

BlackThread

**Akshay Gupta, Evan Downing,
Ashwin Chintaluri, Yash Shah,
Liyan Wan, Deepti Kochhar,
Haoran Ma, Rohit Belapurkar**

Project Requirements

- Web & Search Engine Crawling for Data Discovery
- Web crawler that is able to collect various data points from a site
- Ranking system on the most typical sites
- An analysis of agility

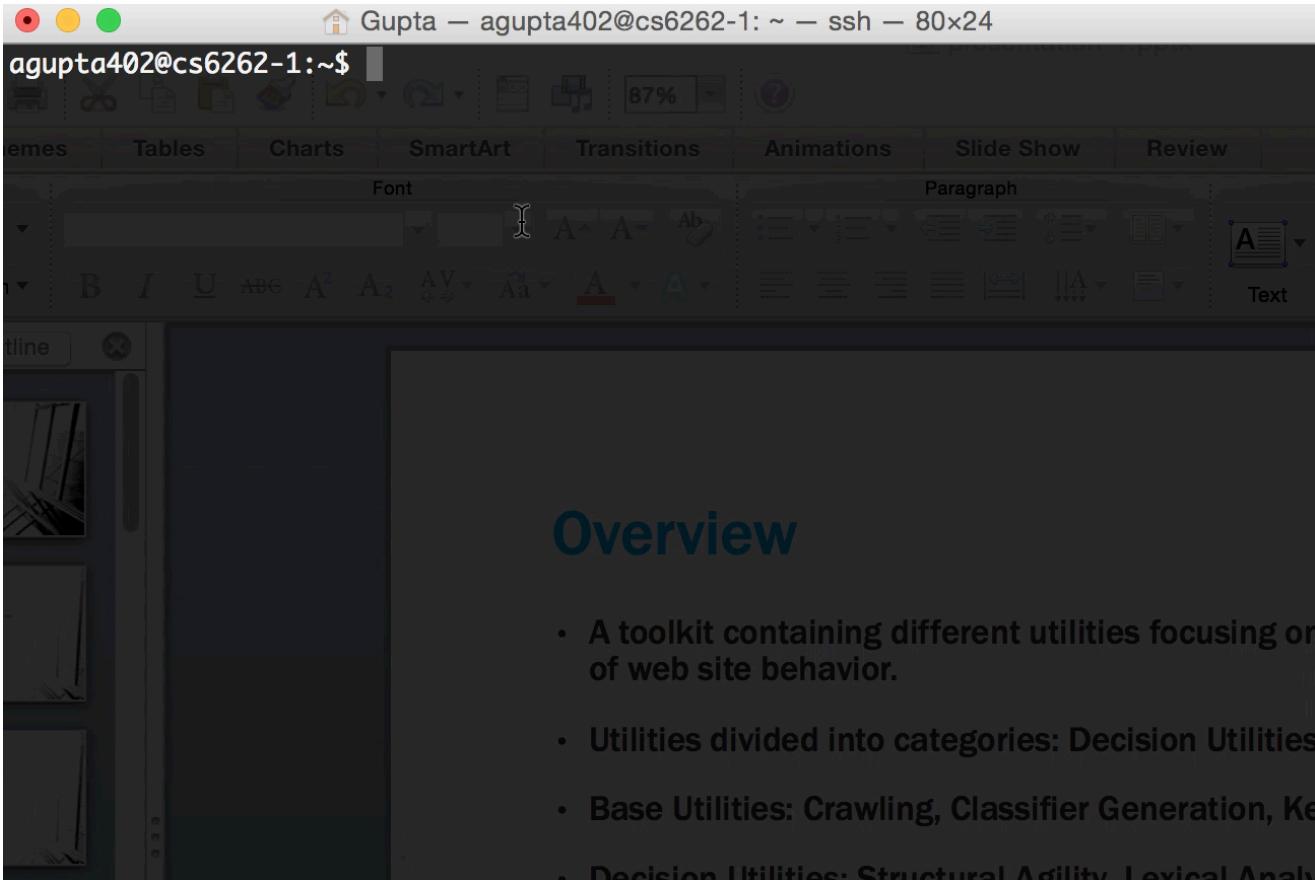
Road Map

- BlackThread Overview
- Machine Learning Background
- Crawling Background
- Web structure agility
- Lexical Analysis of URL
- iFrame and JavaScript
- DNS Agility
- Conclusion

