Using Lexical Features for URL Classification- A Machine Learning Approach

Apoorva Joshi
FireEye

Levi Lloyd
Lawrence Livermore National Lab

Paul Westin
FireEye

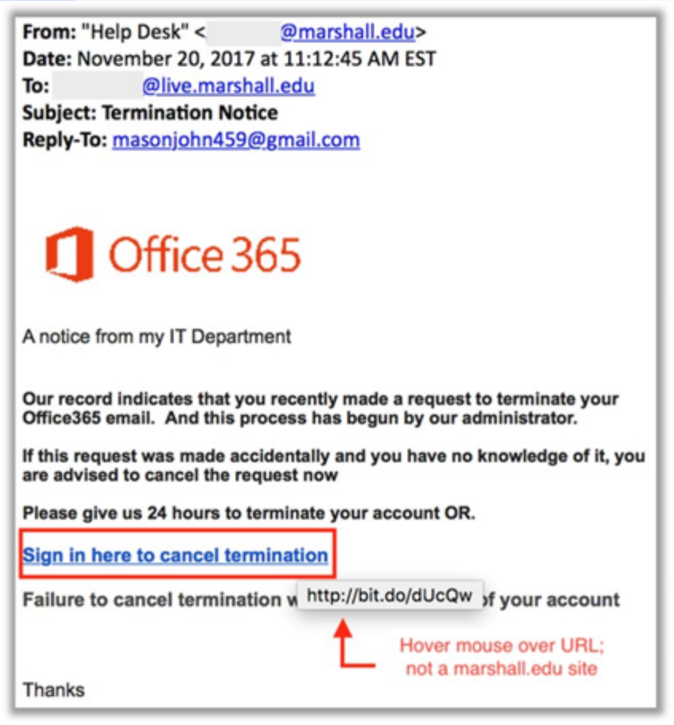
Srini Seethapathy
FireEye

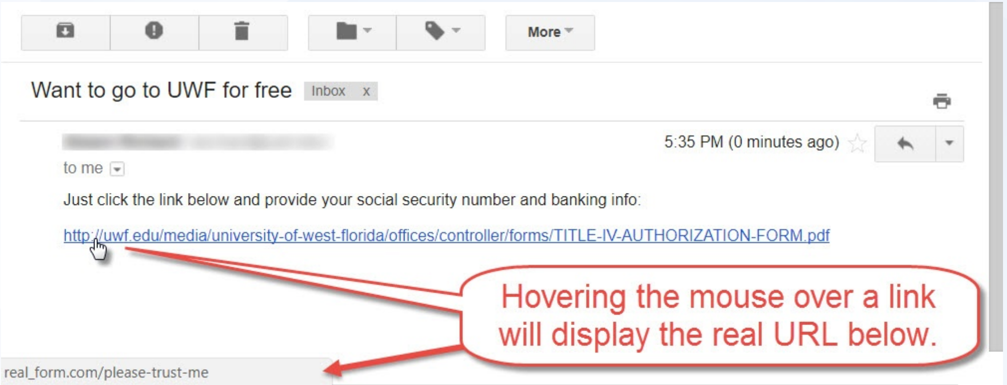
OUTLINE

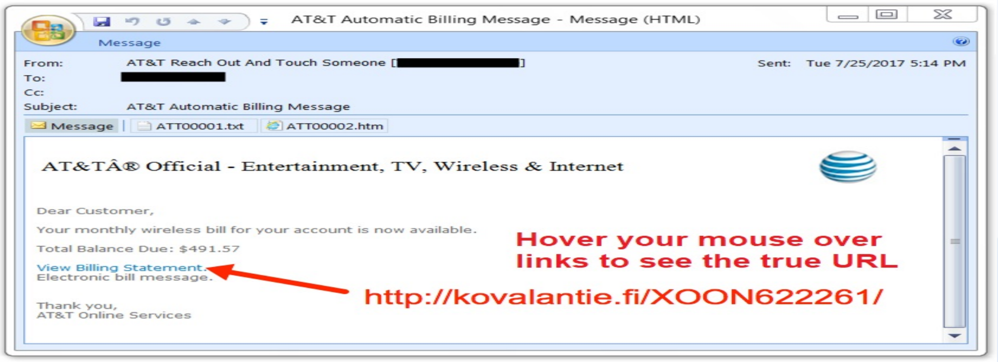
- Motivation
- Previous Work
- Deployment Specification
- Requirements and Goals
- Tasks and Tools
- Observations
- Conclusions
- Future Work

©2019 FireEye

MOTIVATION







©2019 FireEye

PREVIOUS WORK

Blacklists

Host information, network traffic etc.

