NLP Techniques for Security Challenges

Rakesh M. Verma

ReDAS Laboratory

Computer Science Dept.

University of Houston

ReDAS Mission: Cutting-edge research, education in Reasoning, Data Analytics and Security, and address current practical problems in security

NLP for Security

NLP techniques have been applied to:

- **Password Security (lots of work: Bloom Filters, Markov Models, several papers in every CCS etc.)**
- **Spam Detection (e.g., Naïve Bayes or its variants, and words as features)**
- **Malware (e.g., N-grams as seen in ML for Security module)**
- **Phishing - we discuss this in detail**
- **Attack Generation – we touch upon this briefly**

Password Security

- Second order Markov Model
- 28 states (26 for letters - no case distinction, one for space and one for all other characters)
- Database D of known bad passwords
- Transition probability T[i, j, k] = f(i, j, k)/f(i, j, infinity), Max likelihood estimate, f(i, j, k) - frequency of trigram ijk, f(i, j, infinity) – total trigrams beginning with ith and jth character

[BApasswd: A New Proactive Password Checker, 1993 – Best Paper Award]

BApasswd (contd.)

- Whether password is bad reduces to what is the likelihood that it is generated by the Markov Model
- Use log-likelihood function: for password p, llf(p) = sum of log of transition probabilities of p's trigrams
- Normalize: llf(p)/(number of trigrams of p)
- Transform to mean 0 and std dev. 1
- Apply a threshold. All passwords above threshold are bad. Calculated as -2.6

Phish Or Not is The Question

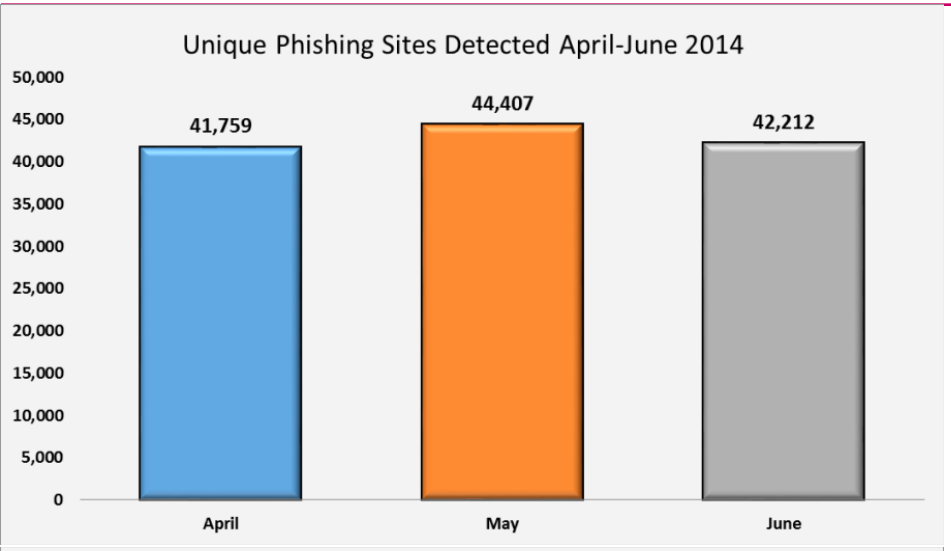
Sampling of the Email Deception Attacks Research - sources

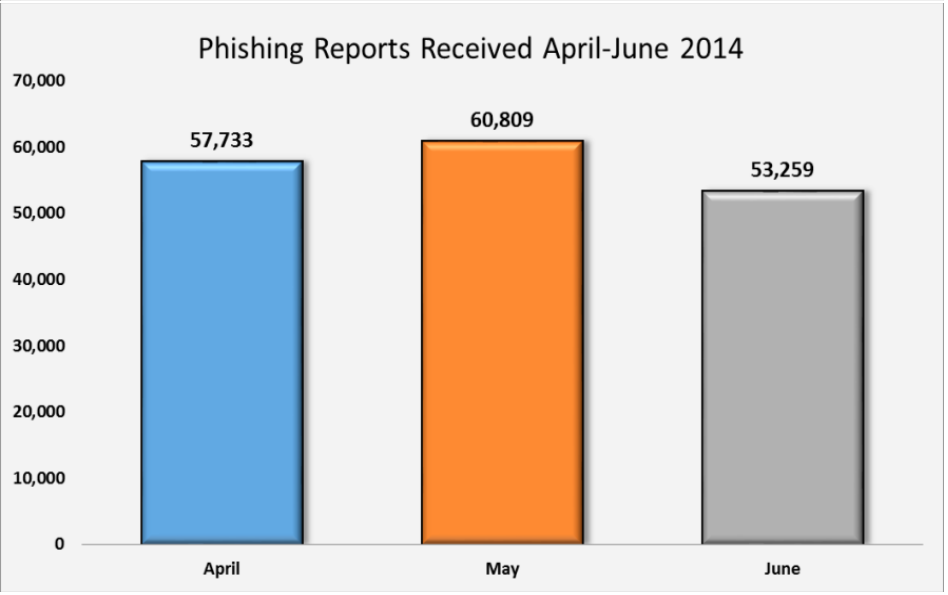
- Phishing email detection [ESORICS 2012, ICISC 2013, SECRYPT 2015, CODASPY 2017]
- Phishing URL detection [CODASPY 2015, IWSPA 2017]
- Phishing web site detection [ICISS 2014]
- Email masquerade attacks [ASIA CCS 2017]

Phishing

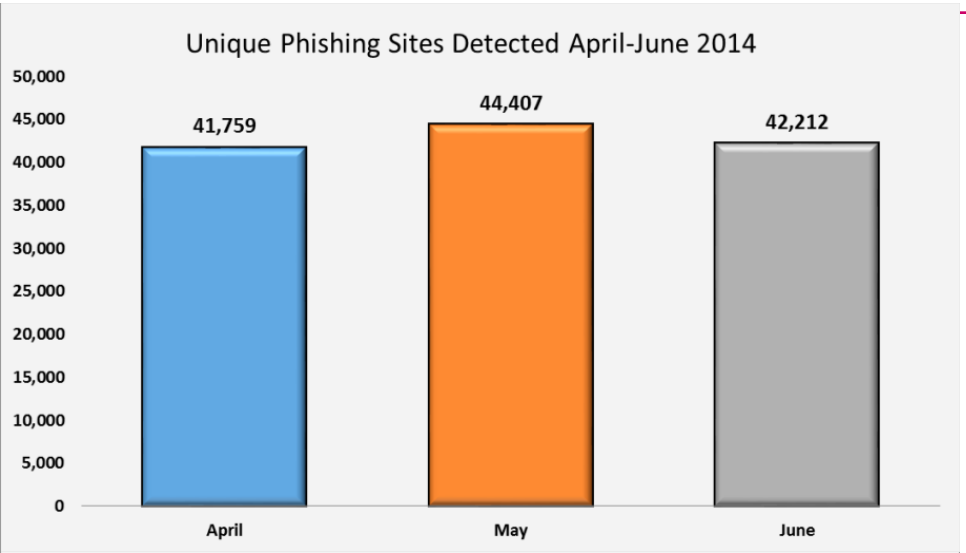
- An acute and persistent problem
- Can be used for stealing user's digital identity or to install malware through attachments (e.g. INR 10 million stolen via Trojan, attack on multiple banks)
- Estimates of losses range from a few billion dollars to tens of billions per year

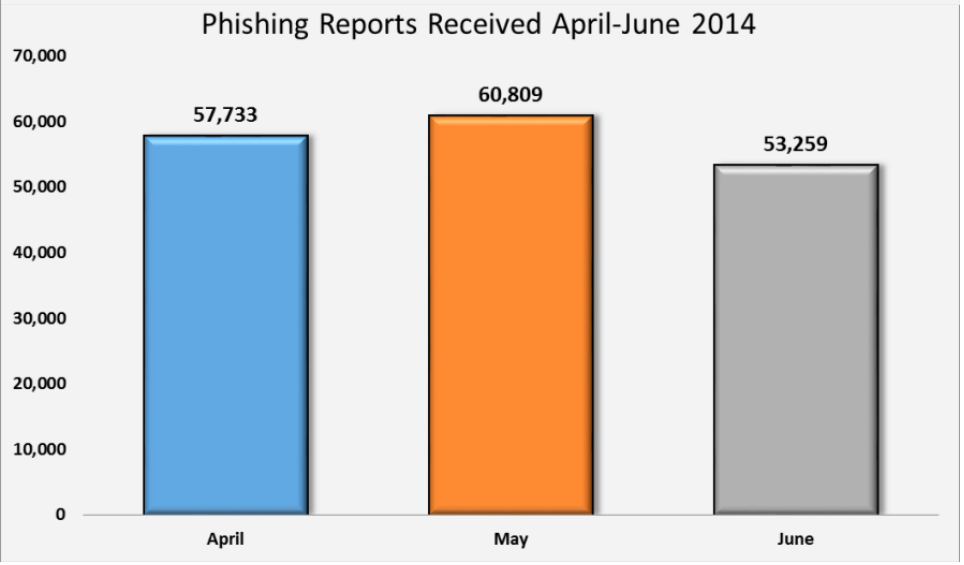
Magnitude of the Menace





- Q2 2014 data obtained from Anti Phishing Working Group (APWG)
- Q2 2016 data is worse!