Overview

- A toolkit containing different utilities focusing on capturing different aspect of web site behavior.
- Utilities divided into categories: Decision Utilities and Base Utilities.
- Base Utilities: Crawling, Classifier Generation, Keyword Generation, Ranking
- Decision Utilities: Structural Agility, Lexical Analysis URL, iFrame and JavaScript, DNS Agility

Demo: Introduction



Machine Learning Background

- Classification
- Cross Validation
- Evaluation Metrics
- Model Selection

Crawling Background

- Crawling - automated visiting of websites
- Scrapy (Python library)
- Our spiders: Tags, Content, iFrames/JS, URLs

Web Structure Agility

Web Structure Agility - Why

- Intuition that legitimate sites will not change structure of website
- Malicious sites may change structural layout of page in order to avoid detection
- Levenshtein Distance
- Does not rely on content of website (which could be anything)

Demo: Crawling

The screenshot shows a Safari browser window with the URL www.antonakakis.org. The page content includes:

- A portrait photo of Manos Antonakakis, Assistant Professor.
- Contact information:
 - School of Electrical and Computer Engineering
 - School of Computer Science (Adjunct)
 - Room 3366A, Klaus Advanced Computing
 - Georgia Institute of Technology
 - 266 First Drive
 - Atlanta, GA 30332-0765
 - manos@gatech.edu
- Links: [Public Key](#) | [Thesis](#) | [Blog](#)
- Georgia Tech Information Security Center (GTISC)**
Messaging, Malware and Mobile Anti-Abuse Working Group (M3AAWG)
- News**
 - CFP 9th eCrime Symposium: <http://ecrimeresearch.org/events/ecrime2014/cfp>
 - Our paper on "DNS noise" was accepted in [DSN 2014](#). Bravo [Yizheng](#)!
 - Our paper on botnet takedowns was accepted in [CCS 2013](#). Bravo [Yacin](#)!
 - Our paper on the detection of sinkholes was accepted in [LEET 2013](#). Bravo [Babak](#)!
 - Our paper on malware downloads was accepted in [ESORICS 2013](#). Bravo [Phani](#)! The source code for Amico is here: <https://code.google.com/p/amico/>
- Research**

My main research interests revolve around network security, computer security and anomaly detection. I am very happy when our research has operational impact ([cso](#), [circleid](#), [darkreading](#), [tnews](#), [issn](#), [darkreading](#), [threatpost](#), [eweek](#), [scmagazine](#), [pcadvisor](#), [NYTimes](#), [crn](#), [scmagazine](#), [ars](#), [iw](#), [The Economist](#)).
- Students**
 - Graduate Level
 - [CS] [Yacin Nadji](#) (co-advise with Wenke Lee)
 - [CS] [Yizheng Chen](#) (co-advise with Wenke Lee)
 - [CS] [Chaz Lever](#)
 - [ECE] [Panagiotis Kintis](#)

At the bottom of the screen, the Mac OS X dock displays various application icons, including Finder, Google Chrome, Terminal, Mail, Zulu, Java, Xcode, Unity, RStudio, Microsoft Word, MySQL Workbench, Safari, and iWork.

Demo: Machine Learning

A screenshot of a Mac OS X desktop environment. At the top, there's a dock with various icons including Finder, Mail, Safari, and others. The system tray shows battery level (17%), signal strength, and the date/time (Thu Nov 20 8:40 PM). Below the dock is a menu bar with 'File', 'Edit', 'View', 'Project', 'Run', 'Edit', 'Build', 'Run', 'Terminal', and 'Help'. The main window is a code editor titled 'BlackThread - [~/Projects/BlackThread]'. It has tabs for 'rawl.py', 'tags/lexical.py', 'tags/database.py', and 'tags/.../tagSpider.py'. The current file, 'tags/.../tagSpider.py', contains Python code related to machine learning classifiers and stratified k-fold cross-validation. A terminal window is open below the code editor, showing a warning message about overflow in the exponential function. The bottom of the screen features the Mac OS X Dock with icons for Mail, Safari, and other applications.

```
cross_validation
t SGDClassifier
adientBoostingClassifier
ndomForestClassifier
t LogisticRegression
_curve, auc

import StratifiedKFold

database
se"
tent(database)

n.py

untimeWarning: overflow encountered in exp
```

