-Mail Filtering

Doug Herbers

Master’s Oral Defense

June 28, 2005

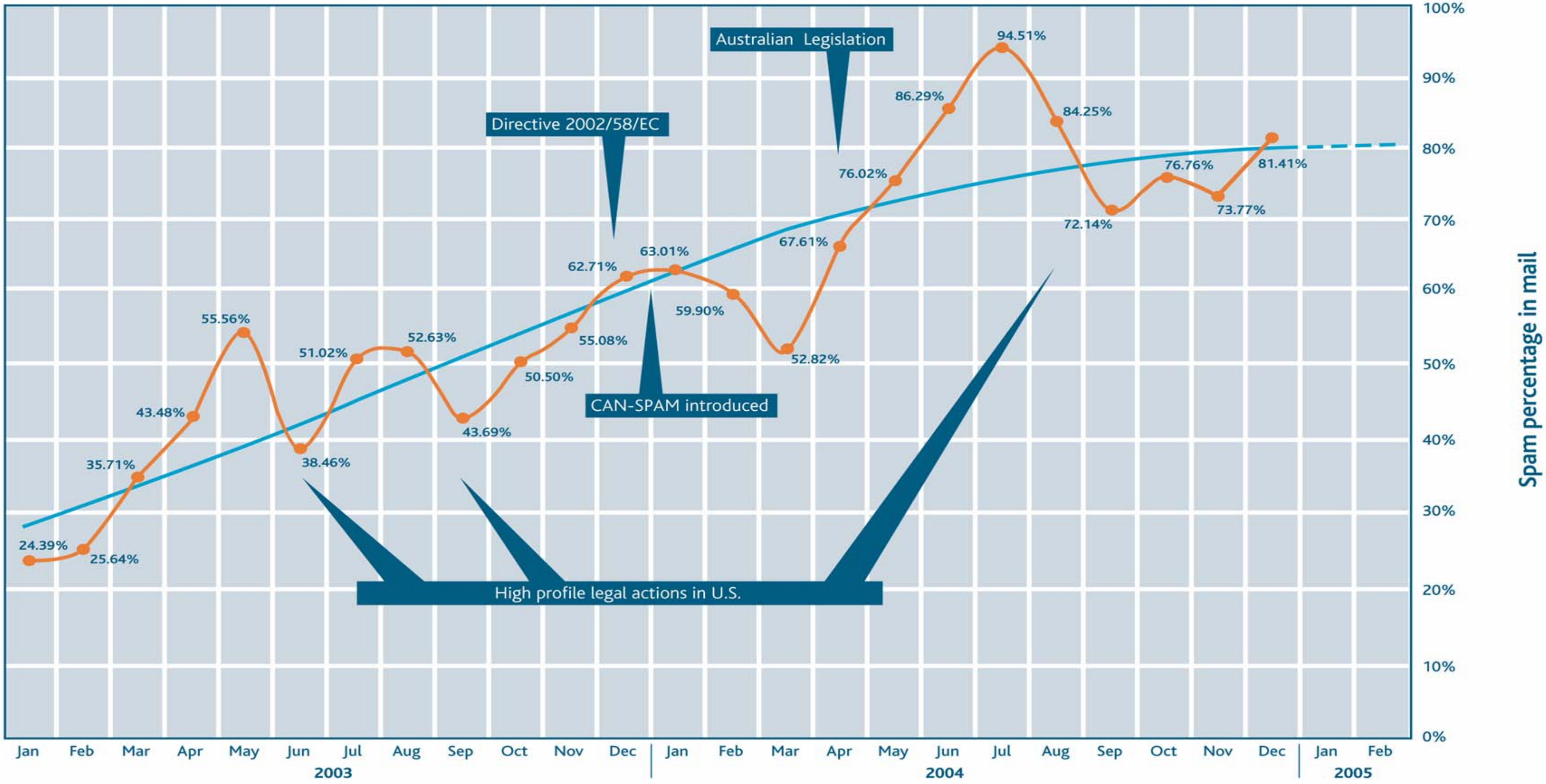
Background

- Spamming – the use of any electronic communications medium to send unsolicited messages in bulk
- E-Mail is the most common medium, also cell phones, text messaging, and pagers
- SPAM has developed a negative reputation
- SPAM ~ Door-to-Door Sales ~ Junk Postal Mail

E-Mail & SPAM Trends

- On average, 31 billion e-mails were sent each day in 2002
- MSNBC reports 66% of World’s E-Mail is SPAM (May 2004)
- MessageLabs reports 76% of e-mail received by their clients in May 2004 was SPAM, projected 81% by February 2005

SPAM Trends



MessageLabs filtering results of E-Mail Worldwide

Legislation & Litigation Make Short-Term Decreases in SPAM

- CAN-SPAM (Controlling the Assault of Non-Solicited Pornography and Marketing Act), December 2003
- Message must have valid headers
- Subject must represent content of e-mail
- Message must include a valid postal address of the sender
- Message must include an unsubscribe notification, by which e-mails will cease in less than 10 days after submission

Why is SPAM Harmful?

- Decreased productivity for employees – according to estimates, a company with 200 employees will waste about 5000 minutes per month, up to $3,000 per month dealing with SPAM
- Exposing children to inappropriate material
- Congestion of Internet Service Provider’s Networks – costs are passed on to consumers
- Scams and devious behavior, virus, denial of service attacks, etc.

Why Do We Need a Filter?

- Decrease the quantity of SPAM messages in our inboxes
- Protect minors from inappropriate content
- Protect from scams
- Protect from viruses