- Global estimated losses for 2014 = $3.2 Billion
  - Infragard report
- Other losses include: lost time and productivity
- Q2 2014 data from Anti Phishing Working Group (APWG)

Phishing Modus Operandi

- Email is by far the most popular medium for phishing
- Others include:
  - Web forums
  - Internet chat rooms
  - Phone conversations
  - …

I. Phishing Email Detection

- July 2011-Aug 2012: 115 phishing emails passed unmarked through my spam filter. Approx. 9/month.

Date: Tue, 13 Sep 2011 09:09:52 -0600
From: XYZ <abc@sw1.k12.wy.us>
To: undisclosed-recipients: ;
Subject: Mail Box Quota Exceeded

Your web mail quota has exceeded the set quota which is 3GB. you are currently running on 3.9 GB.
To re-activate and increase your web mail quota please click the link below.
<CLICK HERE>

Failure to do so may result in the cancellation of your web mail account.
Thanks, and sorry for the inconvenience
Local-host.

Quick Victory Achievable?

"It is *non-trivial* to distinguish phishing messages from legitimate messages, since phishing messages are constructed to resemble legitimate messages as much as possible."

[Irani, Webb, Giffin, Pu – 2008]

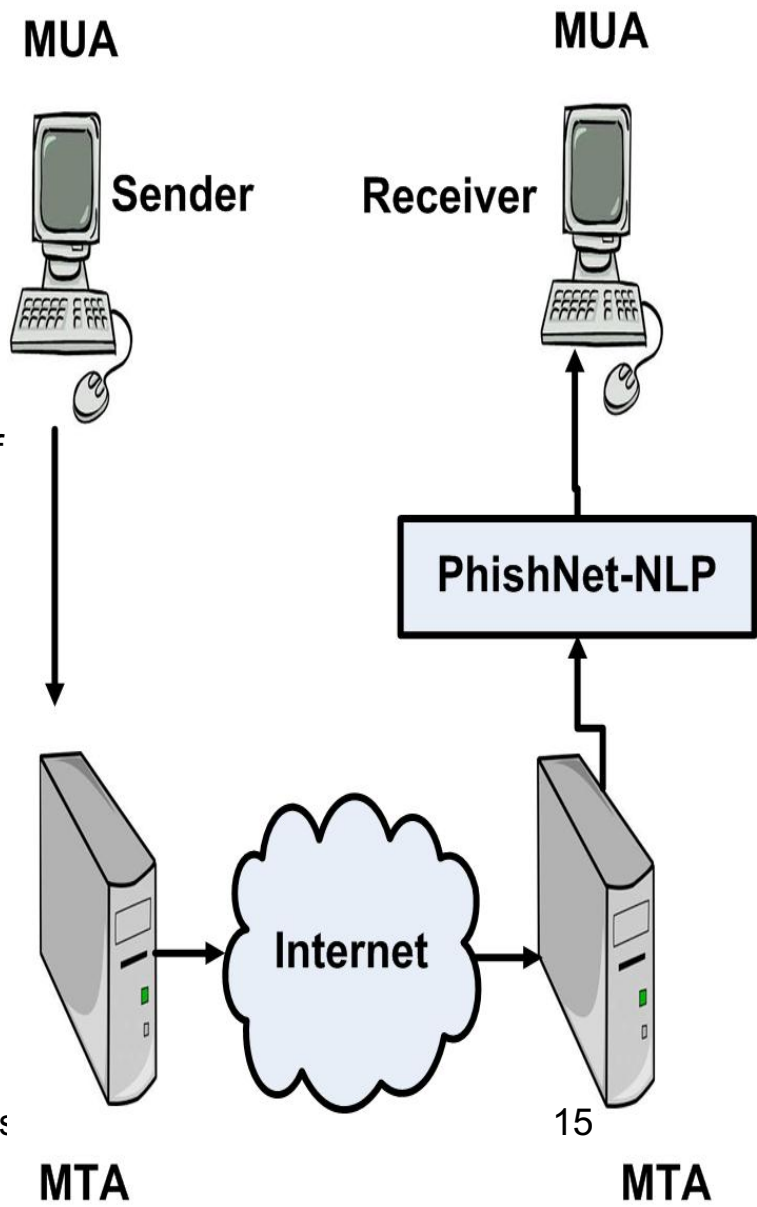
Outline of Phishing Email Detection

- Phishnet-NLP [ESORICS 2012]
- Improving the body text classifier [ICISC 2013]
- Message-Id field [SECRYPT 2015]
- Datasets
- Results and Comparison

