On Attacking Statistical Spam Filters

Greg Wittel & S. Felix Wu

U.C. Davis

CEAS 2004

Outline

- Introduction
- Attack Classes
- Testing A New Attack
- Conclusions & Future

Attack Classes

- Attempted attack methods:
  - Tokenization
    - Works against feature selection by splitting or modifying key message features
    - e.g. Splitting up words with spaces, HTML tricks
  - Obfuscation
    - Use encoding or misdirection to hide contents from filter
    - e.g. HTML/URL encoding, letter substitution

Attack Classes cont.

- Weak Statistical
  - Skew message statistics by adding in random data
  - e.g. Add in random words, fake HTML tags, random text excerpts
- Strong Statistical
  - Differentiated from ‘weak’ attacks by using more intelligence in the attack
  - Guessing v. educated guessing
  - e.g. Graham-Cumming Attack

Attack Classes cont.

- Misc:
  - Sparse Data attack
  - Hash breaking attacks

Testing A New Attack

- Tested two types of attacks:
  - Dictionary word attack (old)
  - Common word attack (new)
- Both attacks add *n* random words to a base message.
- Tested against two filters:
  - CRM114 - Sparse binary poly. + Naïve Bayesian
  - SpamBayes (SB) - Naïve bayesian

Procedure

- Training data
  - 3000 hams from SpamAssassin corpus
  - 3000 spams from SpamArchive-mod corpus
  - CRM114 trained on errors
  - SB using bulk training

Procedure cont.

- Test data
  - Started with a base ‘picospam’ not in training data:

```
From: Kelsey Stone <bouhooh@entitlement.com>
To: submit@spamarchive.org
Subject: Erase hidden Spies or Trojan Horses from your computer

Erase E-Spyware from your computer

http://boozofoof.spywiper.biz
```

Procedure cont.

- Test data cont.
  - Base picospam is detectable by filters
  - Generated 1000 variations with *n* words added.
    - Words selected with and without replacement
    - n = 10, 25, 50, 100, 200, 300, 400
  - Recorded classifications, effect on score