Previous Filtering Techniques

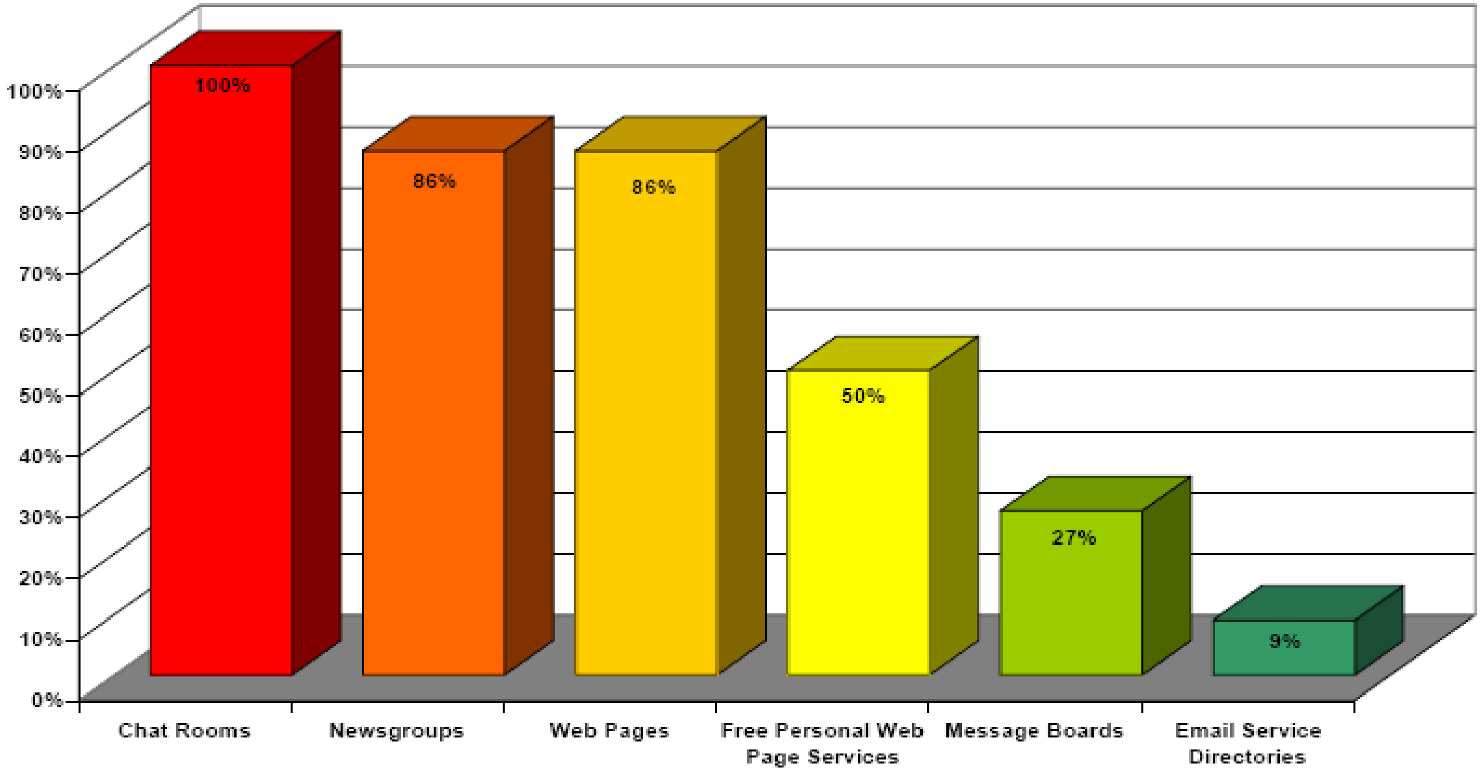
- Rule-Based Systems – Rules that have to be manually changed to adapt to new SPAM
- Statistical (TF-IDF) – Statistical based on word frequencies
- *Naïve Bayes – Probabilistic, trained on e-mails, will adapt and get better over time, false positive rates less than 1 in 1000.
- Memory-Based Filtering – Vector-based, judgments made by considering kNN related e-mail message vectors

Previous Filtering Techniques

- Blacklists and Whitelists – allow or deny specific users to sent you e-mail, usually requires some form of handshaking
- Collaborative Filtering – local vs worldwide model

E-Mail Harvesting

Email Address Harvesting by Forum



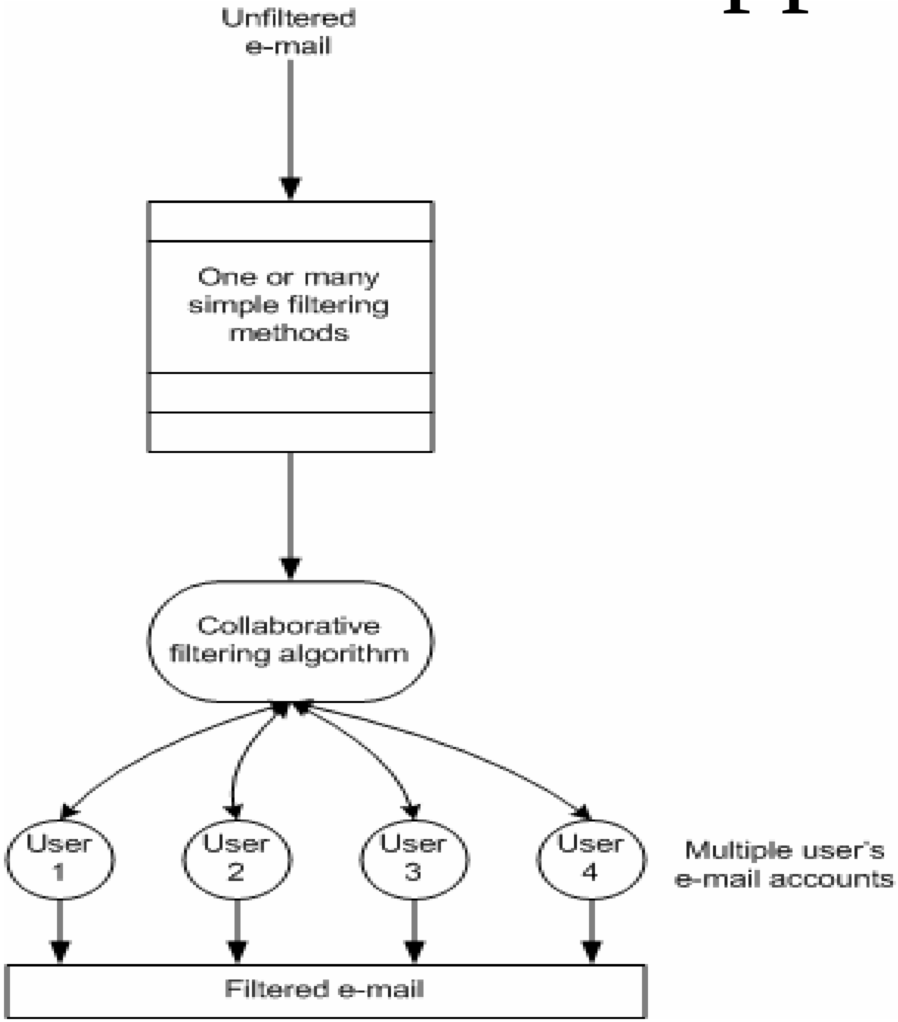
No spam was received at email addresses for "Whois" Domain Name Information,
Instant Message Service User Profile, Online Dating Services and Online Resume Services.

Source: Northeast Netforce investigators seeded 175 different locations on the Internet with 250 new, undercover email addresses and monitored the addresses for 6 weeks.

Goals

- Use multiple user’s e-mail to identify and remove SPAM
- Apply algorithms at the user-level and system-level, and compare results
- Show an improvement over SpamAssassin alone

Approach


Unfiltered e-mail
One or many simple filtering methods
Collaborative filtering algorithm
User 1
User 2
User 3
User 4
Multiple user's e-mail accounts
Filtered e-mail

- Preliminary Filter – SpamAssassin
- Identify duplicated messages
- Remove all duplicate messages
- Evaluate various definitions of duplicate

Data Collection

- Collection of two weeks of e-mail
- Week 1 – Test Data Set
- Week 2 – Validation Data Set

User Set

- Sixteen Volunteers from ITTC
- 2 Professors
- 3 Ph.D. Students
- 7 M.S. Students
- 2 B.S. Students
- 2 Staff Members

E-Mail Classification

- Classification of e-mail determined by what the user did with each message

Location	Read/Unread	Classification
Inbox	Read	Legitimate
Inbox	Unread	Void
SPAM Folder	Read or Unread	SPAM
Trash	Read	Legitimate
Trash	Unread	SPAM