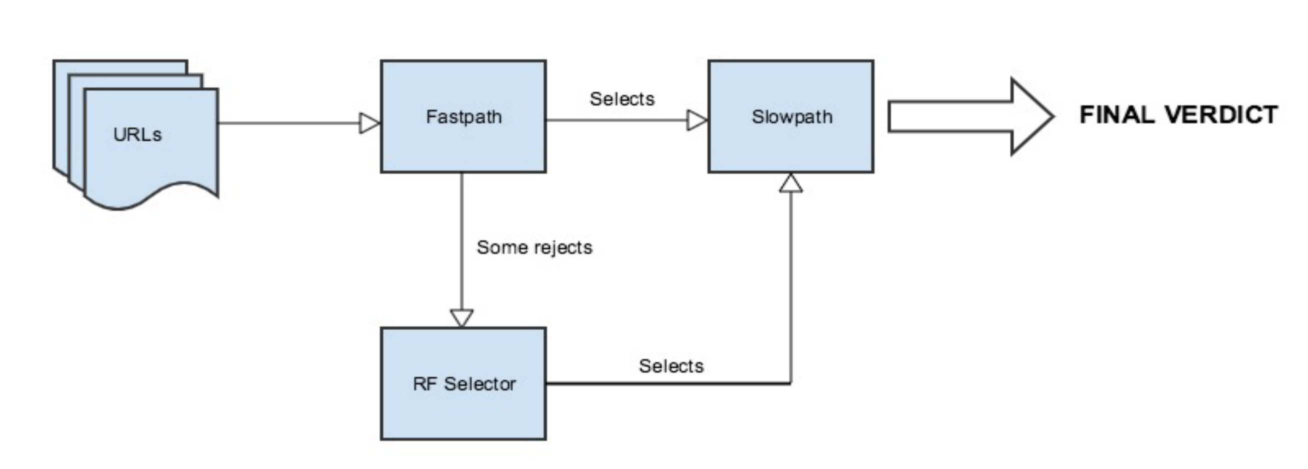
HTML and Javascript content

LEXICAL FEATURES ONLY???

©2019 FireEye

DEPLOYMENT SPECIFICATION

- Model should run as plugin for FAUDE (FireEye Advanced URL Detection Engine)
- Should correct FNs from fastpath analysis
- URLs to be sent for slowpath analysis based on the model verdict



©2019 FireEye

REQUIREMENTS AND GOALS

- Model should act as a means of down selection and/or detection
- False Negative Rate should be very low
- False Positive Rate such that the model results in at most 20% increase in current load
- Model latency should be in the order of 10^{-1} ms

©2019 FireEye

THE DATASET

- ~5.5 million labelled URLs
- 60% benign, 40% malicious URLs
- Collected from different sources – Openphish, Alexa whitelists, FireEye products and honeypots