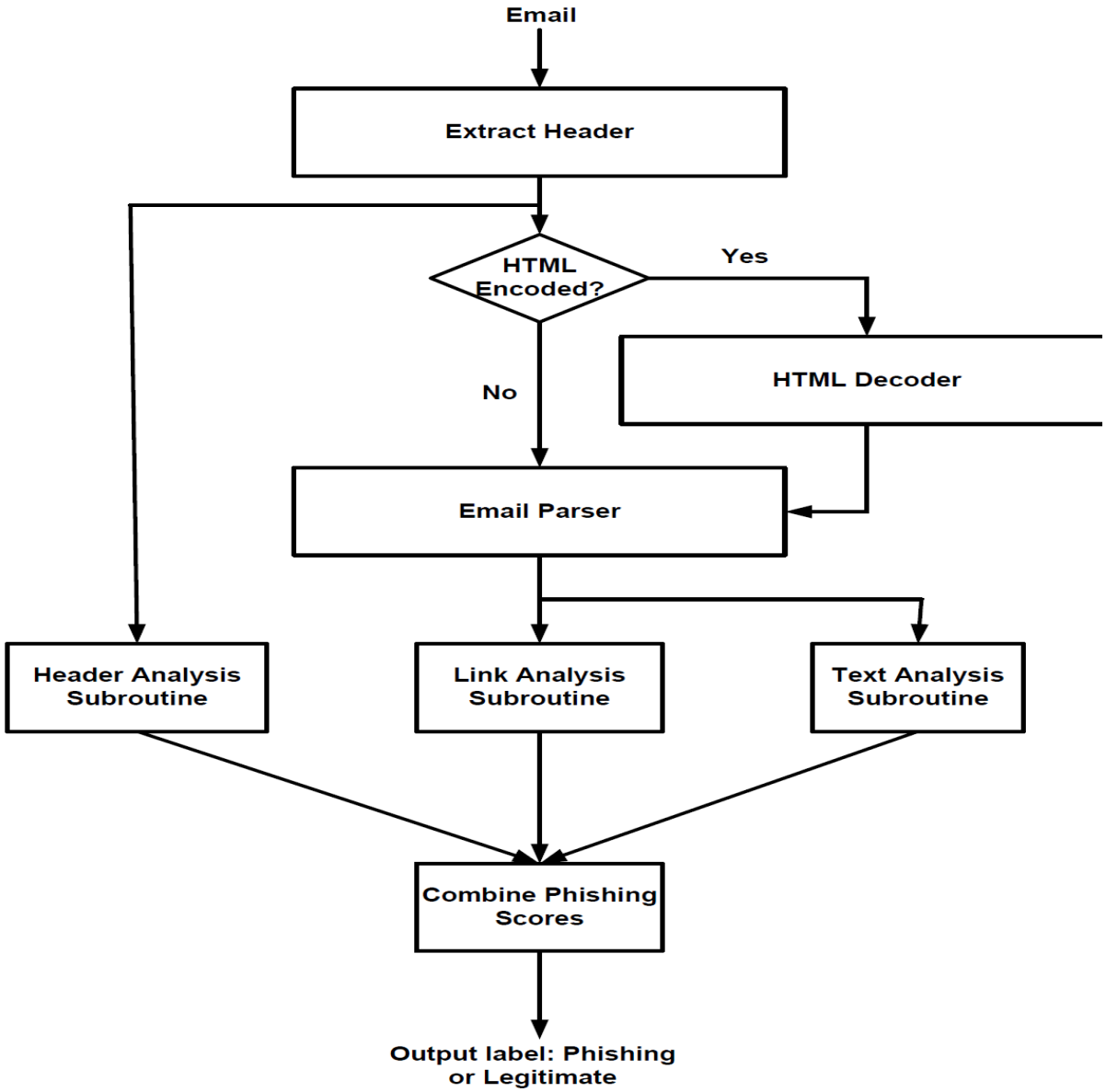
The Principle Behind Our Methods

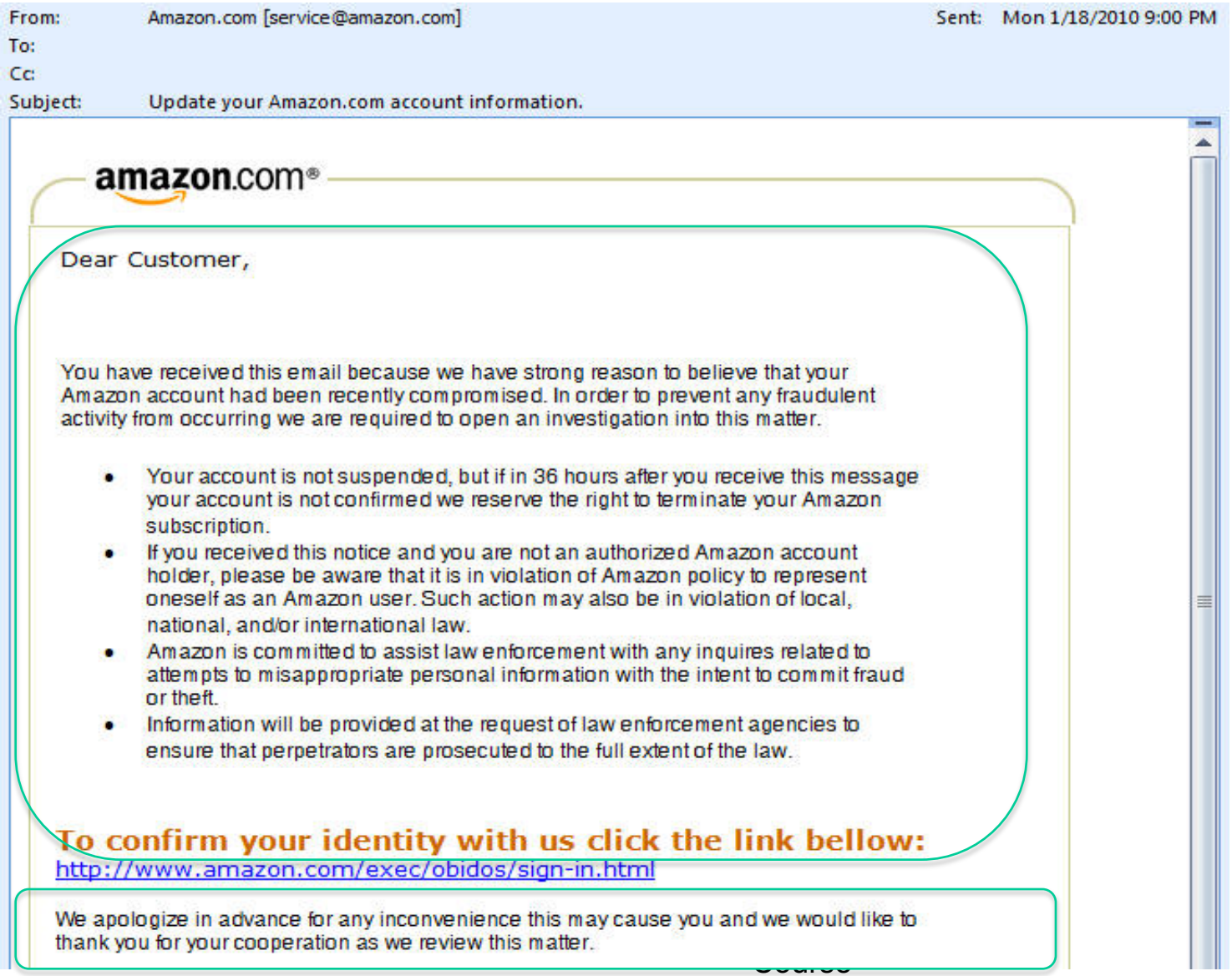
- Focus on key differences between a phishing and a legitimate email
  - Phishing email requires user to take some action
  - Phishing email conveys a sense of urgency, or threat, or deadline, or some financial incentive for maximum reward in a short period of time

- **PhishNet-NLP** [ESORICS 2012]
  - Comprehensive approach based on three boolean classifiers:
    - Text Analysis
    - Header Analysis
    - Link Analysis
  - Combines results from each classifier to decide if email is phishing
  - Analyzes emails before reaching mailbox to prevent attack by malware
  - Use contextual information of links for efficiency
  - No training on or annotation of emails
- **Dataset**
  - 4,550 phishing emails (available online)
  - 1,000 legitimate emails (authors' mailboxes)