Week 1 - Test Data Set

User	Total Messages	Legitimate Messages	%	SPAM	%	SpamAssassin	%	Void Messages
1	818	70	9 %	747	91 %	675	90 %	1
2	17	7	41 %	1	6 %	0	0 %	9
3	51	45	88 %	6	12 %	0	0 %	0
4	922	236	26 %	676	73 %	641	95 %	10
5	434	105	24 %	324	75 %	292	90 %	5
6	11	11	100 %	0	0 %	0	0 %	0
7	8	0	0 %	7	88 %	0	0 %	1
8	8	7	88 %	0	0 %	0	0 %	1
9	54	12	22 %	19	35 %	16	84 %	23
10	305	32	10 %	252	83 %	194	77 %	21
11	3	3	100 %	0	0 %	0	0 %	0
12	8	2	25 %	0	0 %	0	0 %	6
13	106	5	5 %	89	84 %	8	9 %	12
14	1516	228	13 %	1208	85 %	1088	85 %	2
15	0	0	0 %	0	0 %	0	0 %	0
16	48	43	90 %	1	2 %	0	0 %	4
Totals	**4309**	**806**	**19 %**	**3408**	**79 %**	**2914**	**86 %**	**95**

User Selection

- > 20 Messages Total
- > 1 SPAM Message
- > 1 Legitimate Message

- Nine Users Remain

Determination of Baseline

- Remove all void messages (95)
- Remove intra-server e-mail (514)
- Remove all messages tagged as SPAM by SpamAssassin (2883)

	Tagged as Legitimate	Tagged as SPAM
Legitimate According to User	338	0
SPAM According to User	441	2883

Revised Data Set

User	**Total Messages**	**Legitimate**	**%**	**SPAM**	**%**
1	124	52	42 %	72	58 %
3	5	4	80 %	1	20 %
4	115	82	71 %	33	29 %
5	90	58	64 %	32	26 %
9	3	0	0 %	3	100 %
10	59	16	27 %	43	73 %
13	82	4	5 %	78	95 %
14	272	93	34 %	179	66 %
16	29	29	100 %	0	0 %
Totals	**779**	**338**	**43** %	**441**	**57** %

Evaluation Criteria

	Tagged as Legitimate	Tagged as SPAM
Legitimate According to User	Legitimate Passed	Legitimate Removed (False Positive)
SPAM According to User	SPAM Passed (False Negative)	SPAM Removed

$$\text{Recall} = \frac{\text{Legitimate Passed}}{\text{Legitimate Passed} + \text{Legitimate Removed}}$$

$$\text{Precision} = \frac{\text{Legitimate Passed}}{\text{Legitimate Passed} + \text{SPAM Passed}}$$

$$\text{Accuracy} = \frac{\text{Legitimate Passed} + \text{SPAM Removed}}{\text{All Messages in the Data Set}}$$

Evaluation Criteria (cont.)

$$\mathrm{F}_{Measure} = \frac{(1+\beta^2)\times \mathrm{Precision} \times \mathrm{Recall}}{\beta^2 \times \mathrm{Precision} + \mathrm{Recall}}$$

Chose Beta=2.0 to weight recall higher than precision

User-Level – Remove all duplicates within a user's e-mail box

User 1	User 2	User 3
Msg 1	Msg 2	Msg 3
Msg 2	Msg 2	Msg 4
Msg 5	Msg 6	Msg 6

Message 2 counts as one message with two duplicates

System-Level – Remove all duplicates over all e-mail boxes

User 1	User 2	User 3
Msg 1	Msg 2	Msg 3
Msg 2	Msg 2	Msg 4
Msg 5	Msg 6	Msg 6

Message 2 counts as one message with three duplicates

Message 6 counts as one message with two duplicates

Classification of Msg 2

	Number of Copies	Classify as Legitimate	Classify as SPAM
User 2	2	1	1
User 3	1	0	1

User-Level – 1 legitimate removed & 1 SPAM removed

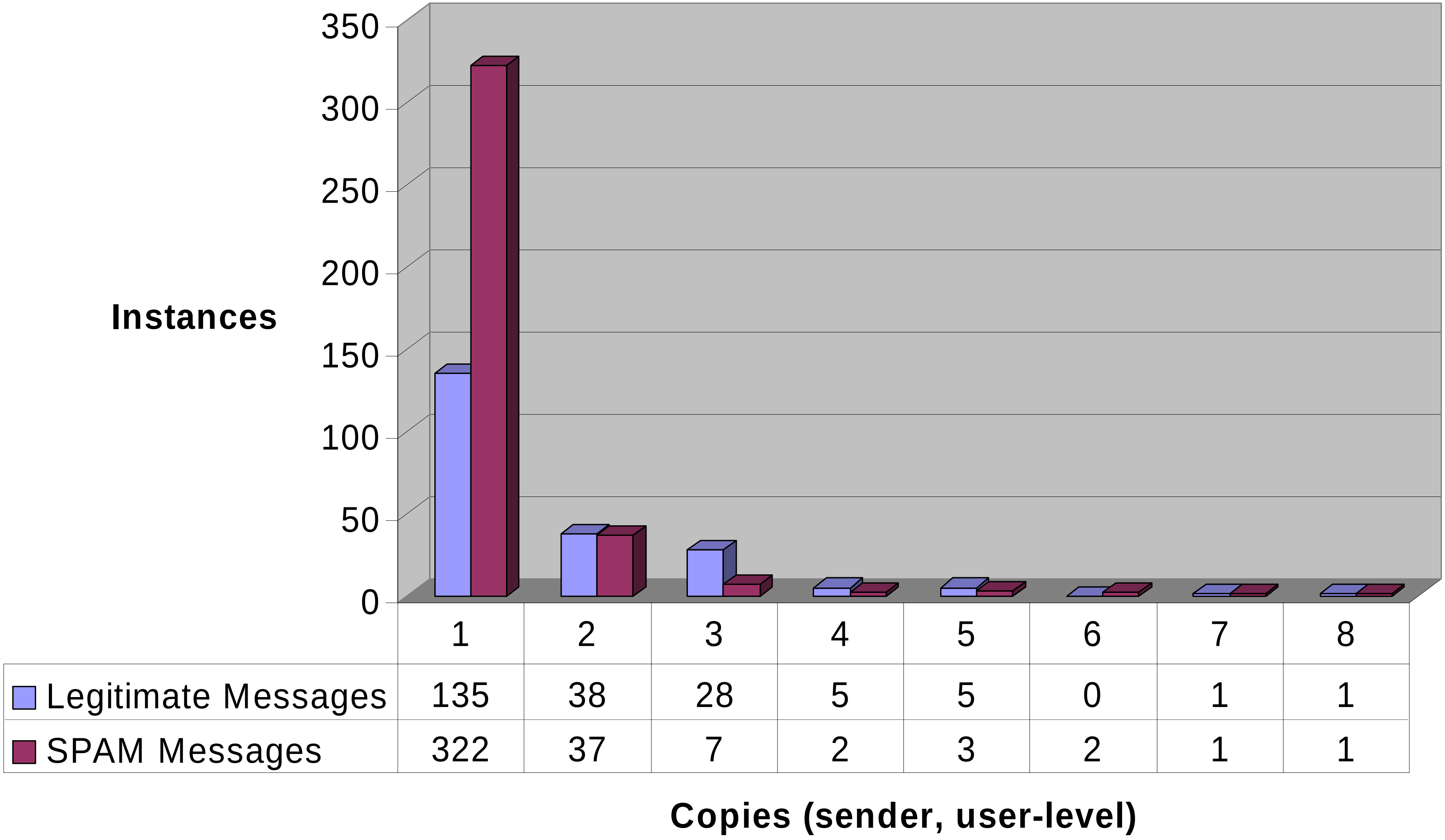
System-Level – 1 legitimate removed & 2 SPAM removed

