

# THE SPAM FILTERING PLATEAU AT 99.9% ACCURACY AND HOW TO GET PAST IT.\*

WILLIAM S. YERAZUNIS, PHD

MITSUBISHI ELECTRIC RESEARCH LABORATORIES  
( MERL )

CAMBRIDGE, MA

[WSY@MERL.COM](mailto:WSY@MERL.COM)

\* INCLUDES CLARIFICATIONS AND EXAMPLES EXPANDED FROM  
PRESENTATION GIVEN AT THE MIT SPAM CONFERENCE 2004

# SPAM FILTERING STATE OF THE ART

BAYESIAN FILTERS HAVE BECOME “THE WAY”.

THERE ARE MORE THAN A DOZEN AVAILABLE ON  
SOURCEFORGE ALONE.

---

MOZILLA MAIL NOW INCLUDES A BAYESIAN  
OPTION

SPAMASSASSIN HAS AN OPTION TO INCLUDE A  
BAYESIAN “HEURISTIC”.

# THE STATE OF THE ART

## PART II

MOST BAYESIAN FILTERS REPORT ACCURACY ON THE ORDER OF 99% TO 99.9%

BUT NONE OF THE FILTERS REPORT ACCURACY PAST THIS LEVEL. THERE'S NO "GAUSSIAN TAIL".

---

HOW TO GET PAST THIS PLATEAU AT  
99.9% ACCURACY  
IS  
THE ULTIMATE GOAL OF THIS TALK

# STATE OF THE ANTI-ART

SPAMMERS HAVE REACTED TO BAYESIAN FILTERS!

(THAT'S GOOD NEWS- IT MEANS THAT FILTERING HAS MADE A SIGNIFICANT IMPACT AGAINST SPAMMERS. SPAMMERS WOULD NOT HAVE REACTED IF IT WASN'T MAKING A DIFFERENCE.)

- “LONG STORY” SPAM\*
- “DICTIONARY SALAD” SPAM
- JOINING WELL-CREDENTIALLED LISTS
- NEWS STORY SPAM
- HABEAS HAIKU SPAM

\* THE AUTHOR FALLS FOR THESE A LOT.

# THE TYPICAL MODERN SPAM FILTER

- BAYESIAN CLASSIFIER
- TRAINING:
  - TRAIN ON ERRORS (**TOE** STRATEGY)
  - TRAIN EVERY THING (**TEFT** STRATEGY, A.K.A. “BULK” TRAINING)
  - TRAIN UNTIL NO ERRORS (**TUNE** STRATEGY)
- ONLY TOP N FEATURES (OR “PEAKS”) ARE USED FOR CLASSIFICATION

IS THIS OPTIMAL?

# TESTING A SPAM FILTER

- USE THE SPAMASSASSIN TEST SET\*
- 4147 MESSAGES (1400 SPAM, REMAINDER GOOD)
- SHUFFLE TEN TIMES TO FORM 10 “STOCK” RUNS
- RESET LEARNING AFTER EACH STOCK RUN
- EACH METHOD SEES THE SAME 10 STOCK RUNS
- RESERVE FINAL 500 MESSAGES OF EACH RUN AS THE “TEST SET”

\* THIS TEST SET IS A SEVERE TORTURE TEST. THE AUTHOR SCORES LESS THAN 90% ACCURACY ON THIS TEST SET.

# WHAT TRAINING METHOD WORKS BEST?

TRAINING METHOD  
(ALL USING SBPH)

ERROR COUNT  
(LOW IS GOOD)

TEFT (TRAIN EVERY THING)

TOE (TRAIN ONLY ERRORS)

TUNE (TRAIN UNTIL NO ERRORS)\*

# WHAT TRAINING METHOD WORKS BEST?

TRAINING METHOD (ALL USING SBPH)	ERROR COUNT (LOW IS GOOD)
TEFT (TRAIN EVERY THING)	1 4 9
TOE (TRAIN ONLY ERRORS)	6 9
TUNE (TRAIN UNTIL NO ERRORS)*	5 4

\*RESIDUAL ERROR DUE TO CUT OFF IN TRAINING AT 3 OUT OF 41470. BECAUSE **TUNE** REQUIRES KEEPING ALL PRIOR EMAILS AS PART OF THE RETRAINING CORPUS, IT BECOMES INTRACTABLE FOR LARGE INSTALLATIONS.

