Learning to Detect Phishing Emails

Ian Fette
Norman Sadeh
Anthony Tomasic
(School of CS, CMU)

Presented by: Ashique Mahmood

Dept of Computer & Information Sciences
University of Delaware

Key Terms

- ***Learning*** (= Machine Learning)
- ***Classifier, training data, testing data, model*** etc.
- ***False positive, False negative***
- ***Phishing attacks***

  Trying to direct web users to spoofed websites that steal information such as credit card, Identity info, SSN, passwords etc.

  Most popular way to "phish" is E-mail.

Key Terms (contd.)

- *Phishing attacks*

An Example:

"

We Recently Upgraded Our Security System with a Newly Established SSL Sever In which Guarantees your maximum Security Protection when Accessing Your Webmail account Online.

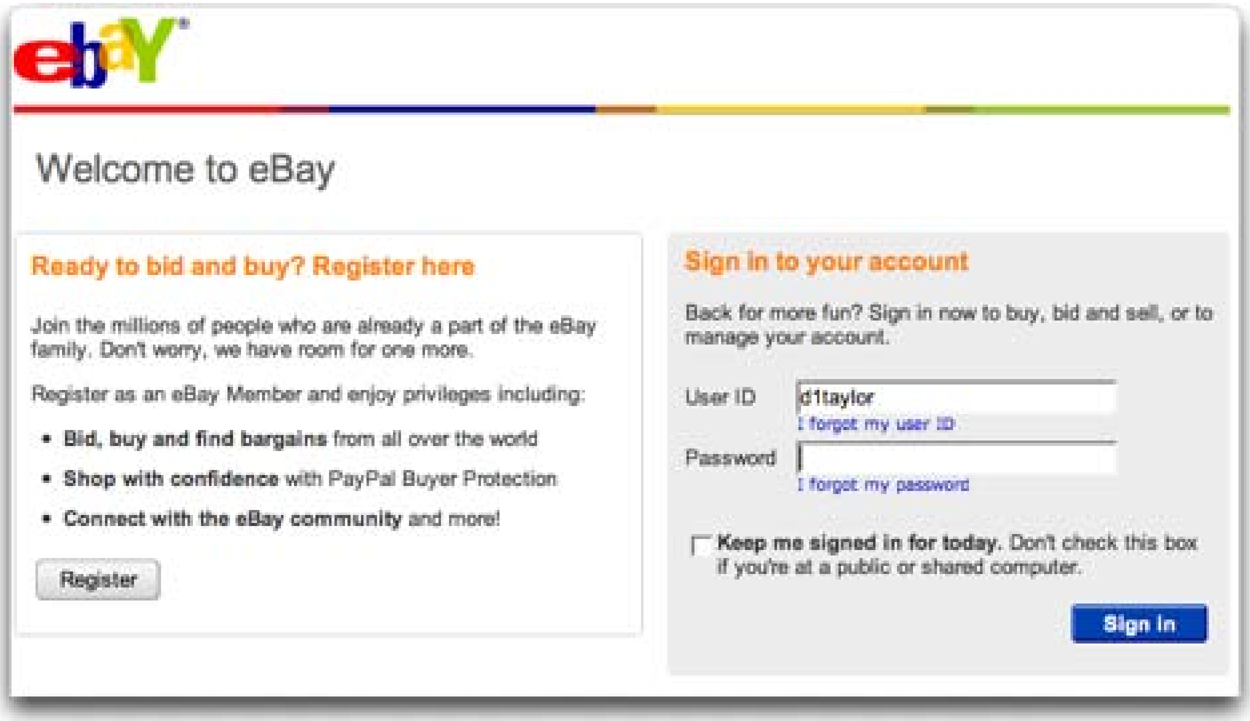
<u>Click here to Upgrade</u>

Regards,
University of Delaware Security Department
"

(March 17, 2010)

Key Terms (contd.)

- *Phishing attacks*



Early attempts

- ***Toolbars***

  Integrated to browsers, prompt user with warning. Can have up to 85% of success.

  - Disadvantage:
    - Less contextual information
    - Users may dismiss or misinterpret warning
    - Loss of productivity