Qualities of Messages

- Algorithm 1: Subject, User-Level
- Algorithm 2: Subject, System-Level
- Algorithm 3: Sender, User-Level
- Algorithm 4: Sender, System-Level
- Algorithm 5: Body, User-Level
- Algorithm 6: Body, System-Level

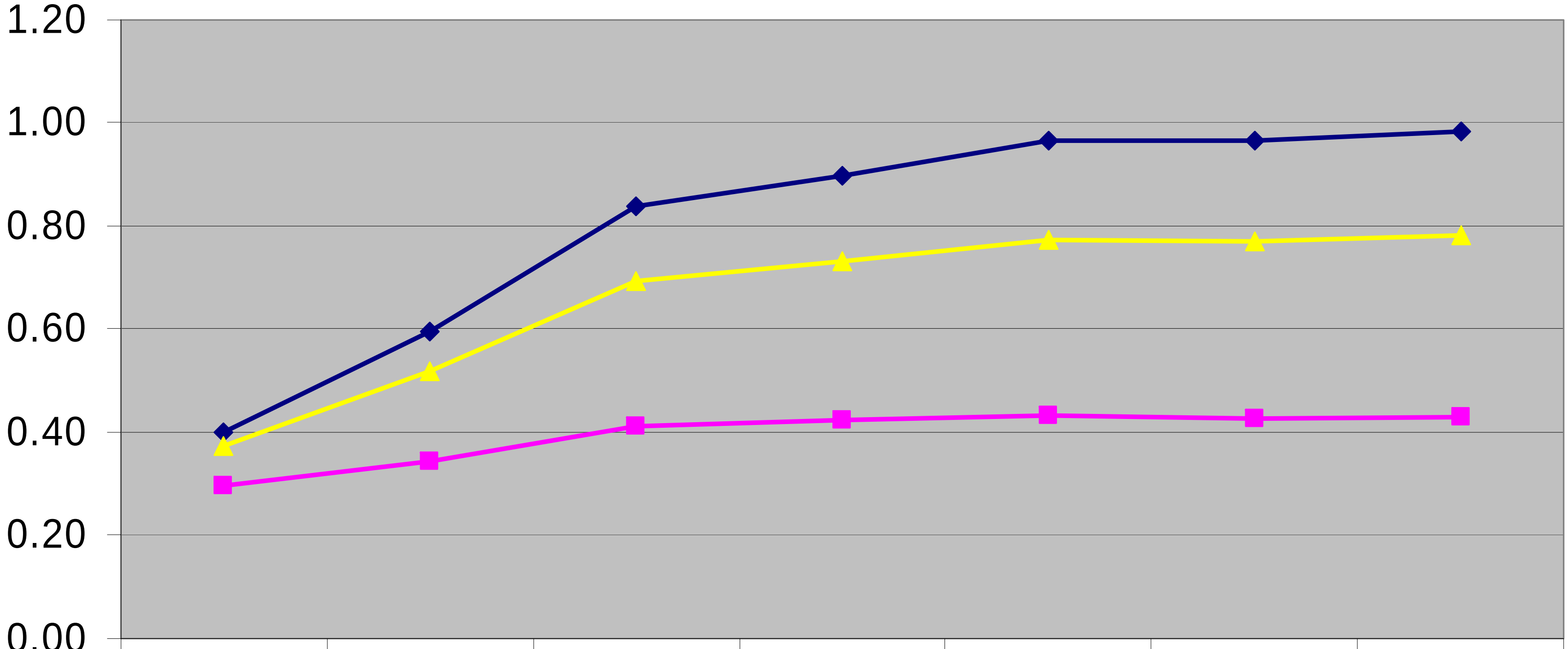
Algorithm 3 – User-Level Sender Duplicates

All Messages not sent from *ku.edu (or *ukans.edu) after Baseline



	1	2	3	4	5	6	7	8
🟦 Legitimate Messages	135	38	28	5	5	0	1	1
🟪 SPAM Messages	322	37	7	2	3	2	1	1