# WHY IS BULK TRAINING SUBOPTIMAL?

HYPOTHESES\*:

- ADDS EXTRANEOUS FEATURES
- POOR EXAMPLES GET THE SAME WEIGHTING AS GOOD EXAMPLES
- OVERLOADS LIMITED-SIZE DATABASES AND FORCES VALUABLE INFORMATION TO BE FLUSHED (BUT FLUSHING OBSOLETE INFORMATION IS NOT NECESSARILY A BAD THING)

\* NOTE THAT CORRECT PLURAL FORM OF “HYPOTHESIS” IS ACTUALLY PRONOUNCED “CONFUSION”

# IS FORGETTING GOOD?

**YES**

FEATURES IN A CORPUS CAN CHANGE POLARITY.  
FORGETTING OLD DATA ALLOWS THE DATABASE  
TO TRACK EVOLUTION IN SPAM MORE ACCURATELY

# IS FORGETTING GOOD?

YES....

BUT

FORGET AS LITTLE AS POSSIBLE.

DON'T GROOM ALL OF THE HAPAXES OUT AN ENTIRE DATABASE. INSTEAD, RANDOMLY DELETE ONLY A FEW, AND ONLY AS NEEDED TO MAKE SPACE FOR INCOMING FEATURES

THIS YIELDS A  $> 3X$  IMPROVEMENT IN ACCURACY OVER "BLOCK PURGE" OR "HAPAX PURGE" DATABASE CLEANING.

# WHAT EVAL ALGORITHM WORKS BEST?

METHOD (TOE TRAINING)	ERROR COUNT (LOW IS GOOD)
--------------------------	------------------------------

FIRST ORDER BAYESIAN*	
PEAK WINDOW VALUE ONLY (W=5)	
TOKEN SEQUENCE SENSITIVE (W=5)	
TOKEN GRAB BAG (W=5)	
SPARSE BINARY POLYNOMIAL HASH	
MARKOVIAN WITH $2^{2N}$ WEIGHTING	

\* USING ALL FEATURES — NOT “TOP 1000”

# WHAT EVAL ALGORITHM WORKS BEST?

METHOD (TOE TRAINING)	ERROR COUNT (LOW IS GOOD)
FIRST ORDER BAYESIAN	92
PEAK WINDOW VALUE ONLY (W=5)	80
TOKEN SEQUENCE SENSITIVE (W=5)	78
TOKEN GRAB BAG (W=5)	71
SPARSE BINARY POLYNOMIAL HASH	69
MARKOVIAN WITH $2^{2N}$ WEIGHTING (THE WINNER IN ALL SINGLE-PASS TECHNIQUES SO FAR)	56

# HOW GOOD IS MARKOVIAN SPAM FILTERING?

MY CURRENT STATISTICS WITH CRM 1 1 4 USING  
A 2<sup>2N</sup> MARKOVIAN HOVER AROUND 99.9%

4 WEEKS (DEC 15 - JAN 12) RAW SCORES:

TOTAL SPAM	4677
TOTAL NONSPAM	4385
TOTAL MAIL	9062

FALSE ACCEPTS	6
FALSE REJECTS	2
HUMAN CAN'T DECIDE EITHER	3

N+1 ACCURACY	99.90%
--------------	--------

# BUT LAST YEAR YOU HAD 99.91% ACCURACY (N+1). WHAT HAPPENED?

- 1) NEW ERROR SOURCE: PENETRATION OF WELL-CREDENTIALIED LISTS
- 2) NEARLY **TRIPLE** THE RATE OF INCOMING SPAMS:  

LAST YEAR:	1 140 SPAMS
THIS YEAR:	4677 SPAMS*
- 3) MY UPSTREAM STARTED DISCARDING DNSRBL SPAM SO **I LOST A LOT OF LOW-HANGING FRUIT.**

\* UPSTREAM DNSRBL DISCARDS AT ~50% OF ALL MAIL

# WHERE DID THE ERRORS HAPPEN ?

FALSE ACCEPTS  $6 - 2 = 4$

(2 SPAMMERS GOT ONTO PREVIOUSLY  
WELL-CREDENTIALLED LISTS)

FALSE REJECTS  $2 - 2 = 0$

(2 "USERS" VIOLATED RULES ON SAID LISTS  
AND WERE SUMMARILY BOUNCED)

NOTE THAT IT'S ALMOST IMPOSSIBLE TO TELL THE  
DIFFERENCE BETWEEN THE TWO CASES!