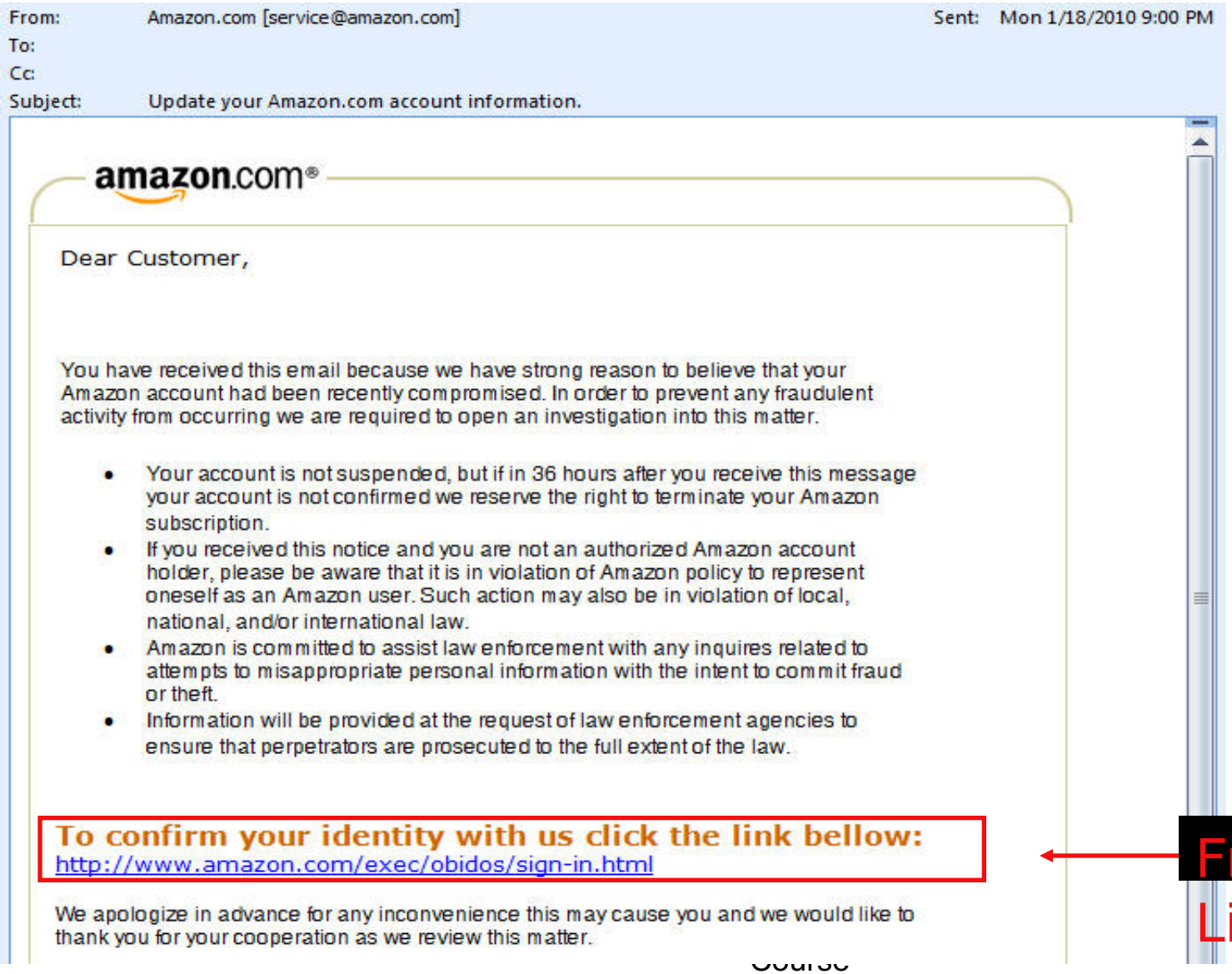
PhishNet-NLP Flowchart





Example

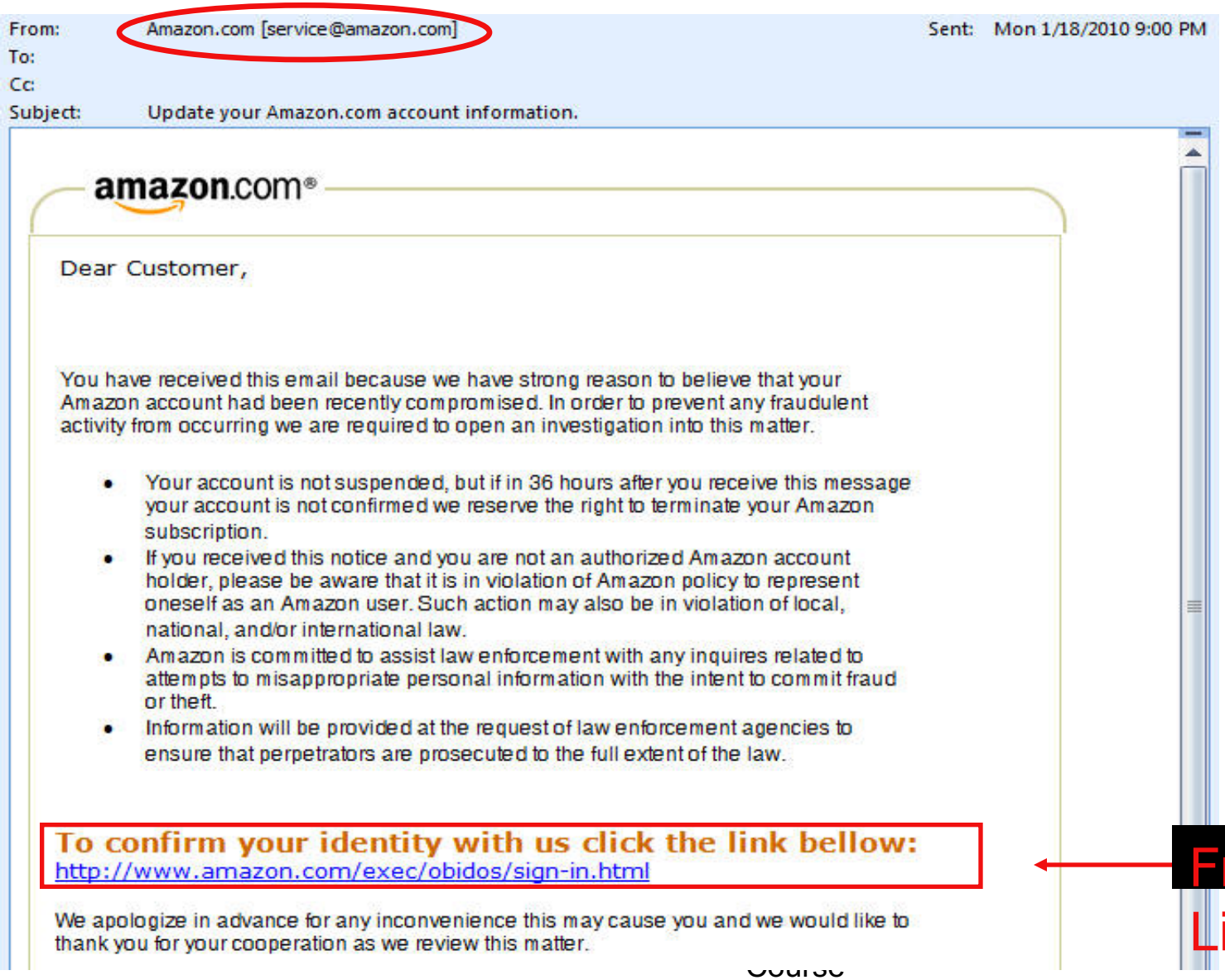
Phishing Email



Example

Phishing Email

Fraudulent Link



Example

Phishing Email

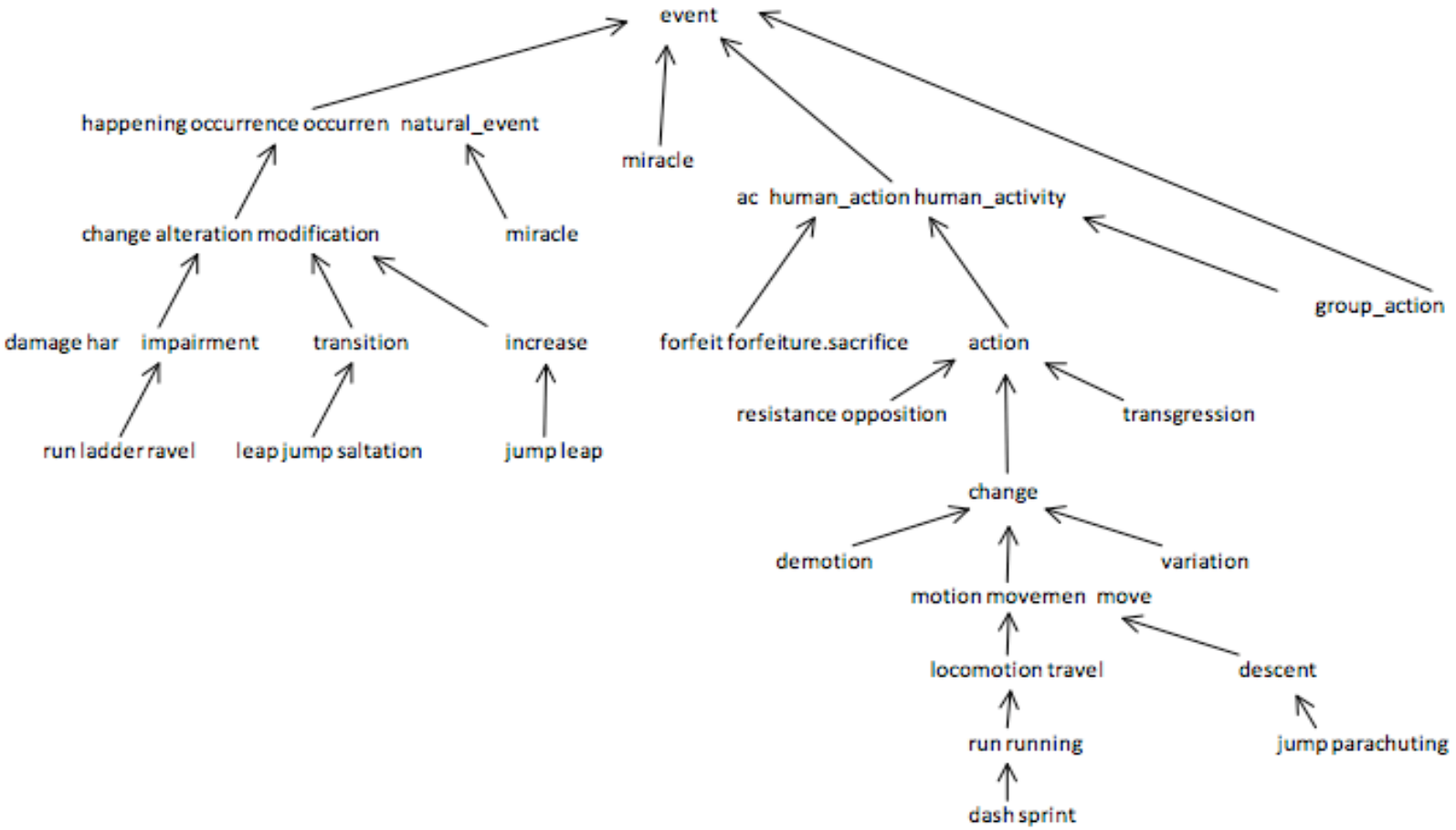
Fraudulent Link

NLP Background

- NLP techniques used
  1. Named-entity extraction (proper noun phrases that denote: person, place, organization, date, or money)

     Bob bought a book at Amazon for 20 dollars.
  1. Part-of-speech (POS) tagging
     Alice/NNP emailed/VBD the/DT certificate/NN.
  2. Word-sense disambiguation for polysemous words Alice *gets (gets/VB/#17)* it. The question got (got/VBD/#33) me.
  3. Stop word removal (removal of function words such as a, an, the, etc.)
  4. WordNet (lexical database of English, needs word + POS + sense)

WordNet Subgraph



Phishnet-NLP : Text Analysis

- Extracts text from email
- Uses NLP Techniques
  1. Part-of-speech tagging
  2. Word-sense disambiguation for polysemous verbs

     (Example: John *gets* it, The child *got* scared, Bob got a speeding ticket)

  3. Stemming