©2019 FireEye

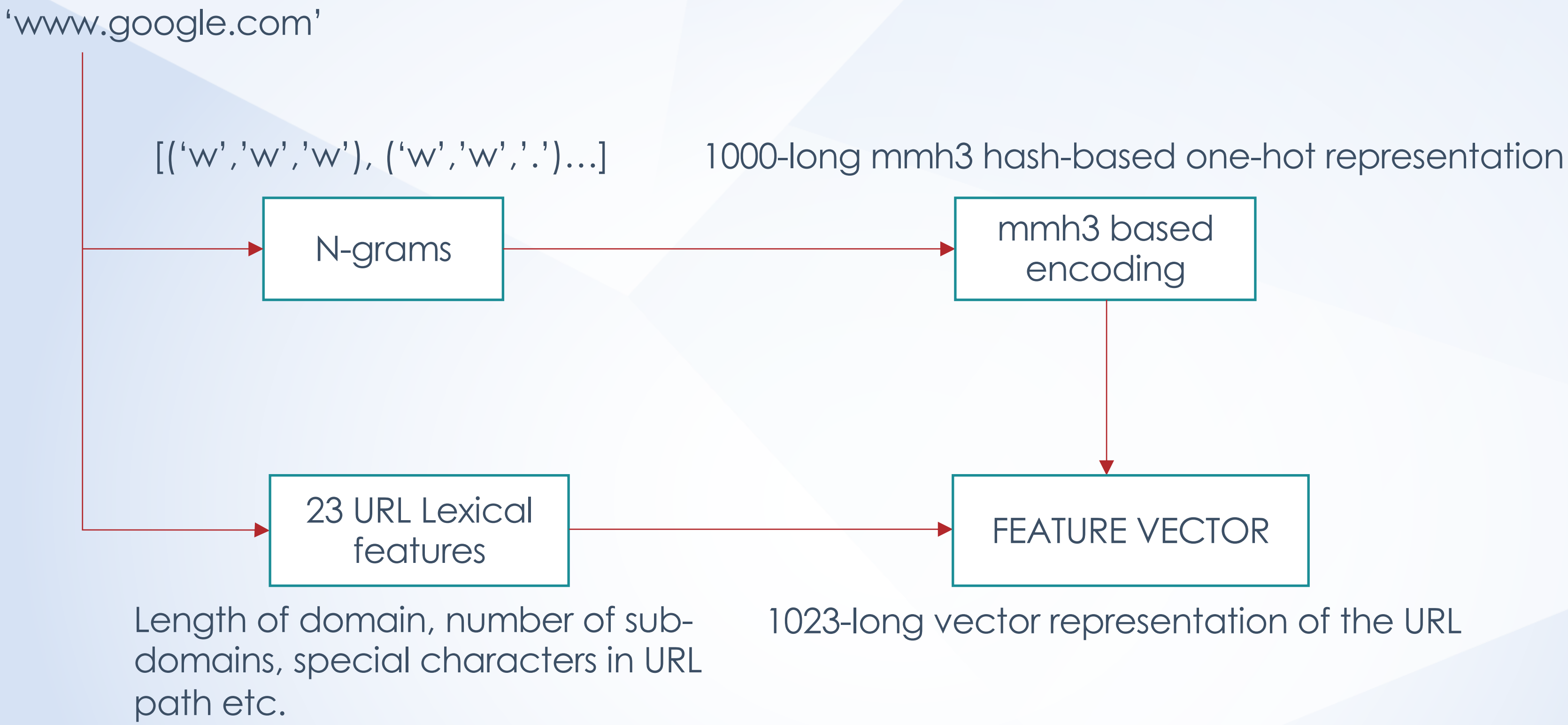
TASKS AND TOOLS

Feature vectors

TASK	TOOLS
N-grams of URLs	NLTK, mmh3
Extract lexical features	urllib
Modelling	Random Forest

©2019 FireEye

FEATURE VECTORS



©2019 FireEye

FEATURE VECTORS

http://www.video.platinumindustrialcoatings.com/wp-content/plugins.php?to=calgaryps3&message=28dd33dc15e8c68934883418341967

Subdomain

Domain

Top Level Domain

Path

Query

Parameters

Complete list of features: https://arxiv.org/abs/1910.06277

©2019 FireEye

MODELLING

- Simple Classifiers- Logistic Regression, Naïve Bayes
- Bagging and Boosting Classifiers- Random Forest, Gradient Boost, Adaboost
- Metrics- Accuracy, AUC, FNR

©2019 FireEye

OBSERVATIONS

ALGORITHM	ACCURACY (%)	AUC	FNR (%)
Logistic Regression	87	0.96	4.75
Naïve Bayes	70	0.74	10.38
Random Forest	92	0.99	0.38
Gradient Boost	90	0.92	9
Adaboost	90	0.9	10

©2019 FireEye

OBSERVATIONS

Suspicious URL patterns:

- TLDs in shady list- .biz, .info, .ru, .cn
- Keywords, special characters in URL path
- IP address in primary domain
- High entropy hostnames
- Uppercase or single character directory

©2019 FireEye

OBSERVATIONS

Number of trigrams	Number of lexical features	Accuracy (%)	FPR (%)	FNR (%)
1000	0	85	29.8	0.4
1000	23	92	16.8	0.38
300	23	93	12.5	0.93
100	23	94	11.5	1.09
0	23	95	8.15	1.11

The Random Forest feature importances also showed that it was focusing on both ngram and lexical features

 ©2019 FireEye

OBSERVATIONS

Max depth	Accuracy (%)	FNR (%)
5	72	1.13
15	88	0.48
20	92	0.38
27	94	0.73
30	95	0.75

Tuning other parameters had no real effect on the evaluation metrics

©2019 FireEye

CONCLUSIONS

- ~22% increase in detections for < 20% increase in load
- Reduction in FNs
- Purely lexical models can be used for fast verdicts on URLs
- Alternative to heuristic-based downselection which needs manual updates

©2019 FireEye

FUTURE WORK

- Deep Learning approach
- Augment the model with new features as necessary
- Cache model verdicts

 ©2019 FireEye

ACKNOWLEDGMENTS

Levi Lloyd

Paul Westin

Srini Seethapathy

FireEye

 ©2019 FireEye

Thank You!

Email: apoorva.joshi@fireeye.com