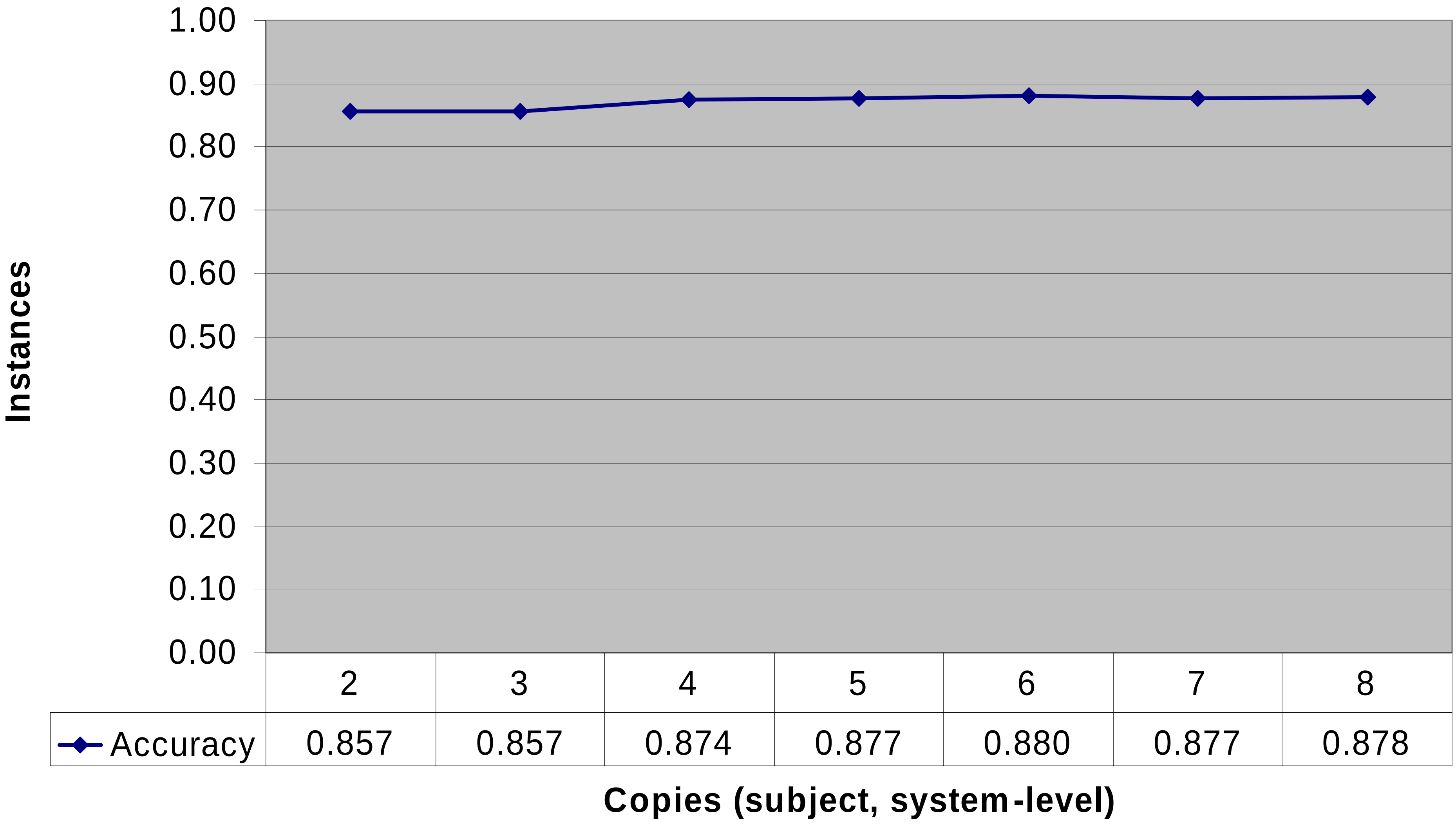
Precision, Recall and F-Measure



	2	3	4	5	6	7	8
◆ Recall	0.399	0.595	0.837	0.896	0.964	0.964	0.982
■ Precision	0.295	0.342	0.411	0.423	0.434	0.427	0.431
▲ F-Measure	0.373	0.518	0.693	0.732	0.775	0.770	0.782

Copies (sender, user-level)

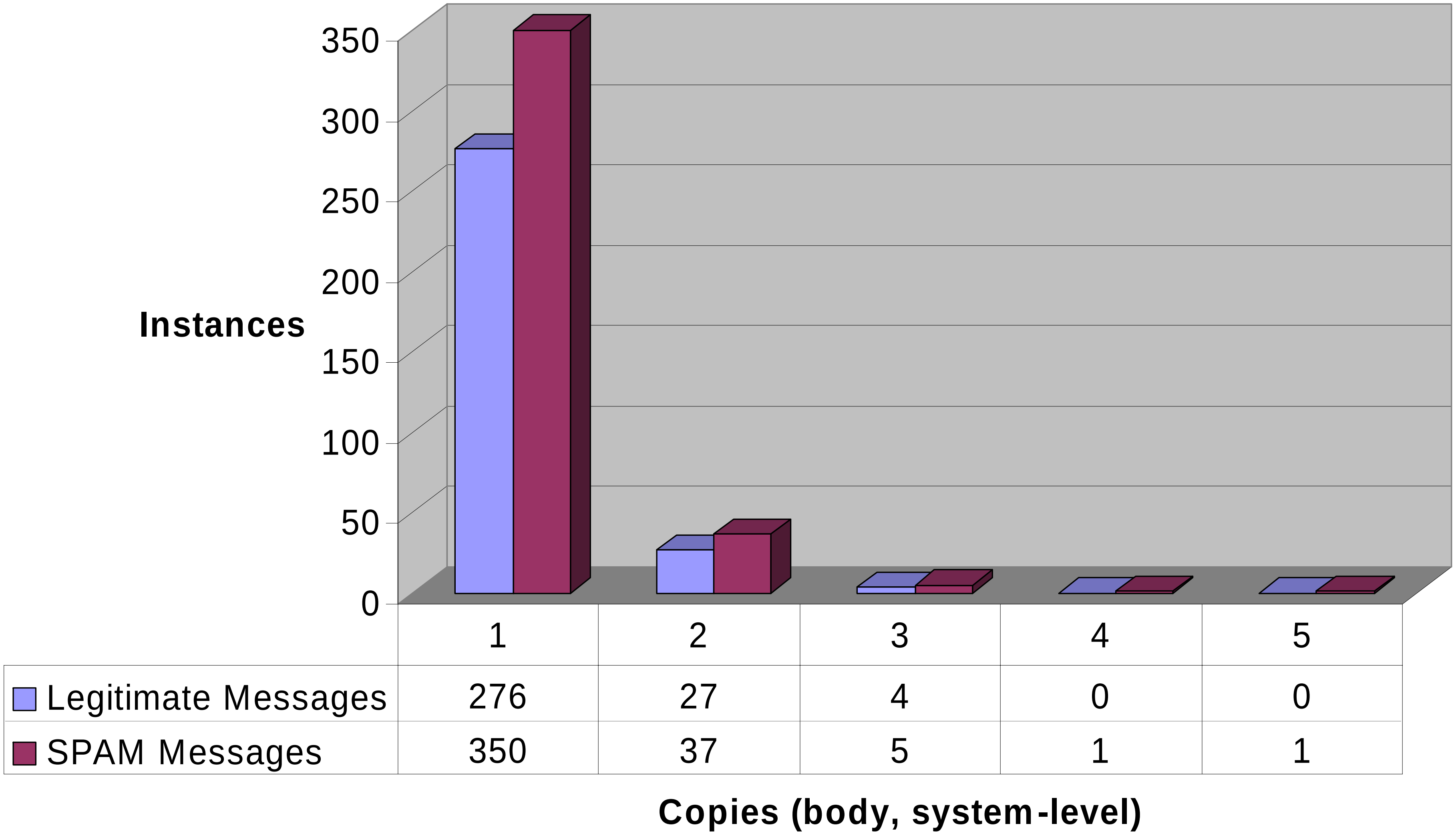
Accuracy of Algorithm 3 and Spam Assassin



	2	3	4	5	6	7	8
◆ Accuracy	0.857	0.857	0.874	0.877	0.880	0.877	0.878

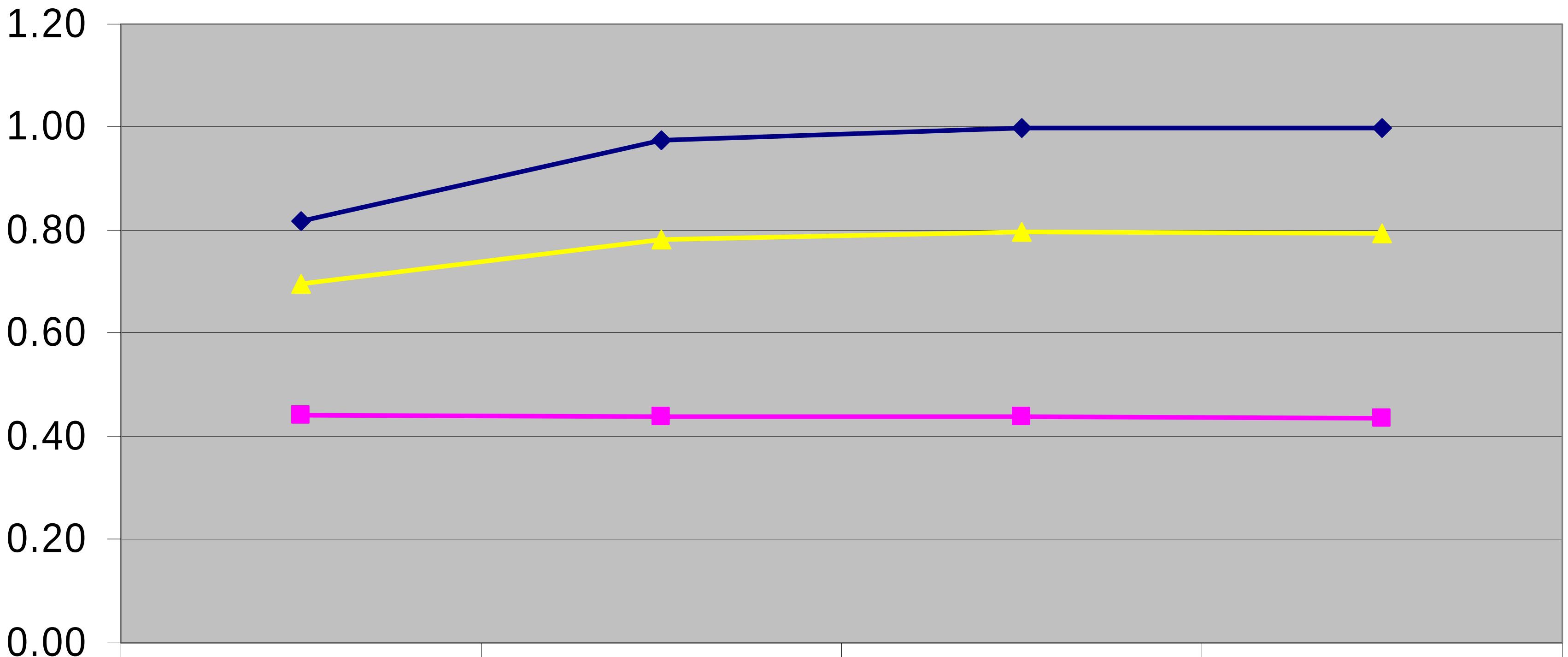
Algorithm 6 – System-Level Body Duplicates