  4. WordNet (needs part-of-speech, stem and sense)
- Scores certain *verbs*, takes maximum score and compares with threshold (set to 1)
  - Score increased with link, urgency, or incentive in same sentence

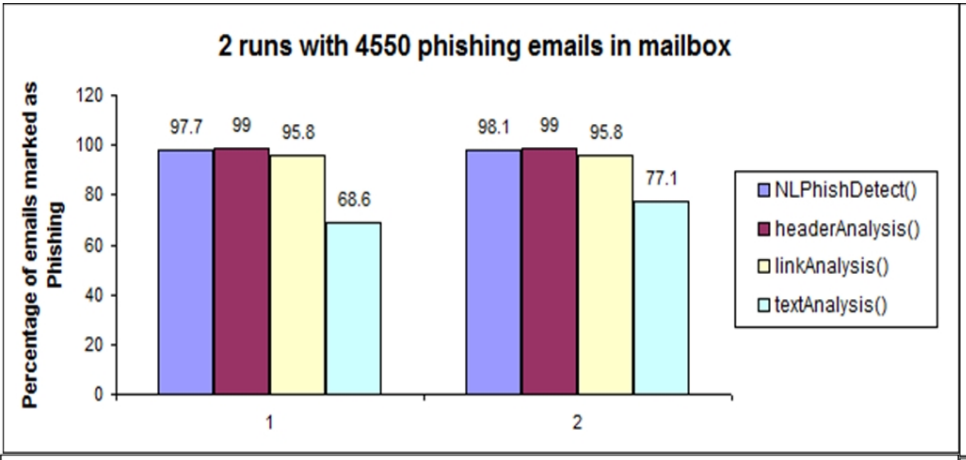
- Semantics
  - Uses *hyponymy* relation on verbs

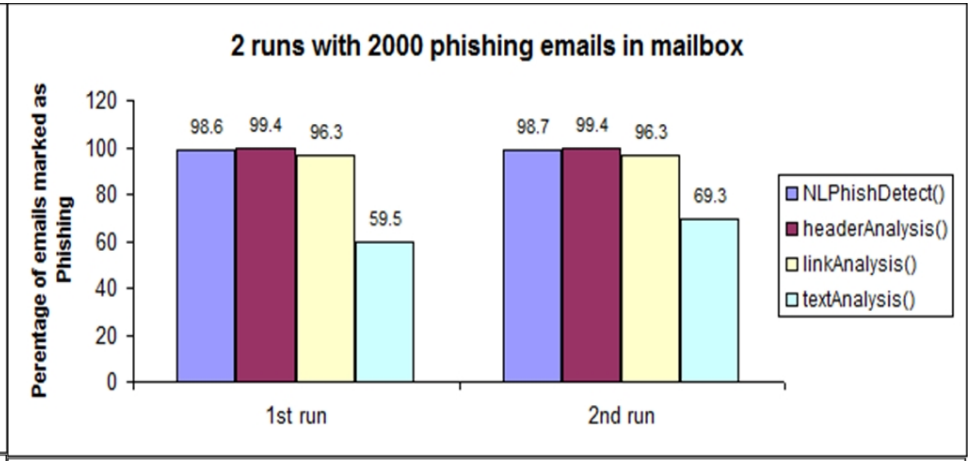
  (Example: verb click is a hyponym of verb move)
- Uses context (user's sent/recd. mail) when available
  - Increases robustness provided phisher does not have access to context
  - Increases detection
- Email scored for similarity and assigned a context-score
- Text score and Context-score combined logically

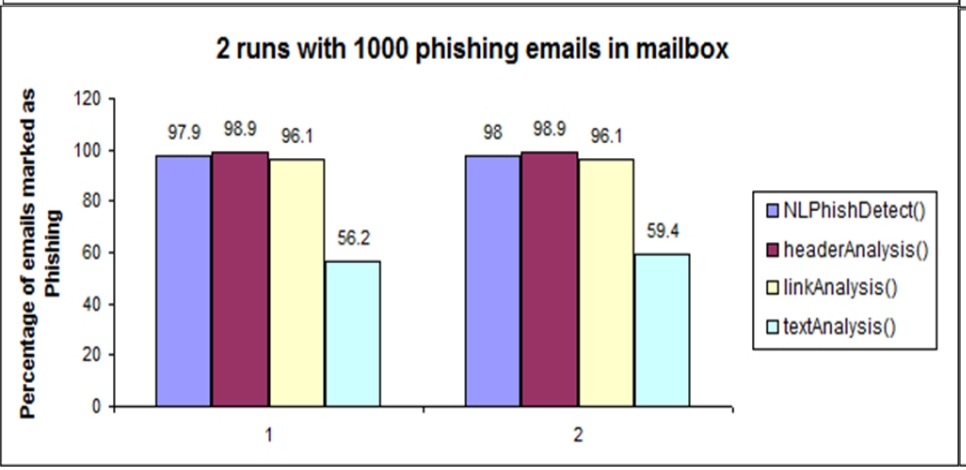
Phishnet-NLP : Context Score

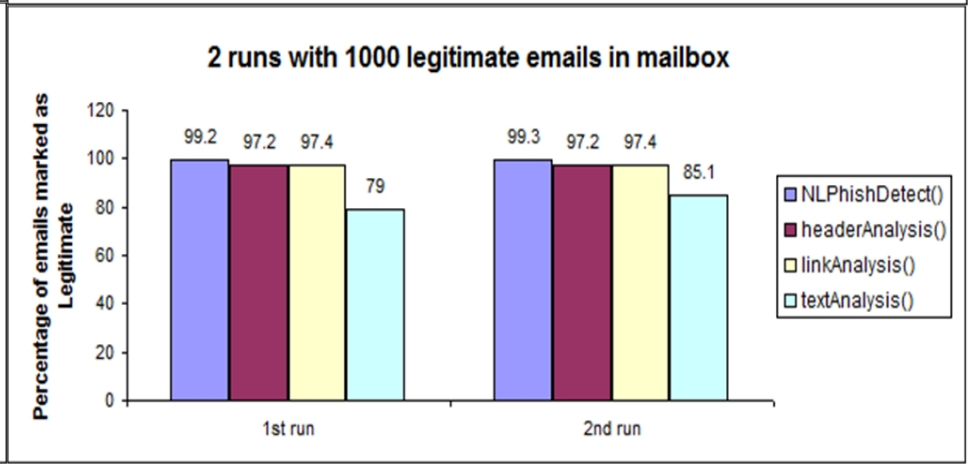
- Email converted to vector using Information Retrieval techniques
  - **TF-IDF**: Term Frequency-Inverse Document Frequency
  - **TF**: No. of occurrences of a word within a document
  - **IDF**: measure of how infrequently the word appears in other documents in the database
- Similarity score: Cosine of the angle between vectors
- Thresholding

