Email Mining Toolkit (EMT)

The Email Mining Toolkit (EMT) is a data mining system that computes behavior profiles or models of user email accounts. This toolkit is useful for report generation and summarization of email archives, as well as for detecting email security violations when incorporated with a real-time violation detection system, such as the MET system.

EMT, which includes approximately 13,200 lines of code, is implemented in Java providing a GUI implementing an interface to an underlying relational database application. It provides the means of loading, parsing and analyzing email messages from a wide range of storage formats. It not only demonstrates the statistics of email account behavior, it also computes the volume and velocity of emails exchanged between parties, analyzes specific content and patterns, and explores social relationships between groups of users, and the relative rankings of importance of different individuals in an organization.

Moreover, EMT extends these kinds of analyses to model "user behavior" at a very fine granularity. It models the behavior of individual user email accounts or groups of accounts, and can be used to detect changes in behavior that may be of interest in forensic analyses. These features of EMT provide the means to detect fraudulent misuse and attacks such as viruses and Spam (unwanted) email.

EMT includes 15 different features and models. The statistical models that include stationary and non-stationary user profile are used to generate user behavior models. These models include

- Message Table where individual emails may be automatically classified by built in machine learning subsystems,

- Usage Histogram revealing a user's typical daily email behavior,

- Similar Users which identifies groups of emails users who behave in similar ways ,

- Recipient Frequency providing a detailed analysis of the typical communicants with a user and

- Attachment Statistics detailing attached files serving as a personal file system of a user, as well as the statistical analyses including the birth rate, lifespan, incident rate, prevalence, threat, spread, and death rate useful in identifying interesting attachments and viral attachments.

The analyses built in to EMT concerning groups of accounts and their communication is provided to detect violations of group behavior. These models include

- Enclave Clique groups of users who frequently pairwise exchange messages,
- User Clique the set of accounts a particular user typically emails as a group,
- Email Flow revealing how a single message produces a web of new communication throughout an organization and
- Average Communication Time that views a user's typical response rates to individuals, indicating the relative importance of communicants.
- These models apply algorithms such as Chi Square, Hellinger Distance, Mahalanobis Distance, N-Gram analysis, Naïve Bayes classifier, TF-IDF categorization and graphical cliques analysis. By combining these features, EMT may be applied to a variety of applications and detection tasks.

EMT's graphical user interface provides an easy to use interface to execute these functions and that visualizes results in tabular form with displays of plots and histograms that are easy to understand.

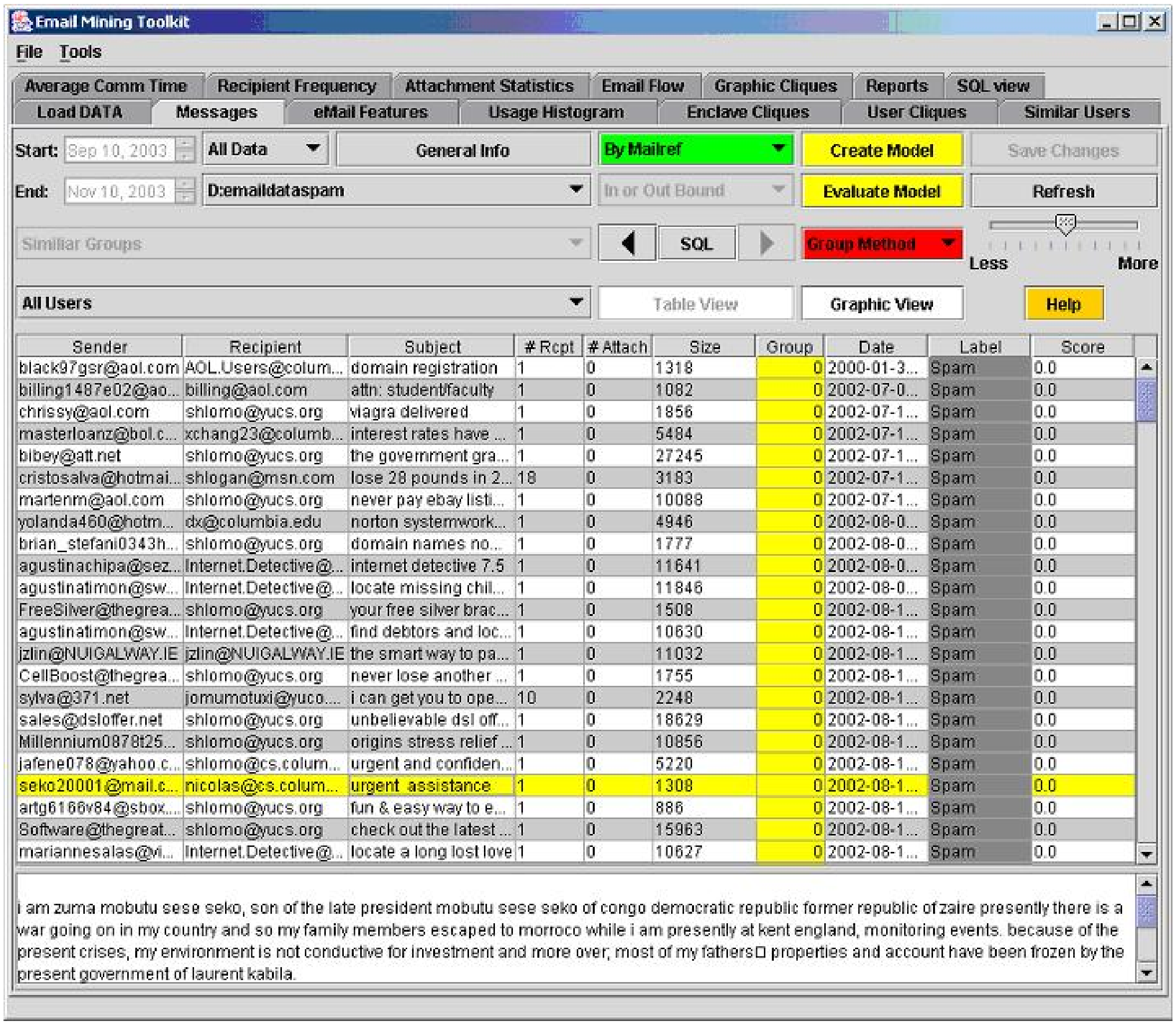
Related publications:

Wei-Jen Li, Shlomo Hershkop, Salvotore J. Stolfo, *Email Archive Analysis Through Graphical Visualization*. ACM CCS VizSEC/DMSEC'04[PDF]

Salvotore J. Stolfo, Wei-Jen Li, Shlomo Hershkop, Ke Wang, Chia-Wei Hu, Olivier Nimeskern, *Detecting Viral Propagations Using Email Behavior Profiles*, *ACM Transactions on Internet Technology (TOIT)*, May 2004. [PDF]

Some EMT screen shots are shown bellow:

General email client window / Machine Learning analysis



Start: Sep 10, 2003 All Data By Mailref

End: Nov 10, 2003 D:emaildataspam In or Out Bound

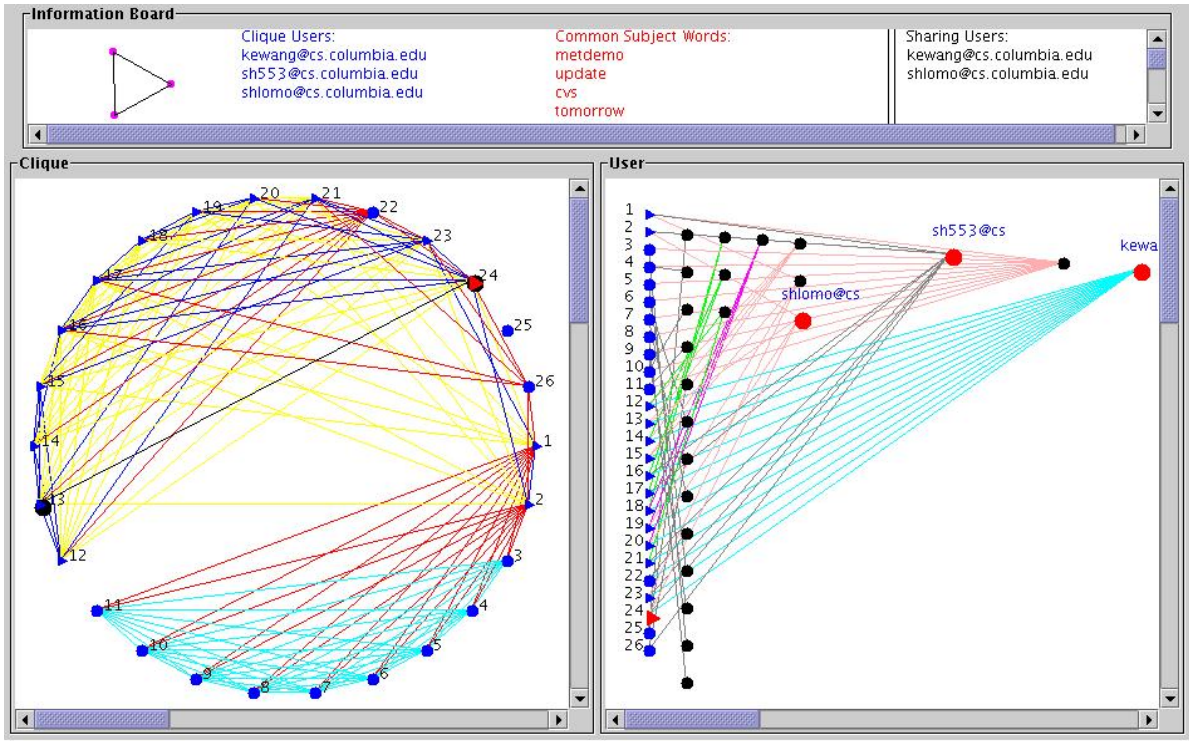
Similiar Groups

All Users

Sender	Recipient	Subject	# Rcpt	# Attach	Size	Group	Date	Label	Score
black97gsr@aol.com	AOL.Users@colum...	domain registration	1	0	1318	0	2000-01-3...	Spam	0.0
billing1487e02@ao...	billing@aol.com	attn: student/faculty	1	0	1082	0	2002-07-0...	Spam	0.0
chrissy@aol.com	shlomo@yucs.org	viagra delivered	1	0	1856	0	2002-07-1...	Spam	0.0
masterloanz@bol.c...	xchang23@columb...	interest rates have ...	1	0	5484	0	2002-07-1...	Spam	0.0
bibey@att.net	shlomo@yucs.org	the government gra...	1	0	27245	0	2002-07-1...	Spam	0.0
cristosalva@hotmai...	shlogan@msn.com	lose 28 pounds in 2...	18	0	3183	0	2002-07-1...	Spam	0.0
martenm@aol.com	shlomo@yucs.org	never pay ebay listi...	1	0	10088	0	2002-07-1...	Spam	0.0
yolanda460@hotm...	dx@columbia.edu	norton systemwork...	1	0	4946	0	2002-08-0...	Spam	0.0
brian_stefani0343h...	shlomo@yucs.org	domain names no...	1	0	1777	0	2002-08-0...	Spam	0.0
agustinachipa@sez...	Internet.Detective@...	internet detective 7.5	1	0	11641	0	2002-08-0...	Spam	0.0
agustinatimon@sw...	Internet.Detective@...	locate missing chil...	1	0	11846	0	2002-08-0...	Spam	0.0