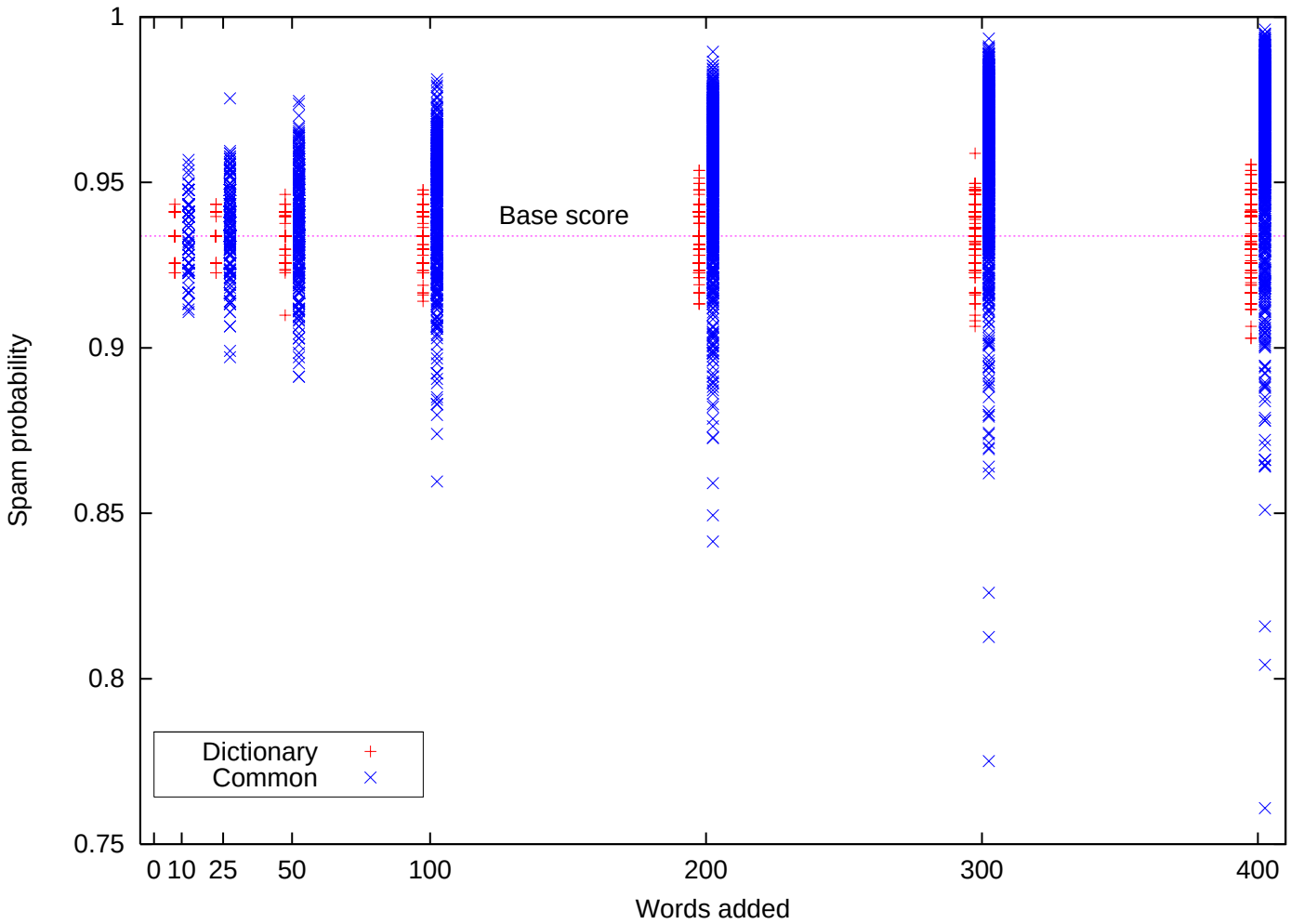
Results

- Using 10,000 variants didn't effect results
- Selection with/without replacement had no effect
- Mixed results

CRM114 Results

- Both attacks failed; 0 false negatives
- Spam score *was* effected…

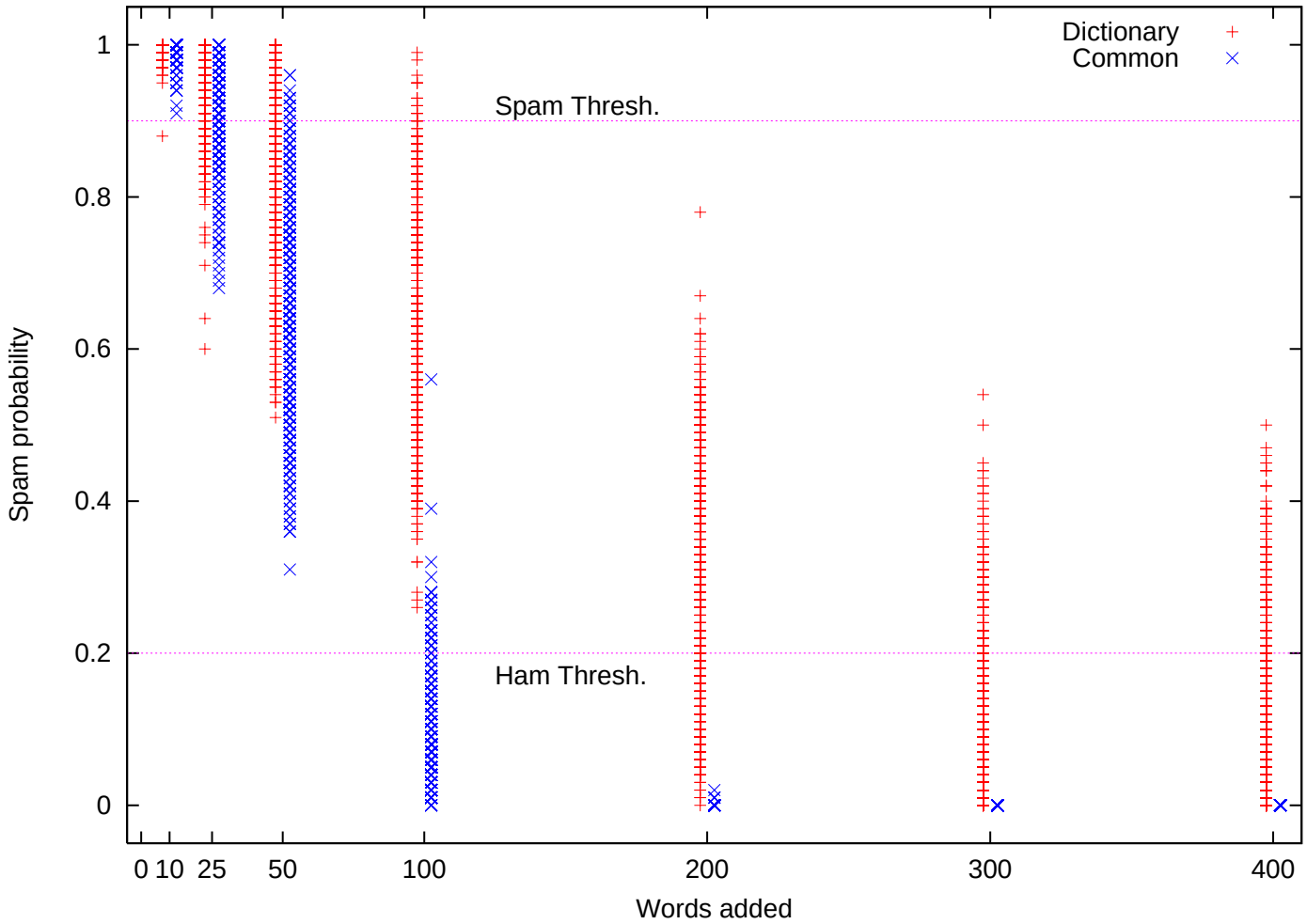
CRM114 Results cont.