All Messages not sent from *ku.edu (or *ukans.edu) after Baseline



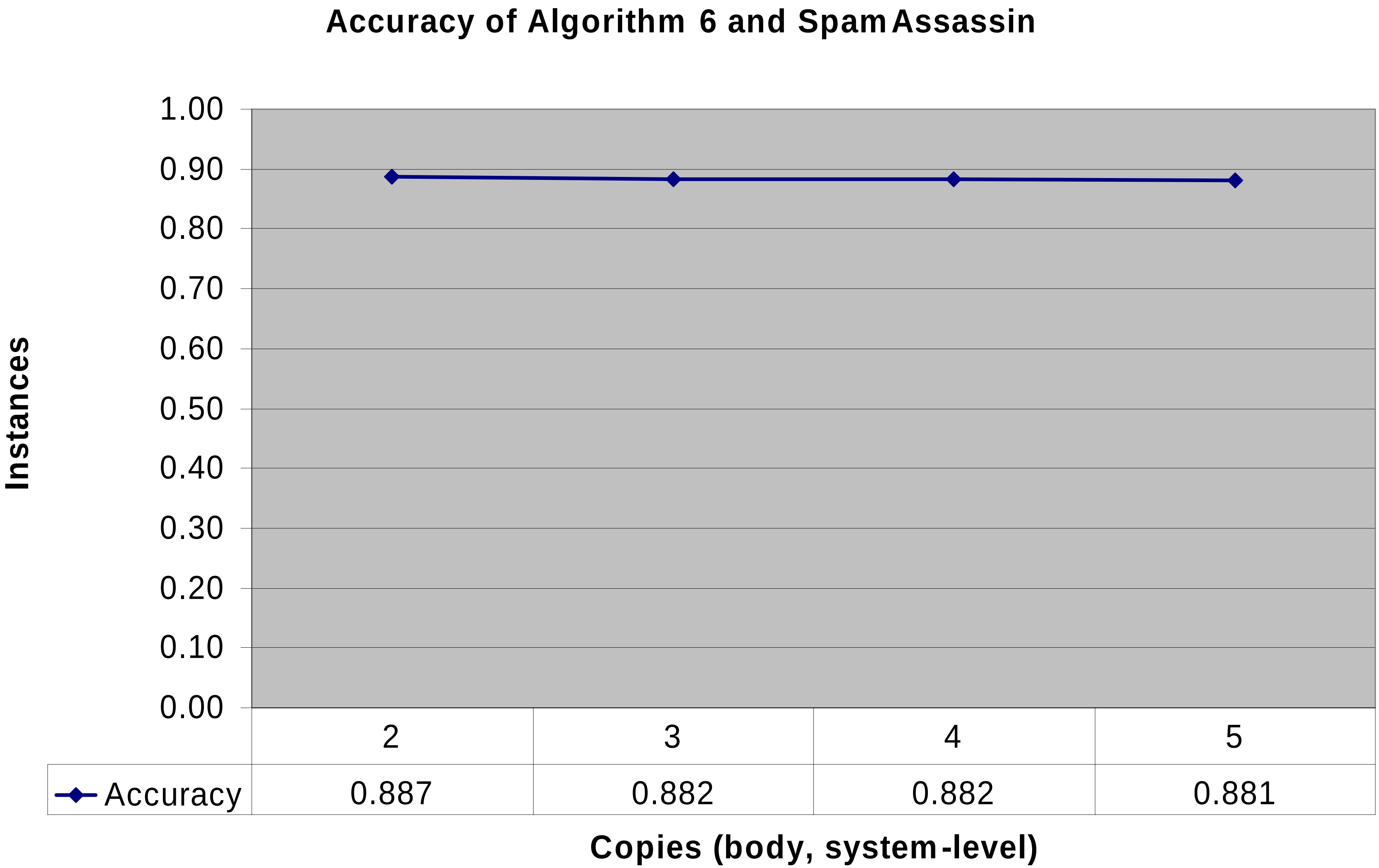
	1	2	3	4	5
▨ Legitimate Messages	276	27	4	0	0
▨ SPAM Messages	350	37	5	1	1