FreeSilver@thegrea...	shlomo@yucs.org	your free silver brac...	1	0	1508	0	2002-08-1...	Spam	0.0
agustinatimon@sw...	Internet.Detective@...	find debtors and loc...	1	0	10630	0	2002-08-1...	Spam	0.0
jzlin@NUIGALWAY.IE	jzlin@NUIGALWAY.IE	the smart way to pa...	1	0	11032	0	2002-08-1...	Spam	0.0
CellBoost@thegrea...	shlomo@yucs.org	never lose another ...	1	0	1755	0	2002-08-1...	Spam	0.0
sylva@371.net	jomumotuxi@yuco....	i can get you to ope...	10	0	2248	0	2002-08-1...	Spam	0.0
sales@dsloffer.net	shlomo@yucs.org	unbelievable dsl off...	1	0	18629	0	2002-08-1...	Spam	0.0
Millennium0878t25...	shlomo@yucs.org	origins stress relief ...	1	0	10856	0	2002-08-1...	Spam	0.0
jafene078@yahoo.c...	shlomo@cs.colum...	urgent and confiden...	1	0	5220	0	2002-08-1...	Spam	0.0
seko20001@mail.c...	nicolas@cs.colum...	urgent assistance	1	0	1308	0	2002-08-1...	Spam	0.0
artg6166v84@sbox....	shlomo@yucs.org	fun & easy way to e...	1	0	886	0	2002-08-1...	Spam	0.0
Software@thegreat...	shlomo@yucs.org	check out the latest ...	1	0	15963	0	2002-08-1...	Spam	0.0
mariannesalas@vi...	Internet.Detective@...	locate a long lost love	1	0	10627	0	2002-08-1...	Spam	0.0

i am zuma mobutu sese seko, son of the late president mobutu sese seko of congo democratic republic former republic of zaire presently there is a war going on in my country and so my family members escaped to morroco while i am presently at kent england, monitoring events. because of the present crises, my environment is not conductive for investment and more over, most of my fathers□ properties and account have been frozen by the present government of laurent kabila.

Graphical clique analysis

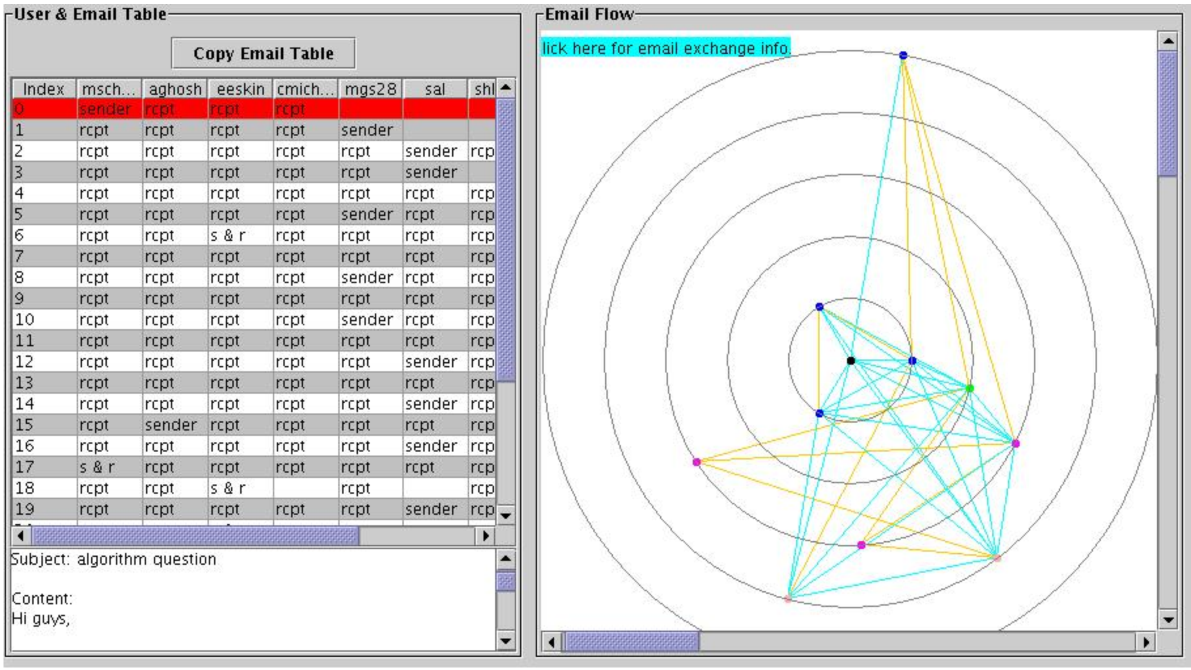


Clique Users:
kewang@cs.columbia.edu
sh553@cs.columbia.edu
shlomo@cs.columbia.edu