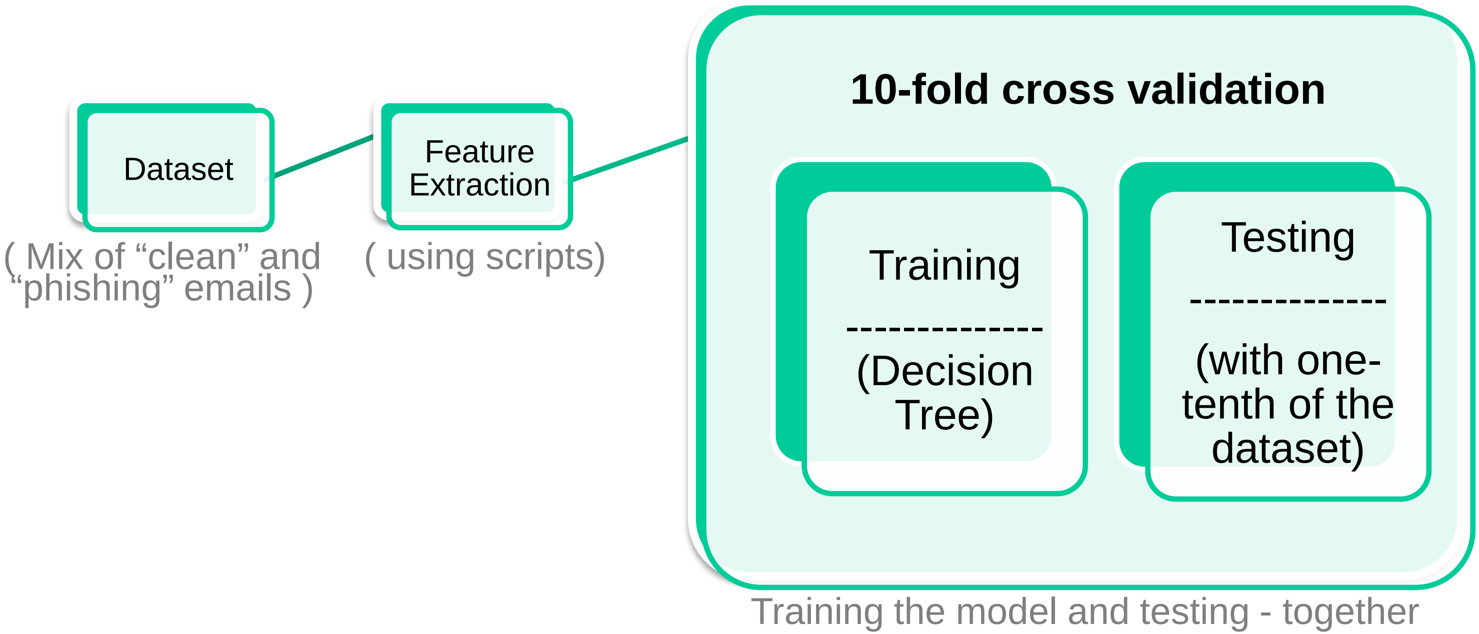
Spam Detection vs Phishing detection

- Why phishing detection is different from spam detection?
- Spam Detection -
  - focuses on the structure/subject of the email.
  - looks at the vocabulary of the email, suspicious words.
  - Blacklisted senders.
- Phishing emails look like legitimate.

Motivation

- Phishing emails and websites are ***identical*** to legitimate ones; hence difficult to detect.
- **Spam filters** are not good for phishing detection.
- **Toolbar based detection** not effective and sufficient.
- So, we need more sophisticated filters for phishing detection, prohibiting phishing emails reaching to inbox.

Overall approach (PILFER)



10-fold Cross-validation :

The dataset is divided into 10 distinct parts. Each part is *Tested* using the other 9 parts as *training* data.

Dataset

- Two publicly available datasets:
  - **The Ham Corpora**
    (SpamAssassin project)
    6950 non-phishing, non-spam “ham” emails
  - **Phishingcorpus**
    approx. 860 “phishing” emails.

Features

- Binary features:
  - Is it an IP-Based URL?

    Ex: http://192.168.0.1/ebay.cgi?fix_account

  - Age of linked-to domain names

    WHOIS query, to detect for how long the domain was active

  - Non-matching URLs

    paypal.com

  - "here" links to non-modal domain

    Non-modal : not the most frequently linked domain

Features(cont'd)

- Binary features:
  - HTML emails?

    MIME type text/html indicates possible phishing attack
  - Contains javascript?

    does the string "javascript" appears in the email?
  - Spam-filter output

    Output from stand-alone spam-filters is also a feature, which indicates "ham" or "spam".

    (SpamAssassin is used for PILFER)

Features(cont'd)

- Continuous features:
  - No. of links

    No. of links in HTML part, defined as <a> tag
  - No. of domains

    Count of how many distinct domains are present in the email, starting with http:// or https://
  - No. of dots in URL

    Maximum no. of dots contained in any of the links.

    http://www.my-bank.update.data.com
    http://www.google.com/url?q=http://www.badsite.com

SpamAssassin

- SpamAssassin
  - Widely used, freely-available spam filter
  - Highly accurate in classifying spams
- SpamAssassin also tested, both
  - Trained
  - Untrained
- SpamAssassin compared with PILFER.