Phishnet-NLP: Results









Improving Text Analysis: Goals

- We investigate 2 basic questions:
  - Can statistical NLP techniques be used effectively for email text analysis?
  - Do NLP techniques such as part-of-speech tagging and semantics help improve the statistical methods, and if so, by how much?

Text Analysis Comparison

- Extracts text from email
- Uses NLP techniques
- Intuitive approach: Scored certain *verbs*, took maximum score and compared with a threshold
  - Score increased with link, urgency, or incentive in same sentence
- Statistical approach: New, more accurate text-based classifiers based on semantic feature selection with t-test

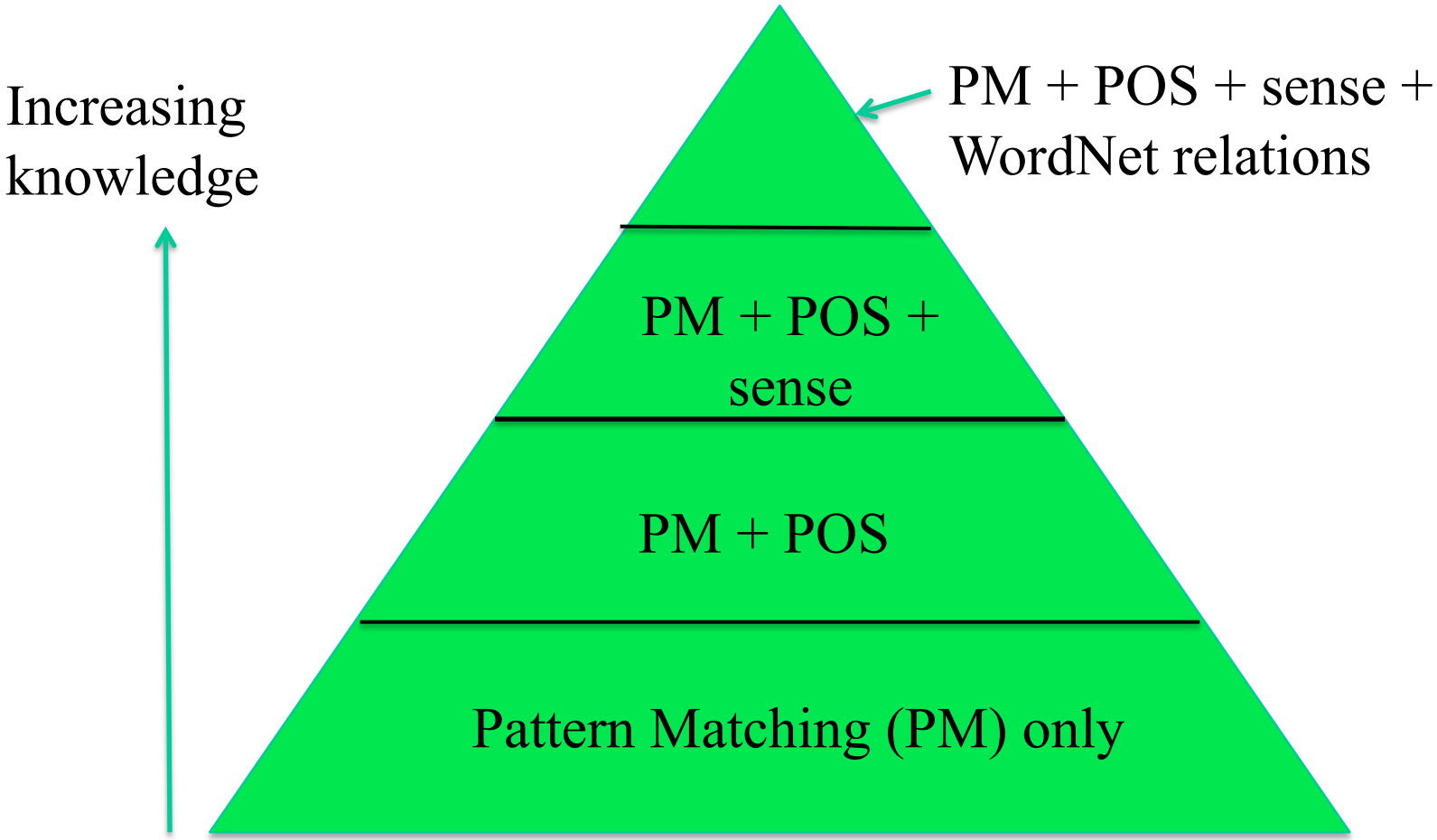
Additional Contributions

- Semantic feature selection for text data
- Careful study of feature selection with increasing "knowledge"
- Testing on bigger public datasets with more variety
- Detailed comparison of statistical versus intuitive approach on same phishing dataset

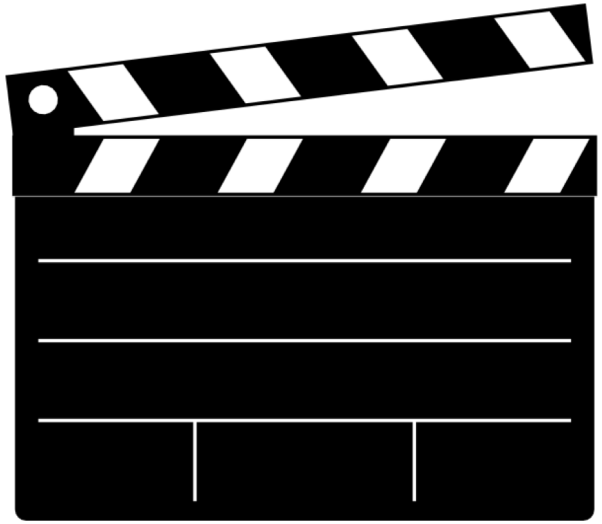
Semantic Feature Selection for Text

- Feature selection is difficult
  - Type: Filters, wrappers, hybrid
  - Too many methods: frequency-based, chi-square, information gain, mutual information, …
- We propose simple, robust and effective method based on semantics with t-test
- T-test: popular, easy to understand
- We use the 2-sample, unequal variance, 2-tailed test and look for statistically significant differences