Common Subject Words:
metdemo
update
cvs
tomorrow

Sharing Users:
kewang@cs.columbia.edu
shlomo@cs.columbia.edu

Email flow analysis

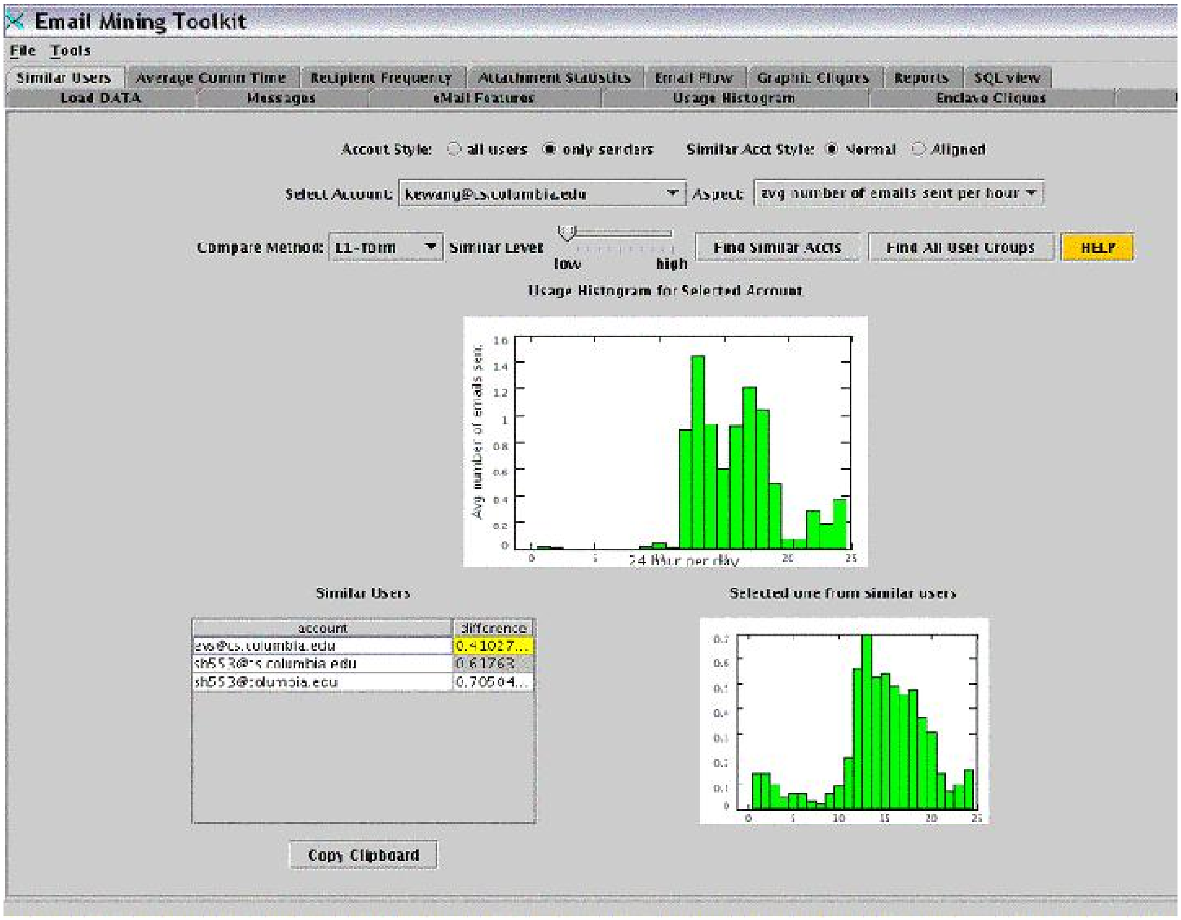


Index	msch...	aghosh	eeskin	cmich...	mgs28	sal	shl
0	sender	rcpt	rcpt	rcpt			
1	rcpt	rcpt	rcpt	rcpt	sender		
2	rcpt	rcpt	rcpt	rcpt	rcpt	sender	rcp
3	rcpt	rcpt	rcpt	rcpt	rcpt	sender	
4	rcpt	rcpt	rcpt	rcpt	rcpt	rcpt	rcp
5	rcpt	rcpt	rcpt	rcpt	sender	rcpt	rcp
6	rcpt	rcpt	s & r	rcpt	rcpt	rcpt	rcp
7	rcpt	rcpt	rcpt	rcpt	rcpt	rcpt	rcp
8	rcpt	rcpt	rcpt	rcpt	sender	rcpt	rcp
9	rcpt	rcpt	rcpt	rcpt	rcpt	rcpt	rcp
10	rcpt	rcpt	rcpt	rcpt	sender	rcpt	rcp
11	rcpt	rcpt	rcpt	rcpt	rcpt	rcpt	rcp
12	rcpt	rcpt	rcpt	rcpt	rcpt	sender	rcp
13	rcpt	rcpt	rcpt	rcpt	rcpt	rcpt	rcp
14	rcpt	rcpt	rcpt	rcpt	rcpt	sender	rcp
15	rcpt	sender	rcpt	rcpt	rcpt	rcpt	rcp
16	rcpt	rcpt	rcpt	rcpt	rcpt	sender	rcp
17	s & r	rcpt	rcpt	rcpt	rcpt	rcpt	rcp
18	rcpt	rcpt	s & r		rcpt		rcp
19	rcpt	rcpt	rcpt	rcpt	rcpt	sender	rcp