**ARGUABLE N+1 ACCURACY** FOR MARKOVIAN  
FILTER:

**99.95%**

# HOW A MARKOVIAN IS DIFFERENT

(1) A MARKOVIAN DISCRIMINATOR TRIES TO MATCH THE INCOMING TEXT AGAINST THE HIDDEN MARKOV MODELS OF THE TWO TEXT CORPI.

(2) WE **DO NOT TRY** TO ACTUALLY CALCULATE THAT HIDDEN MARKOV MODEL (BECAUSE OF TRACTABILITY ISSUES)

(3) THE LONGER A CHAIN WE MATCH (EVEN A CHAIN CONTAINING A FEW ERRORS) THE STRONGER THE EVIDENCE FOR DISCRIMINATION.

# ONE REASON WHY A MARKOVIAN IS BETTER

CONSIDER THE “PERCEPTRON THEOREM”\*

A LINEAR COMBINATIONAL DECISION ALGORITHM  
CAN **NOT** DISCRIMINATE THE CASE:

**A OR B BUT NOT BOTH.**

A CROSS-PRODUCT DECISION ALGORITHM HAS NO  
SUCH LIMITATION.

\* MINSKY AND PAPERT, PERCEPTRONS, 1969

# HANDWAVING MATHEMATICS

IF THE WEIGHTS OF THE MARKOVIAN TERMS ARE SUPERINCREASING (SUCH AS  $2^{2^N}$ ), THEN LONG CORPUS CHAINS CAN OVERRULE SINGLE WORDS AND SHORT CHAINS.

THIS MAKES THE MARKOVIAN FILTER EQUIVALENT TO A CROSS-PRODUCT DECISION ALGORITHM, CAPABLE OF NONLINEAR FILTERING WITHOUT AN INTERMEDIATE LAYER OF COMPUTED METAFEATURES.

# HOW TO TURN A BAYESIAN INTO A MARKOVIAN

(1) CHANGE THE FEATURE GENERATOR FROM SINGLE WORDS TO **SPANNING MULTIPLE WORDS** \*

(2) CHANGE THE WEIGHTING SO THAT **LONGER FEATURES HAVE MORE WEIGHT** (IE. LONGER FEATURES GENERATE LOCAL PROBABILITIES CLOSER TO 0.0 AND 1.0)

(3) THE  **$2^{2N}$  WEIGHTING** MEANS THAT THE WEIGHTS WERE 1, 4, 16, 64, 256, ... FOR SPAN LENGTHS OF 1, 2, 3, 4, 5 ... WORDS

\* ROHAN MALKHARE AT USF HAS A VERY NICE EXTENSION OF THIS TO A STATISTICAL MODEL OF AN ENTIRE MESSAGE..... HE HAS BEEN ADVISED TO PUBLISH AS SOON AS POSSIBLE.

# MARKOVIAN EXAMPLE

GIVEN THE TEXT:

The quick brown fox jumped ....

THE MARKOVIAN FEATURES ARE:

<b>Feature Text</b>	<b>weight</b>
The	1
The quick	4
The <skip> brown	4
The quick brown	16
The <skip> <skip> fox	4
The quick <skip> fox	16
The <skip> brown fox	16
The quick brown fox	64

...AND SO ON

# HOW TO USE THE WEIGHTS

IF YOUR BAYESIAN LOCAL PROBABILITY IS:

$$P_{\text{LOCAL}} = 0.5 + \frac{\text{GOOD} - \text{BAD}}{\text{GOOD} + \text{BAD} + 1}$$

THEN THE EQUIVALENT MARKOVIAN LOCAL PROBABILITY IS:

$$P_{\text{LOCAL}} = 0.5 + \frac{(\text{GOOD} - \text{BAD}) * \text{WEIGHT}}{(\text{GOOD} + \text{BAD} + 1) * \text{WEIGHT}_{\text{MAX}}}$$

BUT EVEN A FULL  
MARKOVIAN IS NOT  
ENOUGH

A MARKOVIAN FILTER MAKES FEWER  
ERRORS THAN A BAYESIAN FILTER

BY ABOUT THE SAME MARGIN AS

A LIGHT BEER HAS FEWER CALORIES  
THAN A REGULAR BEER.

# PREPROCESSING TO HELP FILTERING?

MOST SPAM FILTERS NOW ALSO DO:

- KEY-TOKENIZING (ADDING METAWORDS WHEN A PARTICULAR HEURISTICALLY-DEFINED FEATURE IS FOUND)
- BASE-64 DECONSTRUCTION
- HTML DECOMMENTING AND PARTIAL RENDERING (“EYE-SPACE”\* RATHER THAN E-SPACE)

\* (DARREN LEIGH’S PUN. BLAME HIM)

# CURRENT STATE OF AFFAIRS

WITH ALL OF THESE ASSISTS, THE BEST WE  
HAVE DONE IS 99.95%

WHAT'S THE **NEXT** STEP?