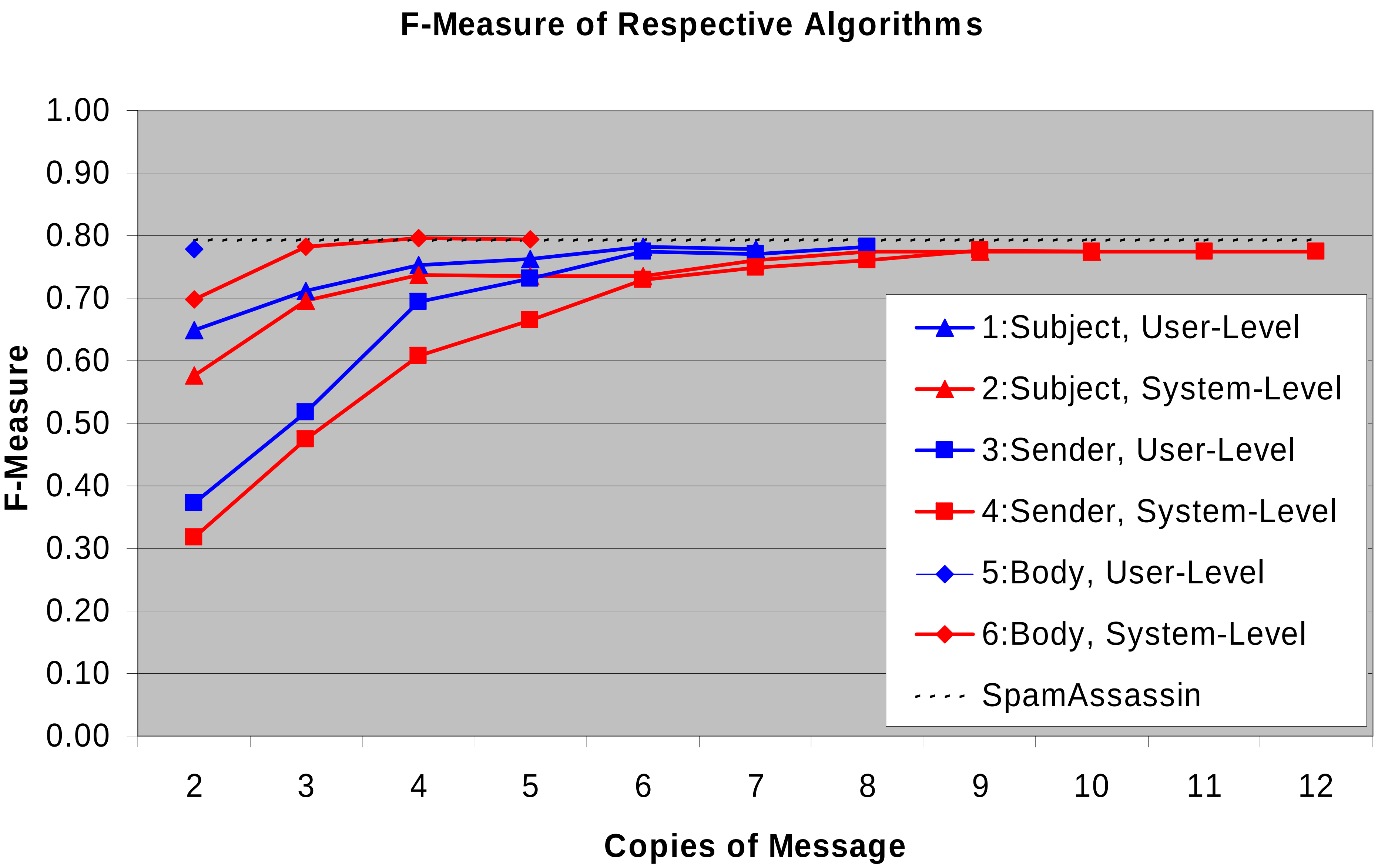
Precision, Recall and F-Measure

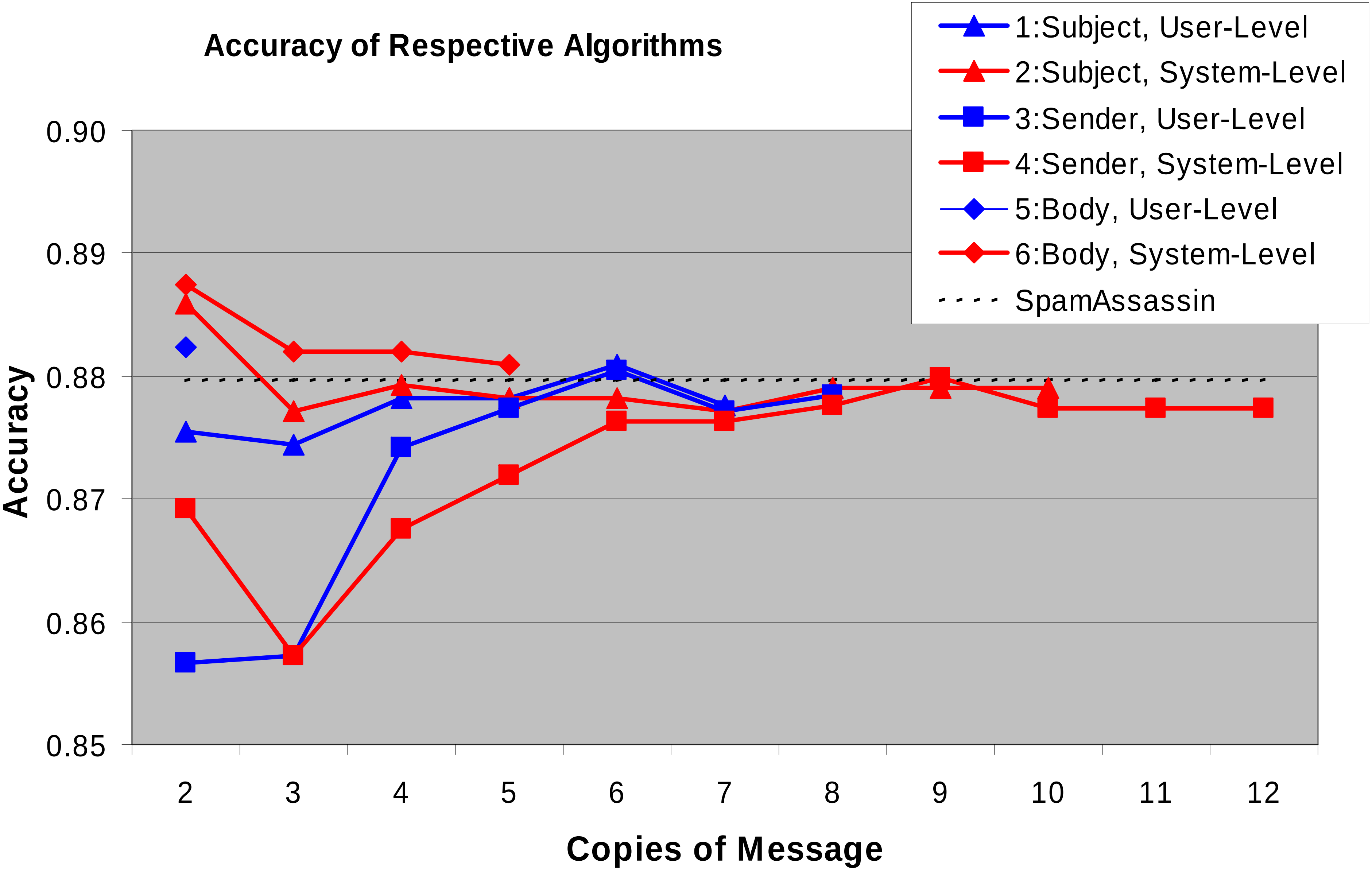


	2	3	4	5
◆ Recall	0.817	0.973	1.000	1.000
■ Precision	0.441	0.438	0.439	0.437
▲ F-Measure	0.698	0.782	0.796	0.795

Copies (body, system-level)


Accuracy of Algorithm 6 and SpamAssassin
Instances
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
2
3
4
5
Accuracy
0.887
0.882
0.882
0.881
Copies (body, system-level)

F-Measure of Respective Algorithms
F-Measure
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
2
3
4
5
6
7
8
9
10
11
12
Copies of Message
1:Subject, User-Level
2:Subject, System-Level
3:Sender, User-Level
4:Sender, System-Level
5:Body, User-Level
6:Body, System-Level
SpamAssassin