Subject: algorithm question

Content:
Hi guys,

Similar Users



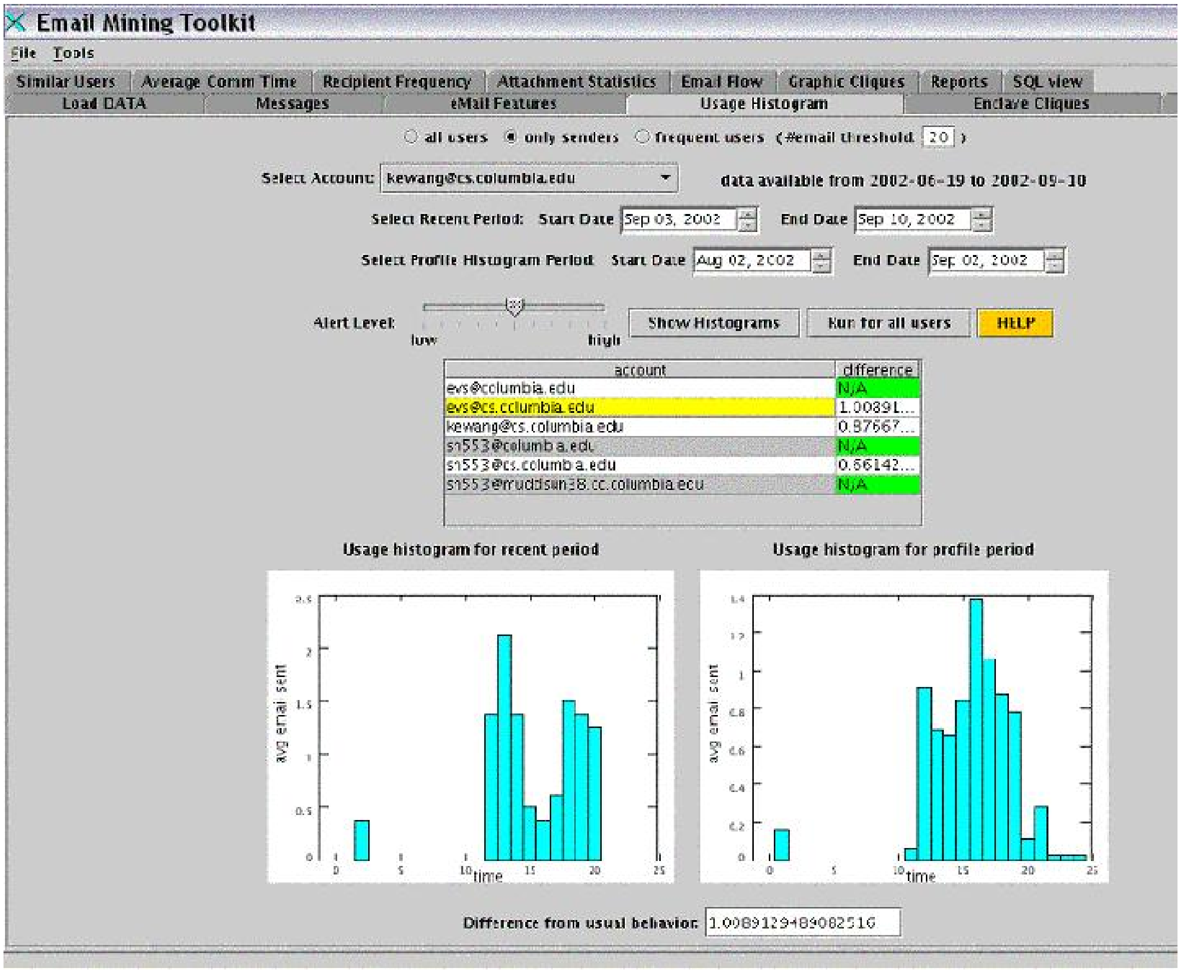
Accout Style: ○ all users ● only senders Similar Acct Style: ● Normal ○ Aligned

Select Account: kewang@cs.columbia.edu Aspect: avg number of emails sent per hour

Compare Method: L1-form Similar Level

account	difference
evs@cs.columbia.edu	0.41027...
sh553@cs.columbia.edu	0.61763
sh553@columbia.edu	0.70504...

Usage Histogram



○ all users ● only senders ○ frequent users (#email threshold 20)

Select Account: kewang@cs.columbia.edu

data available from 2002-06-19 to 2002-09-10

Select Recent Period: Start Date Sep 03, 2002 End Date Sep 10, 2002

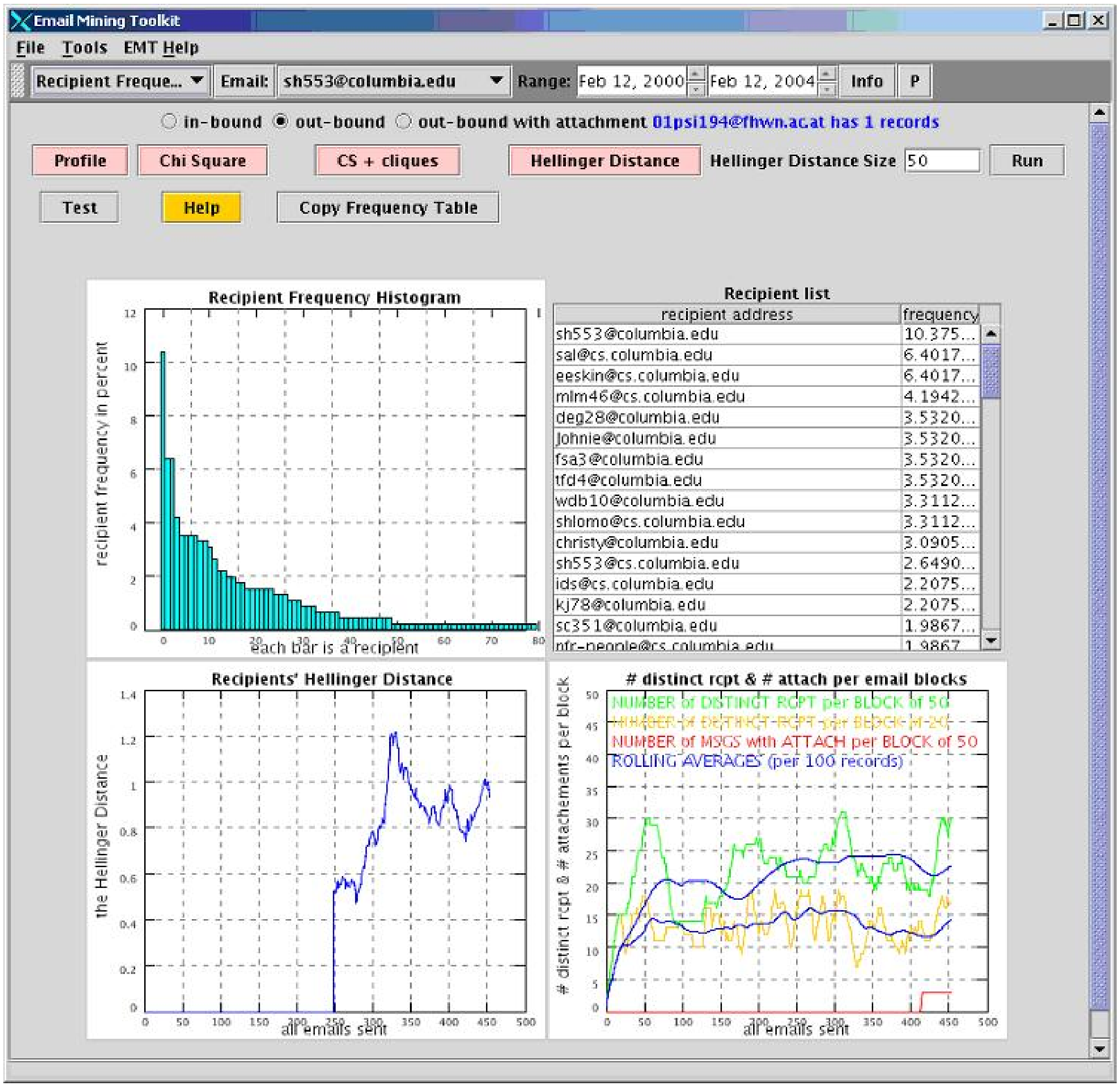
Select Profile Histogram Period Start Date Aug 02, 2002 End Date Sep 02, 2002