# A FEW POSSIBILITIES FOR THE FUTURE

- AUTHENTICATED SENDERS (??)
- DEFENSE IN DEPTH (MULTIPLE LAYERS OF FILTERING)
- EMAIL INOCULATION
- EMAIL MINEFIELDS
- JUST-IN-TIME FILTERING

# AUTHENTICATED SENDERS

- PLENTY OF BUSINESS MODELS; PLENTY OF COMPETITION FOR STANDARDS
  - PLENTY OF LEGAL ISSUES
    - IF A COMPANY CLAIMS **CAN-SPAM** LEGAL COMPLIANCE, HOW CAN AN AUTHENTICATION AUTHORITY DENY AN AUTHENTICATION TOKEN TO A KNOWN SPAMMER, LET ALONE A “FRONT”?
- LOSS OF INTERNET ANONYMITY ( A SIGNIFICANT LOSS OF INTERNET SOCIAL EQUALITY )
- ABILITY OF CORPORATIONS TO CENSOR UNPOPULAR POINTS OF VIEW WITHOUT OVERSIGHT

# DEFENSE IN DEPTH (MULTIPLE LAYERS)

- AUTOMATIC WHITE/BLACKLIST MAINTENANCE
- AUTOMATIC SENDER AUTHENTICATION
- BAYESIAN/MARKOVIAN LAYER
- AUTOMATIC MICROPAYMENT OR HASH-CASH AS THE FINAL ARBITER.

AN INTEGRATED SYSTEM USING THE ABOVE IS CALLED **CAMRAM**\* AND IS UNDER TEST.

\* RESULTS WILL BE PRESENTED IN ANOTHER PAPER BY ERIC JOHANSSON.

“ONE MAN’S PAIN IS  
ANOTHER MAN’S  
PLEASURE”

—MARQUIS DE SADE

“ONE MAN’S PAIN IS  
ANOTHER MAN’S  
PLEASURE”

—MARQUIS DE SADE

INOCULATION IS A MEANS OF USING THE PAIN  
OF ONE SPAM RECIPIENT TO PROTECT A LARGE  
NUMBER OF OTHER RECIPIENTS.

# INOCULATION BASICS

- INOCULATION IS BASED ON THE OBSERVATION THAT SPAM IS WRITTEN ONCE AND THEN SENT TO MILLIONS OF USERS REPEATEDLY.\*

EVEN A PREVIOUSLY UNSEEN SPAM WILL BE STOPPED BY A FILTER **IF** THE FILTER CAN BE PRE-INOCULATED TO REJECT THE SPAM

\*MODULO \$RANDOM\_STR INSERTION TO FOIL SIMPLE CHECKSUMMING FILTERS

# INOCULATION MECHANICS

- USER A RECEIVES A MIS-FILTERED SPAM
- USER A LABELS THE SPAM AND FORWARDS TO B
- USER B'S MAIL AGENT VERIFIES A AS PRIVILEGED
- USER B'S MAIL FILTER LEARNS THE PARTICULARS OF THIS NEW SPAM
- USER B'S FILTER IS NOW INOCULATED AGAINST THE SPAM.B

# INOCULATION RESULTS

- INOCULATION APPEARS TO HAVE VERY GOOD CHARACTERISTICS, ESPECIALLY AMONG OVERLAPPING CIRCLES OF KNOWN FRIENDS.
- JONATHAN A. ZDZIARSKI AND I ARE PROPOSING AN RFC TO STANDARDIZE THE FORMAT FOR CROSS-PLATFORM FILTER INOCULATIONS

FURTHER DETAILS AND RESULTS WILL BE PRESENTED IN JONATHAN'S TALK.(\*)

\* BLATANT TEASER

# THE EMAIL MINEFIELD

- INOCULATION DEPENDS ON HUMAN INTERVENTION TO RECOGNIZE THE FIRST OCCURRENCE OF EACH AND EVERY SPAM
- **MINEFIELDS** ONLY REQUIRE THE CREATION OF NEW ACCOUNTS THAT ARE PURPOSELY “LEAKED” TO SPAMMERS, AND THEN OPERATE AUTOMATICALLY.
- ANY EMAIL TO SUCH MINEFIELD ACCOUNTS IS KNOWN A PRIORI TO BE SPAM.