Classifiers



Pattern Matching Classifier



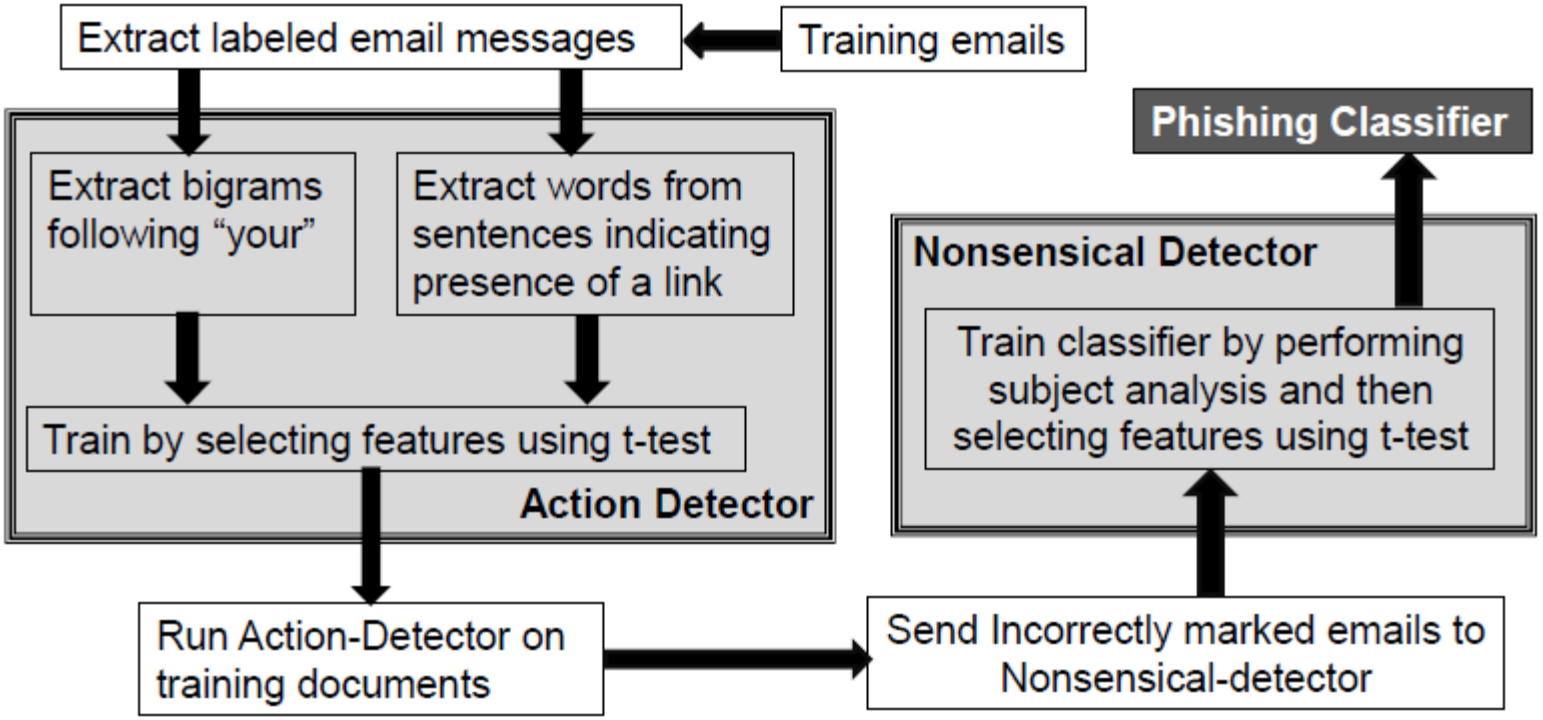
+

Detector

+ …

Detector

Flowchart of Training



Datasets

- Phishing: 4,550 emails [same as in ESORICS 2012]
- Legitimate emails: 10,000 emails from public Enron Inbox email database + 4,000 emails from Enron Sent Mail boxes
- Dataset divided into:
  - training (70% of phishing and Inbox emails)
  - testing (remaining 30% of phishing, remaining inbox, and 100% of sent emails)

Results

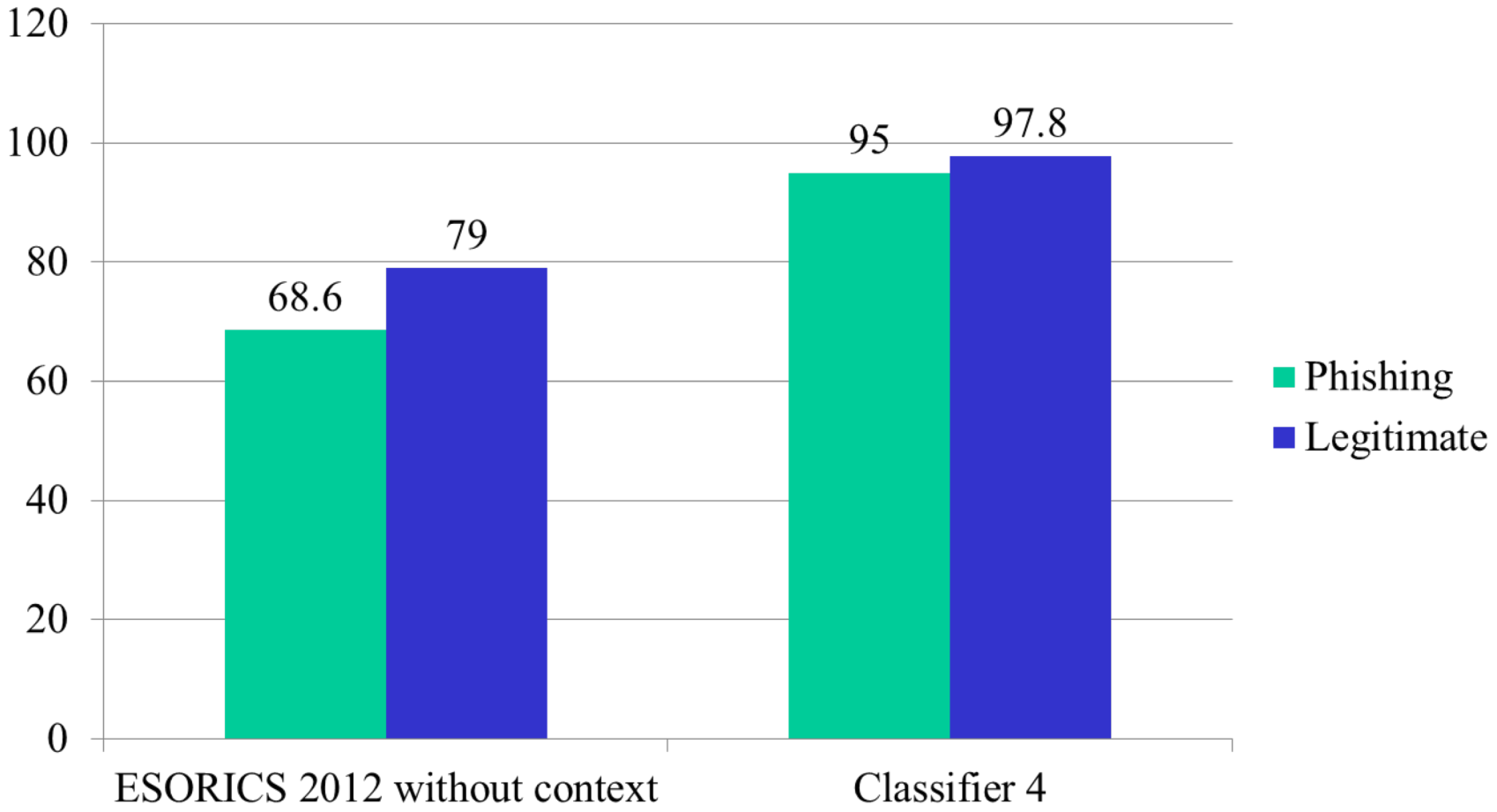
P - Phishing testing data set

I – legitimate testing data set (Inbox)

S – all legitimate sent emails (no training was done on this)

Classifier	*P*	*I*	*S*
Classifier 1	**92.88**	**4.96**	**4.17**
Action-Detector	73.6	1.92	1.96
Nonsensical-Detector	12.84	2.87	2.21
Other	6.44	0.17	0
Classifier 2	**92.01**	**4.88**	**3.9**
Action-Detector	72.23	1.4	1.76
Nonsensical-Detector	13.34	3.31	2.14
Other	6.44	0.17	0
Classifier 3	**94.8**	**2.16**	**2.37**
Action-Detector	75.1	0.5	0.72
Nonsensical-Detector	13.3	1.49	1.65
Other	6.44	0.17	0
Classifier 4	**95.02**	**2.24**	**2.42**
Action-Detector	75.82	0.57	0.77
Nonsensical-Detector	12.74	1.5	1.65
Other	6.44	0.17	0

Text Analysis Comparison



II. Phishing URL Analysis: Motivation

In our working on phishing email detection [ESORICS 2012], we used link analysis based on Internet search with the domain of the link + some terms extracted from the email

The method gave us 95% detection rate and low false positives

But, Google kept blocking our searches

- So we decided to do link analysis based on features extracted from the link alone
- Can a classifier based on link features and statistical tests detect a high percentage of phishing sites with low false positives?