Results

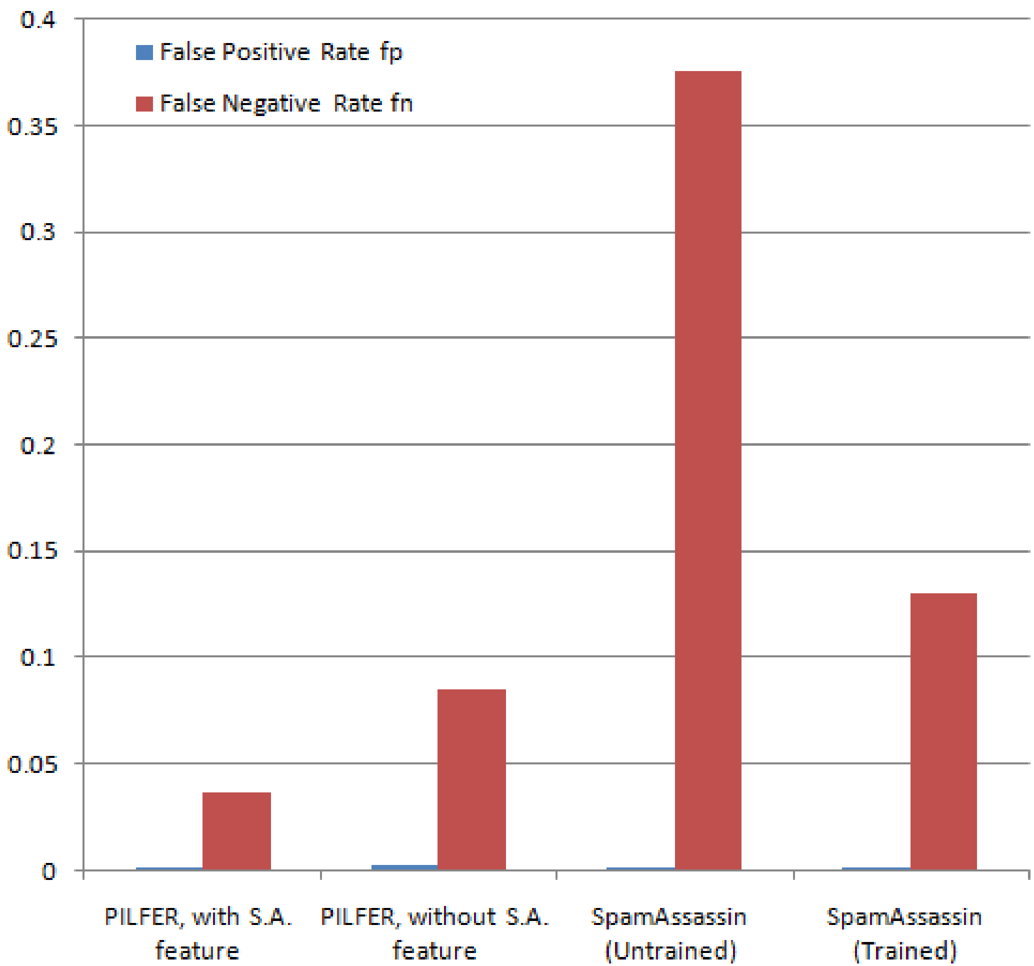
- PILFER
  - Overall accuracy of 99.5%
  - False positive rate, fp= 0.0013 (approx.)
  - False negative rate, fn= 0.035 (approx.)

Results (cont'd)

Table 1: Accuracy of classifier compared with baseline spam filter

Classifier	False Positive Rate fp	False Negative Rate fn
PILFER, with S.A. feature	0.0013	0.036
PILFER, without S.A. feature	0.0022	0.085
SpamAssassin (Untrained)	0.0014	0.376
SpamAssassin (Trained)	0.0012	0.130

Results (cont'd)



Results (cont'd)

Table 2: Percentage of emails matching the binary features

Feature	Non-Phishing Matched	Phishing Matched
Has IP link	0.06%	45.04%
Has "fresh" link	0.98%	12.49%
Has "nonmatching" URL	0.14%	50.64%
Has non-modal here link	0.82%	18.20%
Is HTML email	5.55%	93.47%
Contains JavaScript	2.30%	10.15%
SpamAssassin Output	0.12%	87.05%

Results (cont'd)

Table 3: Mean, standard deviation of the continuous features, per-class

Feature	μ_{phishing}	σ_{phishing}	$\mu_{\text{non-phishing}}$	$\sigma_{\text{non-phishing}}$
Number of links	3.87	4.97	2.36	12.00
Number of domains	1.49	1.42	0.43	3.32
Number of dots	3.78	1.94	0.19	0.87

Conclusion

- PILFER is exhibits almost accurate results, because it exploits few unique features that spam detectors don't use.

- Phishing detection along with spam detection provides best results.

- Future direction:
  - Phishing techniques evolve over time very quickly, so continuous research expected.

That's all, folks!

Questions ???

That's all, folks!

Thank you.

Tiny Appendix

- False positive rate,

$$fp = \frac{ham_{phish}}{ham_{phish} + ham_{ham}}$$

- False negative rate,

$$fn = \frac{phish_{ham}}{phish_{ham} + phish_{phish}}$$