SpamBayes Results

- Baseline Dictionary attack: mild success
- Common word attack…

SpamBayes Results cont.



SpamBayes Results cont.

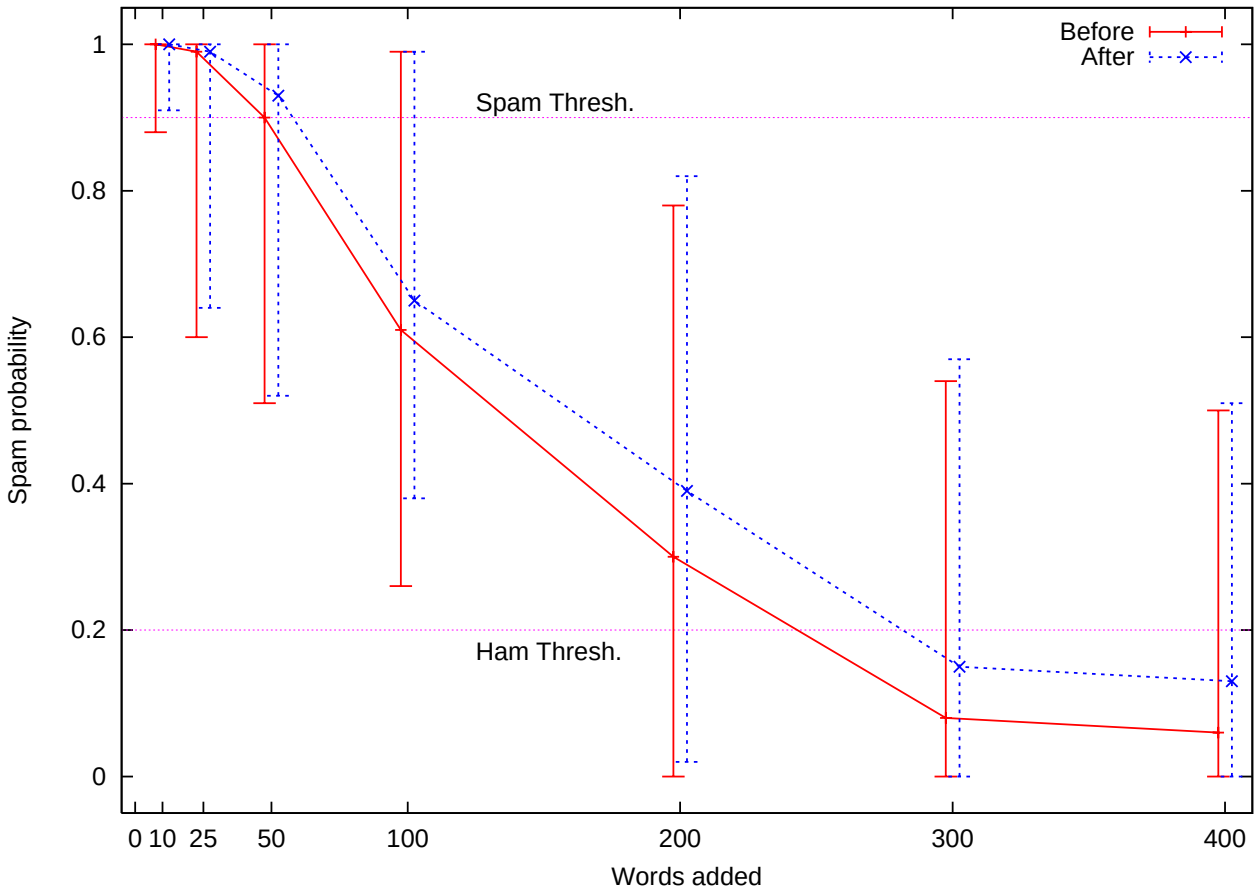
- Common word attack reduces attack size by up to 4x
- What Happened? Why such poor performance on either attack?
- Hypothesis: Basis picospam was not in training data.
- Added the basis spam to SB's training data…

