CERIAS

The Center for Education and Research in Information Assurance and Security

Natural Language IAS: The Problem of Phishing

Students: Lauren M. Stuart, Gilchan Park. Advisors: Prof. Julia M. Taylor, Prof. Victor Raskin

Examples: snippets from actual phishing emails[1], with Ontological Semantics Technology[2] –based analysis.

Existing Strategies

Several proposed policies and implemented tools exist for separating likely phishing emails from legitimate emails.

- Blacklists/whitelists for domains and addresses [3]
- Link analysis: target/text mismatch, features of common bad URLs [4,5]
- Language analysis: common keywords, expanded language consideration [5,6,7]
- Visual/DOM analysis: Page elements of fake and real login pages [5,8,9]

Proposed Direction

Expand the use of linguistic semantics in information security.

- Quantify and/or canonicize *linguistic and logical* hallmarks of phishing emails for detection
  - Similar methods in stylometric analysis, automatic characterization of network events
- Semantic analysis of message content: comparisons and thresholds
  - Ontological Semantic Technology: semantic scripts, text meaning representation, fuzzy logic

References

[1] Nazario, J. The online phishing corpus: http://monkey.org/~jose/wiki/doku.php

[2] Raskin, V. and Taylor, J. M., The (not so) unbearable fuzziness of natural language: The ontological semantic way of computing with words. NAFIPS 2009.

[3] Xiang, G., Pendleton, B. A., Hong, J. I., and Rose, C. P. A hierarchical adaptive probabilistic approach for zero hour phish detection. ESORICS'10.

[4] Garera, S., Provos, N., Chew, M., and Rubin, A. D. A framework for detection and measurement of phishing attacks. WORM'07.

[5] Xiang, G., Hong, J., Rose, C. P., & Cranor, L. CANTINA+: a feature-rich machine learning framework for detecting phishing web sites. TISSEC'11.

[6] Abu-Nimeh, S., Nappa, D., Wang, X., and Nair, S. 2007. A comparison of machine learning techniques for phishing detection. APWG 2007.

[7] Park, G. Text-based phishing detection using a simulation model. Masters' Thesis, Computer and Information Technology, Purdue University, 2013.

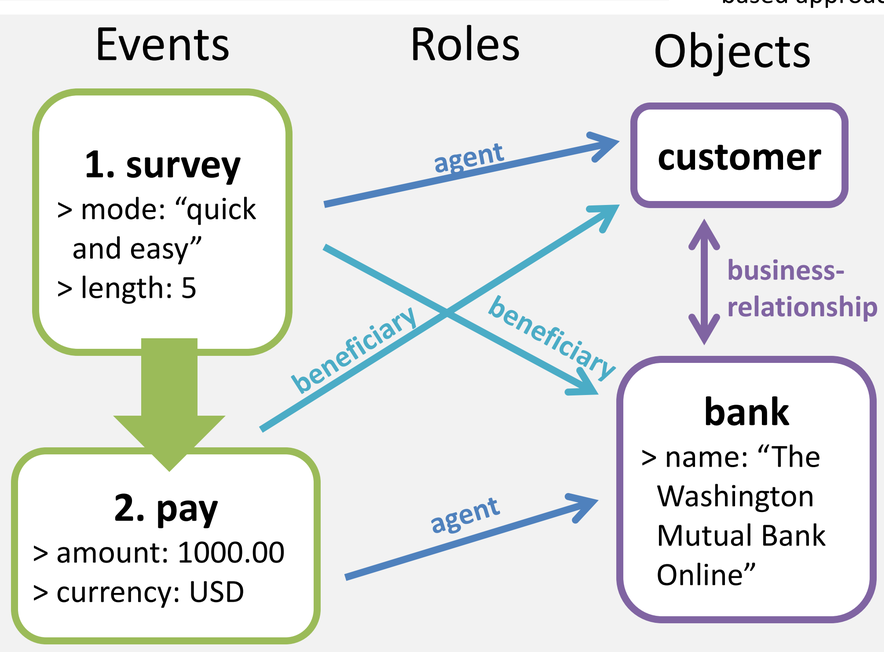
[8] Chen, T.-C., Dick, S., and Miller, J. Detecting visually similar web pages: Application to phishing detection. ACM TOIT 2010.

[9] Rosiello, A. P. E., Kirda, E., Kruegel, C., and Ferrandi, F. A layout-similarity-based approach for detecting phishing pages. SecureComm'07.

Work/Compensation Mismatches

Phishing emails sometimes promise a great deal of money or benefits in return for seemingly little work in order to lure people into divulging sensitive information.

"The Washington Mutual Bank Online department kindly asks you to take part in our quick and easy 5 questions survey. In return we will credit $1000.00 to your account - Just for your time!"