Alert Level:

account	difference
evs@columbia.edu	N/A
evs@cs.columbia.edu	1.00891...
kewang@cs.columbia.edu	0.87667...
sn553@columbia.edu	N/A
sn553@cs.columbia.edu	0.66142...
sn553@ruddsun38.cc.columbia.edu	N/A

Difference from usual behavior: 1.008912948908251G

Usage Frequency Analysis


Email Mining Toolkit
File
Tools
EMT Help
Info
P
Profile
Chi Square
CS + cliques
Hellinger Distance
Run
Test
Help
Copy Frequency Table
Recipient Frequency Histogram
recipient frequency in percent
each bar is a recipient
0
2
4
6
8
10
12
0
10
20
30
40
50
60
70
80
Recipients' Hellinger Distance
the Hellinger Distance
all emails sent
0
0.2
0.4
0.6
0.8
1
1.2
1.4
0
50
100
150
200
250
300
350
400
450
500
# distinct rcpt & # attach per email blocks
# distinct rcpt & # attachements per block
all emails sent
NUMBER of DISTINCT RCPT per BLOCK of 50
NUMBER of DISTINCT RCPT per BLOCK of 20
NUMBER of MSGS with ATTACH per BLOCK of 50
ROLLING AVERAGES (per 100 records)
0
5
10
15
20
25
30
35
40
45
50
0
50
100
150
200
250
300
350
400
450
500

Recipient Freque... ▼ Email: sh553@columbia.edu ▼ Range: Feb 12, 2000 Feb 12, 2004

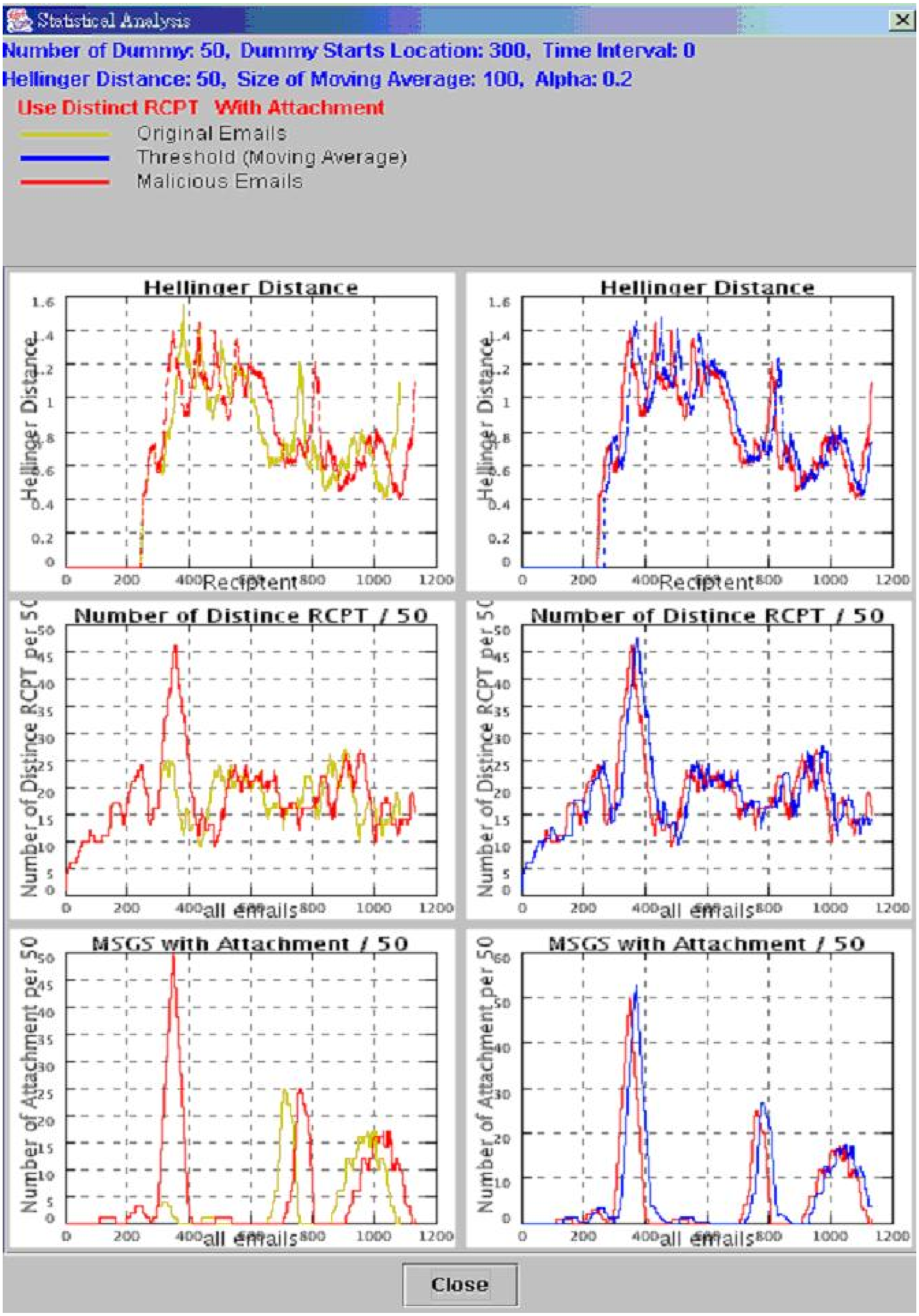
○ in-bound ● out-bound ○ out-bound with attachment 01psi194@fhwn.ac.at has 1 records