# INTEGRATING EMAIL MINEFIELDS

- MINEFIELD ACCOUNTS ARE A GOOD SOURCE FOR AUTOMATIC INOCULATION.
- INOCULATION IS NOT RESTRICTED TO THE TEXT OF A SPAM.

# INTEGRATING EMAIL MINEFIELDS (2)

- CONSIDER THE OTHER INFORMATION AVAILABLE WHEN A MINEFIELD ACCOUNT IS TRIGGERED:

THE IP OF THE CALLER IS KNOWN

AND IT'S NOT SPOOFABLE

- ANY IP ADDRESS OR DOMAIN SENDING TO A MINEFIELD ACCOUNT CAN BE INSTANTLY AND AUTOMATICALLY BLACKHOLED, NOT JUST BY A USER, BUT BY AN ENTIRE SET OF COOPERATING SITES.
- THE BLACKHOLING CAN BE TIME-LIMITED, OR PERMANENT FOR REPEATED SPAMMERS

# MINEFIELD RESULTS

HOW WELL DO EMAIL MINEFIELDS WORK?

- WE DON'T KNOW! WE'RE STILL WORKING THROUGH HOW WELL INOCULATION ITSELF WORKS.

-- BUT WE'LL LET YOU KNOW....

THEORETICALLY\*, ACCURACY SHOULD IMPROVE LINEARLY WITH THE NUMBER OF PEOPLE YOU SHARE INOCULATION DATA WITH (E.G. 10 PEOPLE GIVES YOU 10X ACCURACY)

\* BUT THAT'S ONLY THEORY.

# JUST-IN-TIME FILTERING

CURRENT EMAIL DELIVERY SYSTEMS FILTER UPON ARRIVAL (SO-CALLED “SMTP TIME”).

THIS IS SUBOPTIMAL FOR SYSTEMS WITH INOCULATION OR MINEFIELDING

OBSERVATION- SOME OPTIMIZED SPAMMERS WILL HIT EVERY ACCOUNT ON A SMALL SITE IN LESS THAN TEN SECONDS.

THIS ISN'T ENOUGH TIME TO ALLOW AN INOCULATION TO PROPAGATE

# JUST-IN-TIME FILTERING (2)

IF YOU DON'T HAVE CROSS-SITE HIGH-BANDWIDTH  
MINEFIELDING CONNECTIONS, YOU NEED TO  
**FILTER TWICE:**

- **FIRST FILTER — SMTP TIME - REJECT**  
ANYTHING THAT YOU ARE SURE IS A SPAM.
- **SECOND FILTER — USER-READ TIME - WHEN A**  
USER ACTUALLY IS PULLING EMAIL FROM THE  
SPOOL, FILTER AGAIN.

THIS DELAY ALLOWS THE GREATEST POSSIBLE  
TIME WINDOW FOR INOCULATIONS AND MINEFIELD  
MESSAGES TO ARRIVE.

# CONCLUSIONS:

- BAYESIANS ARE VERY GOOD
- MARKOVIANS ARE EVEN BETTER
- NEITHER BY ITSELF IS SUFFICIENT
- JUST AS BAYESIANS/MARKOVIANS USE ALL INFORMATION AVAILABLE -PER USER-, INOCULATION, MINEFIELDING, AND JUST-IN-TIME FILTERING GAIN INFORMATION (AND ACCURACY) BY LOOKING ACROSS AN ENTIRE SITE OR ACROSS MULTIPLE SITES.

# UNPROVEN HYPOTHESIS

BAYESIAN/MARKOVIAN WITH 100'S OF USERS  
SHARING INOCULATIONS, WITH MINEFIELDS, AND  
WITH JUST-IN-TIME FILTERING COULD  
REASONABLY GET TO FIVE-NINES (99.999%)  
ACCURACY, AND POSSIBLY APPROACH  
99.9999% (ONE ERROR PER MILLION EMAILS)  
ACCURACY.

# THANK YOU ALL!

ARE THERE ANY QUESTIONS?

HANDY WEB SITES:

[HTTP://WWW.CAMRAM.ORG](http://WWW.CAMRAM.ORG)

[HTTP://WWW.PAULGRAHAM.COM](http://WWW.PAULGRAHAM.COM)

[HTTP://CRM114.SOURCEFORGE.NET](http://CRM114.SOURCEFORGE.NET)

SUMMER (THAT MEANS IT'S WARM) SPAM  
CONFERENCE:

[HTTP://WWW.CEAS.CC](http://WWW.CEAS.CC)