Triggering Comparison: Scripts

The two events are causally related, and complementary, so they may fit a rule for work-for-reward.

script: work-for-reward
> sequence
  > 1: action
  > 2: compensation
> consistency requirements
  > complementary roles
  > magnitude(action) ~= magnitude(compensation)

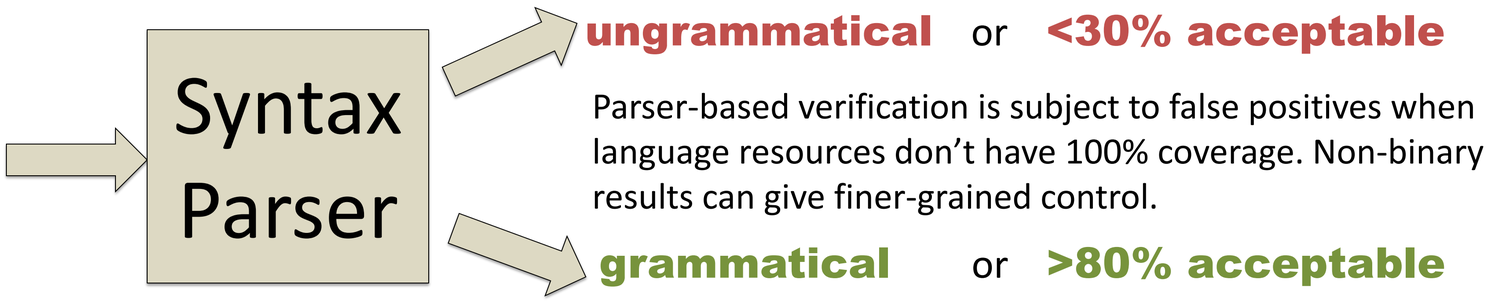
Possible candidates for ~= function: human-set threshold, mined threshold,-- but how to express proportionality?

Awkward/Unprofessional Phrasing

Though many professional, legitimate emails do have grammatical mistakes and awkward phrasing, it may pay to be more skeptical of an "official" communication that does have these mistakes.

"Why you become a PowerSeller?"

"If you agree, please within 24 hours."

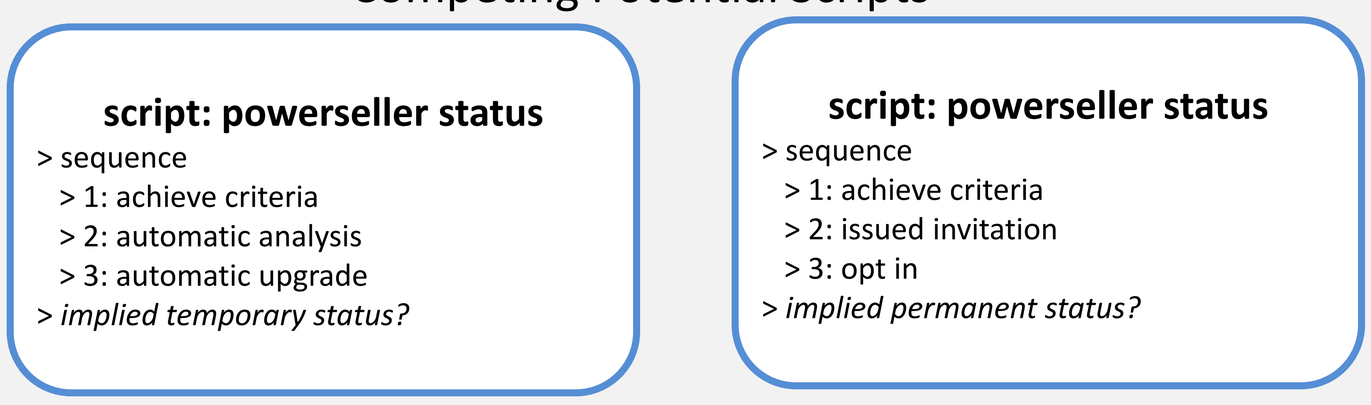


Inconsistency with Company Policy

Some phishing emails directly contradict publicly-available company policies and can be fact-checked if presented as fact. Checking inconsistency in semantic structures is potentially complicated but one simplification of the idea could look like this:

"You have to click the highlighted fields below and in few days you will become an eBay power seller."

"You don't need to apply for the PowerSeller program. If you qualify, you'll automatically be included." (http://pages.ebay.com/help/sell/sell-powersellers.html)



Comparison against an authoritative source is common; in this case, semantic analysis allows for direct comparison of texts.

Research supported in part by a Purdue Research Foundation Ph.D. Fellowship grant to Prof. Taylor.

PURDUE UNIVERSITY