Hellinger Distance Size 50

Recipient list

recipient address	frequency
sh553@columbia.edu	10.375...
sal@cs.columbia.edu	6.4017...
eeskin@cs.columbia.edu	6.4017...
mlm46@cs.columbia.edu	4.1942...
deg28@columbia.edu	3.5320...
Johnie@columbia.edu	3.5320...
fsa3@columbia.edu	3.5320...
tfd4@columbia.edu	3.5320...
wdb10@columbia.edu	3.3112...
shlomo@cs.columbia.edu	3.3112...
christy@columbia.edu	3.0905...
sh553@cs.columbia.edu	2.6490...
ids@cs.columbia.edu	2.2075...
kj78@columbia.edu	2.2075...
sc351@columbia.edu	1.9867...
nfr-people@cs.columbia.edu	1.9867...

Virus simulation and detection

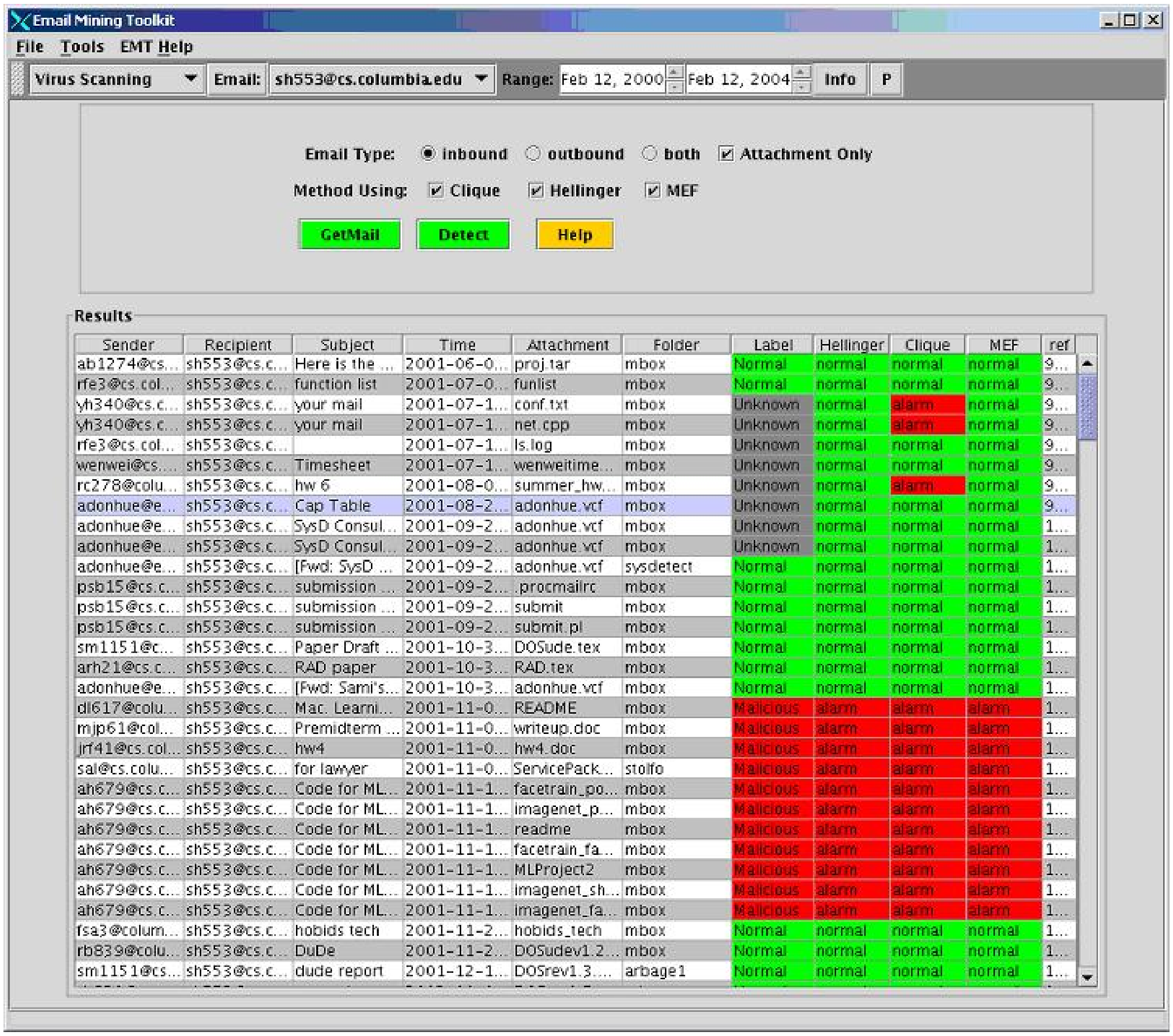


Number of Dummy: 50, Dummy Starts Location: 300, Time Interval: 0

Hellinger Distance: 50, Size of Moving Average: 100, Alpha: 0.2

Use Distinct RCPT With Attachment

Virus detection



Virus Scanning ▼ Email: sh553@cs.columbia.edu ▼ Range: Feb 12, 2000 Feb 12, 2004

Email Type: ● inbound ○ outbound ○ both ☑ Attachment Only

Method Using: ☑ Clique ☑ Hellinger ☑ MEF

Results

Sender	Recipient	Subject	Time	Attachment	Folder	Label	Hellinger	Clique	MEF	ref
ab1274@cs…	sh553@cs.c…	Here is the …	2001-06-0…	proj.tar	mbox	Normal	normal	normal	normal	9…
rfe3@cs.col…	sh553@cs.c…	function list	2001-07-0…	funlist	mbox	Normal	normal	normal	normal	9…
yh340@cs.c…	sh553@cs.c…	your mail	2001-07-1…	conf.txt	mbox	Unknown	normal	alarm	normal	9…
yh340@cs.c…	sh553@cs.c…	your mail	2001-07-1…	net.cpp	mbox	Unknown	normal	alarm	normal	9…
rfe3@cs.col…	sh553@cs.c…		2001-07-1…	ls.log	mbox	Unknown	normal	normal	normal	9…
wenwei@cs.…	sh553@cs.c…	Timesheet	2001-07-1…	wenweitime…	mbox	Unknown	normal	normal	normal	9…
rc278@colu…	sh553@cs.c…	hw 6	2001-08-0…	summer_hw…	mbox	Unknown	normal	alarm	normal	9…
adonhue@e…	sh553@cs.c…	Cap Table	2001-08-2…	adonhue.vcf	mbox	Unknown	normal	normal	normal	9…
adonhue@e…	sh553@cs.c…	SysD Consul…	2001-09-2…	adonhue.vcf	mbox	Unknown	normal	normal	normal	1…
adonhue@e…	sh553@cs.c…	SysD Consul…	2001-09-2…	adonhue.vcf	mbox	Unknown	normal	normal	normal	1…