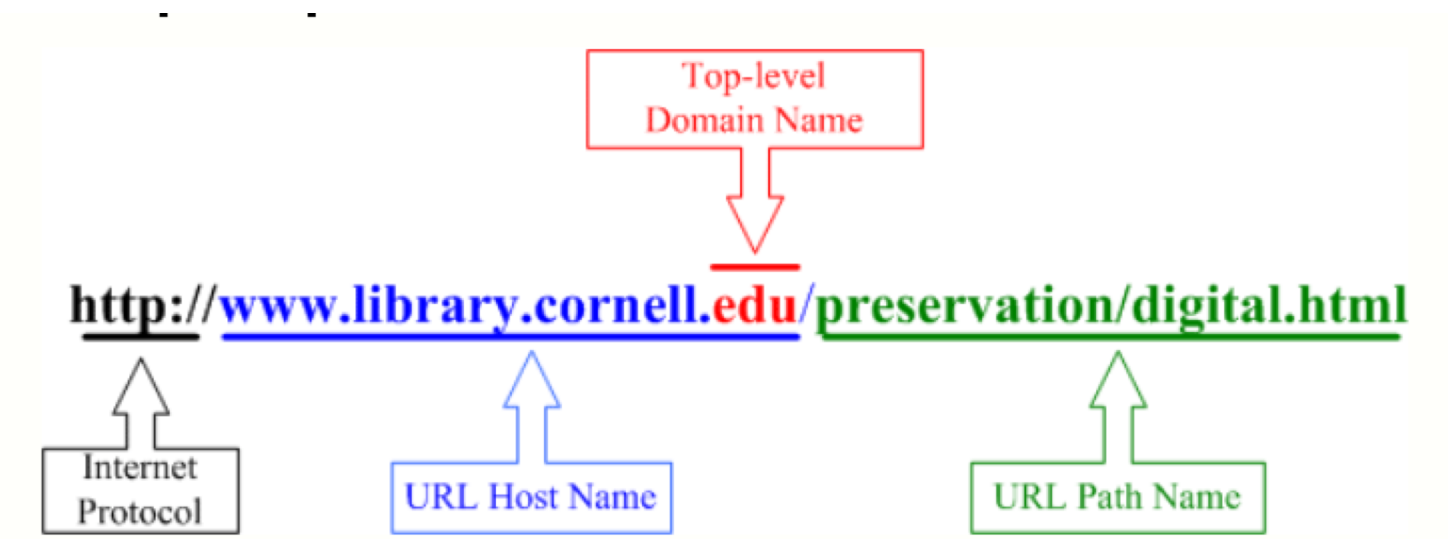
Outline of Phishing URL Analysis

- Features and statistical tests
- Classifiers
- Datasets
- Results
- Security Analysis

The Principle Behind Our Methods

- Focus on key differences between a phishing and a legitimate link
  - Phisher must convince user that link is genuine
  - At the same time it cannot be the legitimate web site
- Robust combination of features
- No host-based features

Parts of a URL



Features and Statistical Tests

- Normalized letter distributions
  - Cumulative distribution function
  - KS-test statistic
  - KL-divergence
  - Euclidean distance
- Edit distance
- Length of URL/Length of domain

- Sum of @'s and hyphens – e.g. www.paypal.com@phishing.com
- Number of punctuation symbols from {

  . ! # $ % & * , ; : `}
- Number of Top Level Domains in the path of the URL – e.g. www.xyz.com/YourBank.com.php

- Number of targets in URL, using Aho-Corasick Automaton for efficiency
- IP address in URL
- Suspicious words in URL

Machine Learning Algorithms

- PART
- SVM with p = 1 (SMO)
- C4.5 Decision tree (J48)
- Logistic regression
- Naïve Bayes
- Random forest

Datasets

- I – 25K URLs: 11K phishing URLs from PhishTank.com, 2/12/2014, and13K legitimate URL's taken from Alexa.com 2/11/2014.
- II* – 15K phishing URLs from Huawei Digital's repository, 15K legitimate URLs

- III* - 18K phishing URLs from APWG with 20K less popular legitimate websites gathered using the original authors' crawler.
- 11K random legitimate URLs from the DMOZ Open Directory Project, 11K phishing URLs from PhishTank.
- [* - Zhang, Wang 2012]

Dataset Characteristics Summary

- Legitimate URLs from 3 sources: Alexa.com, [Zhang, Wang 2012] , and DMOZ.
- Phishing URLs from 3 sources: PhishTank, Huawei's phishing repository, and APWG.
- Diversity of our data we analyzed the breakdown of unique TLDs as well as unique domains in the four sets.

- Our datasets are diverse with one exception, the DMOZ set.
- The top 50 most popular domain names account for roughly 10% of the data for the other three groups.
- DMOZ-PhishTank group's top 50 domains account for nearly 25% of the URLs.

Results for all features combination, Dataset I

5-fold cross validation, Weka

Classifier	Accuracy	False-positive rate
PART	99.0	0.89
Logistic	97.7	2.68
J48	**99.0**	0.83
Random Forest	98.9	**0.51**
SMO	98.5	0.80
Naïve Bayes	79.9	2.19

Results for all features combination, Dataset II

5-fold cross validation

Classifier	Accuracy	False-positive rate
PART	95.4	4.80
Logistic	94.7	4.66
J48	95.0	4.45
Random Forest	**95.7**	**3.25**
SMO	94.8	4.55
Naïve Bayes	83.9	8.09

Ada-boosted Results

Dataset I, 5 fold cross validation

Base Classifier	Accuracy	False positive rate
PART	**99.3**	0.48
Logistic	98.7	1.15
J48	**99.3**	0.54
Random Forest	**99.3**	**0.32**
SMO	98.5	0.76
Naïve Bayes	79.9	2.19

Stacked Classifier Results

Data Set	Accuracy	False-positive rate
Dataset I	99.2	0.58
Dataset II	96.3	3.17

5 base classifiers – PART, Logistic, J48, Random Forest and SMO

Default meta classifier

All Sets Combined

Classifier	Accuracy	False-positive rate
PART	93.2	7.08
J48	94.0	6.13
Random Forest	95.2/96.2*	4.69/3.61*

Three best classifiers, 5 fold cross validation

Approx. 115K URLs with 60K legitimate and 56K phishing

* - Adaboosted results

Security Analysis

- Phisher may try to imitate legal URLs closely to defeat character distribution analysis
- Not so easy, since there have been many such attempts already for popular targets such as PayPal
- Even if possible, features such as edit distance or targets in URL

Why Are Character Distributions Different?

- Presence of digits, special symbols, use of words such as ‘login’, ‘.js’, ‘.php’
- URL obfuscations using various encoding schemes
- Preset session attacks
- Cross-site scripting through URL formatting

III. Phishing Web Site Detection

- Based on
  - URL analysis
  - Content analysis
  - Behavior analysis
- For classical (all three methods) and hijack-based attacks (only content and behavior)

Dataset & Results

- Phishtank – 17,200 URLs
- Legitimate – 17, 200 URLs
- Whitelist of top 5,000 domains from Alexa
- 99.97% detection with 3.5% false positives
- 93.3% detection with 0.5% false positives
- Hijack-based attacks: 92.8% detection with 0.5% false positives

Attacker vs Defender – current state

Attacker vs Defender – ideal state

Conclusions, Future Work

- A sampling of robust schemes for email and URL deception attacks
- Much scope for work in security analytics
- The best is yet to come …

Acknowledgments

- Students: N. Hossain, K. Dyer, T. Thakur, A. Dunbar, N. Rai and S. Baki
- Collaborators: N. Shashidhar, C. Kari, A. Mukherjee and O. Gnawali
- ReDAS group: L. DeMoraes, A. Das, D. Lee, A. Nguyen (Alum: V. Vuppuluri)
- Sponsor: National Science Foundation

Contacts

CS Department: http://www.cs.uh.edu

ReDAS Lab: http://www.cs.uh.edu/~rmverma

Rakesh Verma - rverma@uh.edu