SpamBayes Results Part 2

- Retrained filter offered greater resistance to ‘weak’ dictionary attack.
- Small performance gain against common word attack.
- Gains not big enough to resist attack

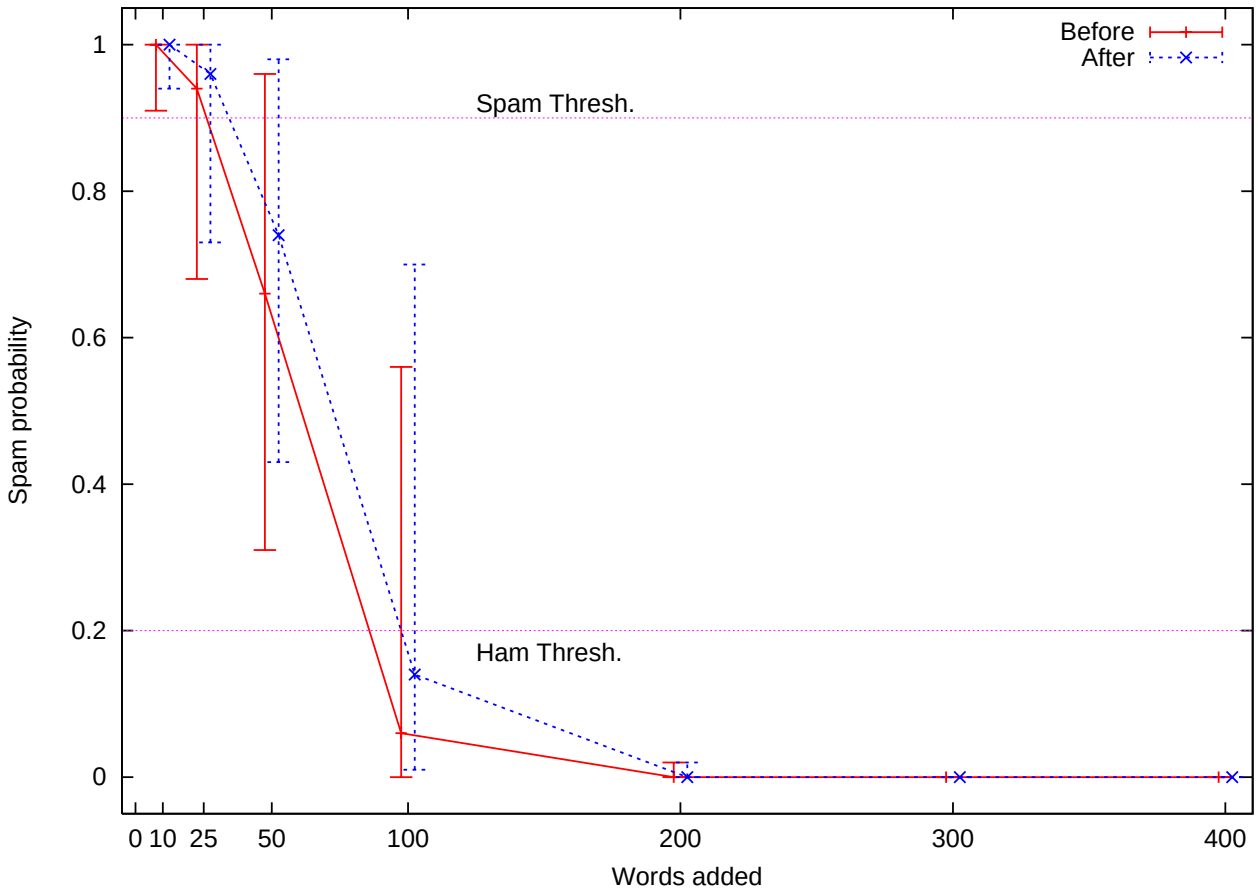
SpamBayes Results Part 2 cont.

Dictionary Word Attack



SpamBayes Results Part 2 cont.

Common Word Attack



Conclusion & Future…

- Mixed success of common word attack shows need for further study
- Other filters
  - Bogofilter shows similar vulnerability
- Effect of re-training on attack msgs v.
  - False negative, false positive rate
- Testing other basis picospams

Future cont.

- What makes a filter hard to distract?
- Relevance of independence assumption
- More advanced attacks
  - Natural language generation
- Traditional software flaws
  - Exploitable buffer overflows
  - Remote code execution

Colophon

- Contact information:
  - Greg Wittel (wittel at cs . ucdavis . edu)
  - S. Felix Wu (wu at cs . ucdavis . edu)
- Questions?