adonhue@e…	sh553@cs.c…	[Fwd: SysD …	2001-09-2…	adonhue.vcf	sysdetect	Normal	normal	normal	normal	1…
psb15@cs.c…	sh553@cs.c…	submission …	2001-09-2…	.procmailrc	mbox	Normal	normal	normal	normal	1…
psb15@cs.c…	sh553@cs.c…	submission …	2001-09-2…	submit	mbox	Normal	normal	normal	normal	1…
psb15@cs.c…	sh553@cs.c…	submission …	2001-09-2…	submit.pl	mbox	Normal	normal	normal	normal	1…
sm1151@c…	sh553@cs.c…	Paper Draft …	2001-10-3…	DOSude.tex	mbox	Normal	normal	normal	normal	1…
arh21@cs.c…	sh553@cs.c…	RAD paper	2001-10-3…	RAD.tex	mbox	Normal	normal	normal	normal	1…
adonhue@e…	sh553@cs.c…	[Fwd: Sami's…	2001-10-3…	adonhue.vcf	mbox	Normal	normal	normal	normal	1…
dl617@colu…	sh553@cs.c…	Mac. Learni…	2001-11-0…	README	mbox	Malicious	alarm	alarm	alarm	1…
mjp61@col…	sh553@cs.c…	Premidterm …	2001-11-0…	writeup.doc	mbox	Malicious	alarm	alarm	alarm	1…
jrf41@cs.col…	sh553@cs.c…	hw4	2001-11-0…	hw4.doc	mbox	Malicious	alarm	alarm	alarm	1…
sal@cs.colu…	sh553@cs.c…	for lawyer	2001-11-0…	ServicePack…	stolfo	Malicious	alarm	alarm	alarm	1…
ah679@cs.c…	sh553@cs.c…	Code for ML…	2001-11-1…	facetrain_po…	mbox	Malicious	alarm	alarm	alarm	1…
ah679@cs.c…	sh553@cs.c…	Code for ML…	2001-11-1…	imagenet_p…	mbox	Malicious	alarm	alarm	alarm	1…
ah679@cs.c…	sh553@cs.c…	Code for ML…	2001-11-1…	readme	mbox	Malicious	alarm	alarm	alarm	1…
ah679@cs.c…	sh553@cs.c…	Code for ML…	2001-11-1…	facetrain_fa…	mbox	Malicious	alarm	alarm	alarm	1…
ah679@cs.c…	sh553@cs.c…	Code for ML…	2001-11-1…	MLProject2	mbox	Malicious	alarm	alarm	alarm	1…
ah679@cs.c…	sh553@cs.c…	Code for ML…	2001-11-1…	imagenet_sh…	mbox	Malicious	alarm	alarm	alarm	1…
ah679@cs.c…	sh553@cs.c…	Code for ML…	2001-11-1…	imagenet_fa…	mbox	Malicious	alarm	alarm	alarm	1…
fsa3@colum…	sh553@cs.c…	hobids tech	2001-11-2…	hobids_tech	mbox	Normal	normal	normal	normal	1…
rb839@colu…	sh553@cs.c…	DuDe	2001-11-2…	DOSudev1.2…	mbox	Normal	normal	normal	normal	1…
sm1151@cs…	sh553@cs.c…	dude report	2001-12-1…	DOSrev1.3.…	arbage1	Normal	normal	normal	normal	1…