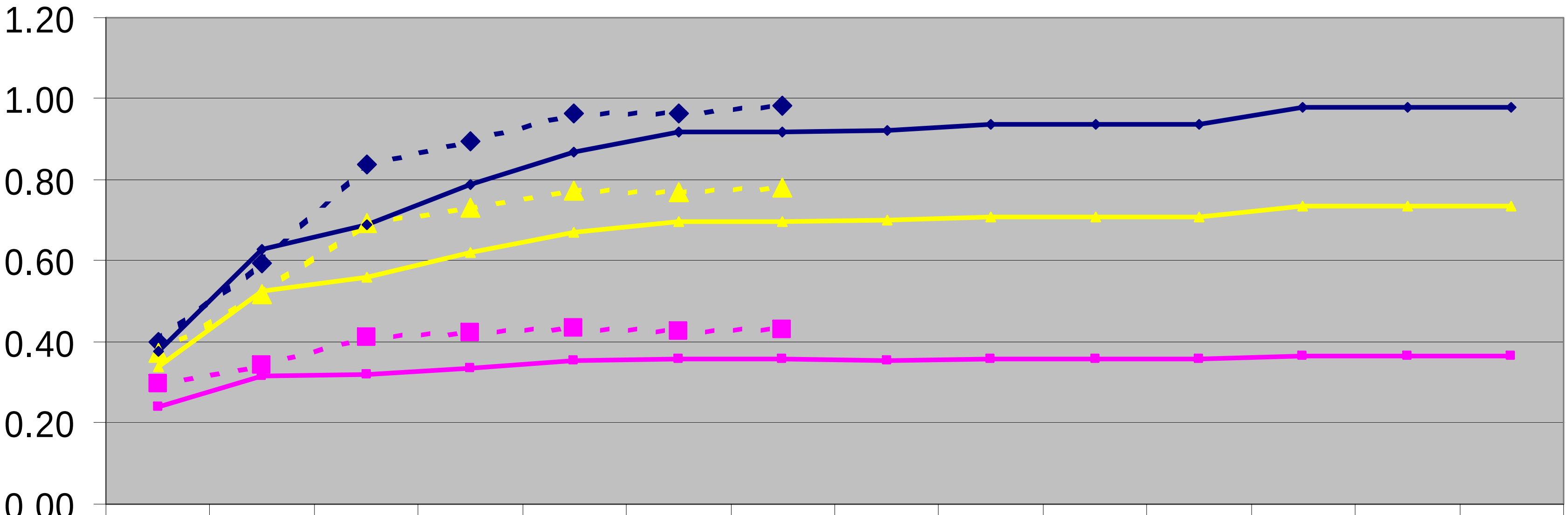

Accuracy of Respective Algorithms
1:Subject, User-Level
2:Subject, System-Level
3:Sender, User-Level
4:Sender, System-Level
5:Body, User-Level
6:Body, System-Level
SpamAssassin
Accuracy
0.90
0.89
0.88
0.87
0.86
0.85
2
3
4
5
6
7
8
9
10
11
12
Copies of Message

Validation

- Chose Best User-Level Algorithm: Algorithm 3, Duplicates Based on Sender

- Chose Best System-Level Algorithm: Algorithm 6, Duplicates Based on Body

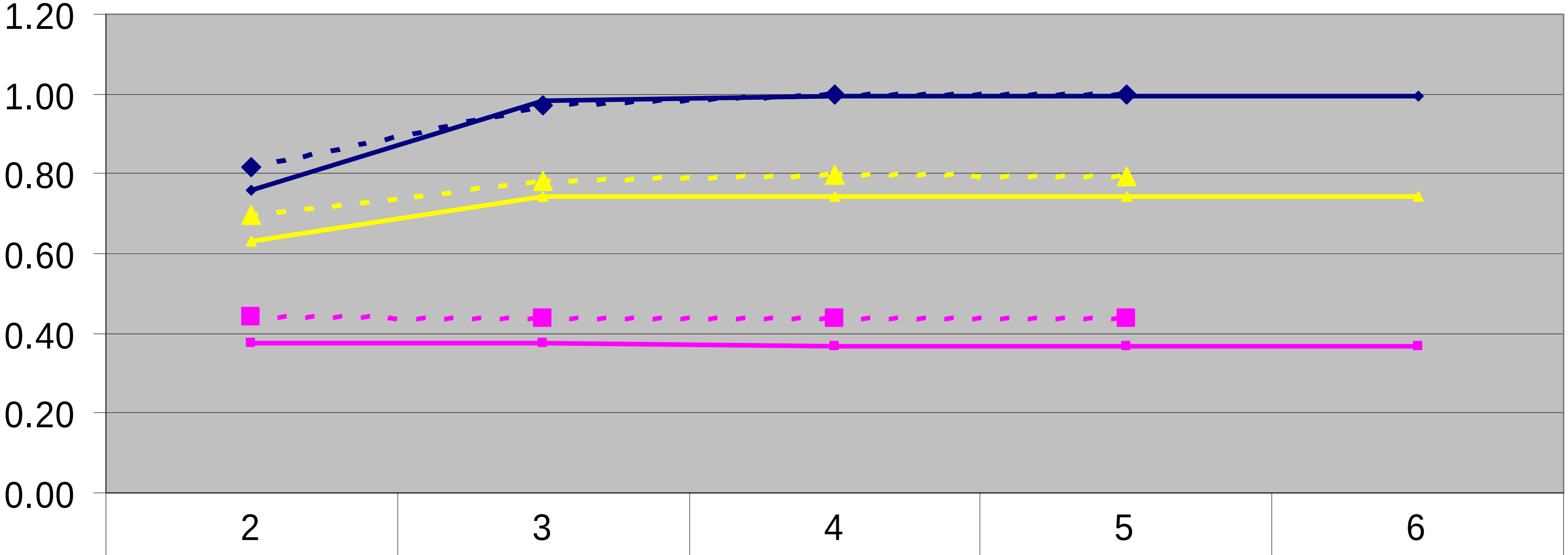
Precision, Recall and F-Measure



	2	3	4	5	6	7	8	9	10	11	12	13	14	15
- ◆ - Recall	0.399	0.595	0.837	0.896	0.964	0.964	0.982							
- ■ - Precision	0.295	0.342	0.411	0.423	0.434	0.427	0.431							
- ▲ - F-Measure	0.373	0.518	0.693	0.732	0.775	0.770	0.782							
—◆— Recall'	0.378	0.628	0.691	0.788	0.868	0.917	0.917	0.924	0.938	0.938	0.938	0.979	0.979	0.979
—■— Precision'	0.242	0.317	0.320	0.337	0.353	0.357	0.357	0.356	0.357	0.357	0.357	0.367	0.367	0.367
—▲— F-Measure'	0.340	0.525	0.561	0.622	0.672	0.698	0.698	0.700	0.708	0.708	0.708	0.734	0.734	0.734

Copies (sender, user-level)

Precision, Recall and F-Measure



	2	3	4	5	6
- ◆ - Recall	0.817	0.973	1.000	1.000	
- ■ - Precision	0.441	0.438	0.439	0.437	
- ▲ - F-Measure	0.698	0.782	0.796	0.795	
—◆— Recall'	0.760	0.983	0.997	0.997	0.997
—■— Precision'	0.375	0.377	0.369	0.369	0.369
—▲— F-Measure'	0.631	0.744	0.744	0.744	0.744

Copies (body, system-level)

Conclusions

- Probability of a message to be SPAM increases as the number of copies of the messages increases
- Since all algorithms improve with more duplicates, a larger collection of participating users in our study would likely have shown more convincing improvements

Conclusions

- Only dealing with duplicate messages limited the overall effectiveness
- One average, the algorithms performed 90% or better as compared to the maximum achievable F-measure
- SPAM is subjective, and a more personalized filter might be a better solution

Future Work

- Try the algorithms on a larger community
- Learning Collaborative Filter – create a long-term database of e-mails
- Collaborative Voting Filter – allow users to classify e-mail via mail reader