0 azlyrics.com
1 chaturbate.com
2 conduit.com
3 domaintools.com
4 ero-advertising.com
5 exoclick.com
6 ink361.com
7 issuu.com
8 justdial.com
9 npr.org
10 pixiv.net
11 odesk.com
12 reimageplus.com
13 subscene.com
14 teepublic.com
15 ziddu.com
16 zedo.com
17 zippyshare.com
18 irctc.co.in
19 paytm.com
20 bookmyshow.com
21 jagran.com
22 comsec.com.au
23 bigpond.com
24 www.vcommission.com
25 subaonet.com

Lexical Analysis

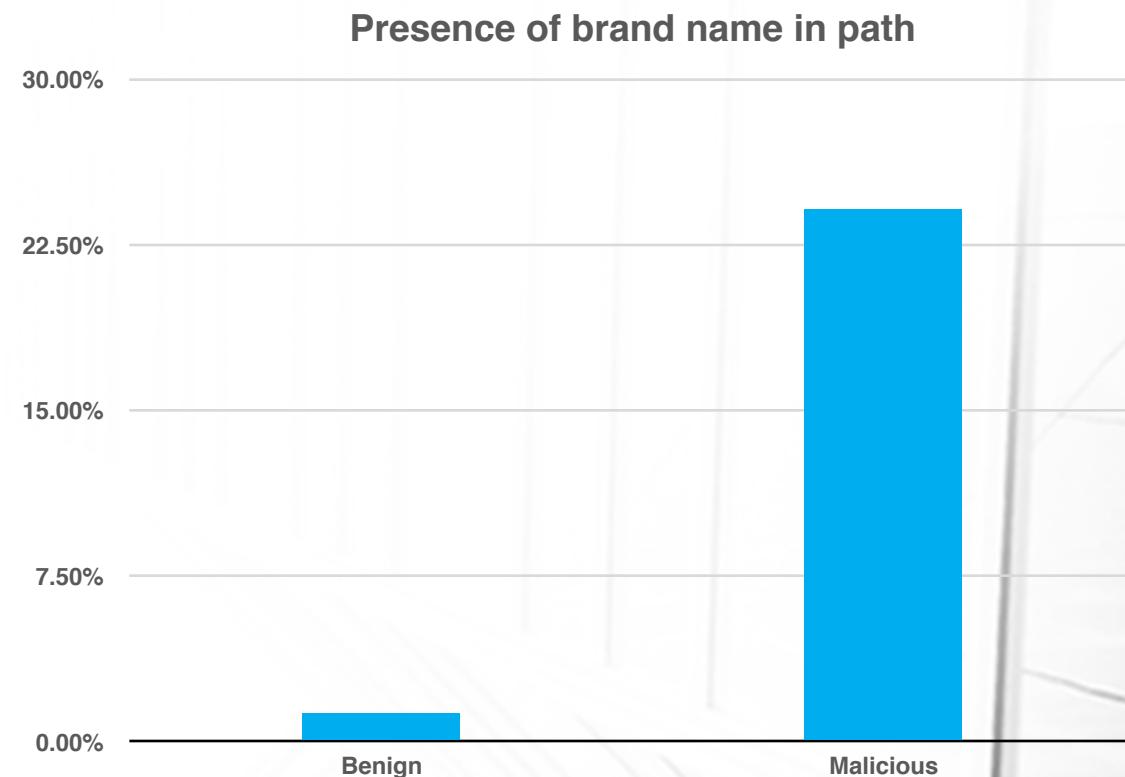
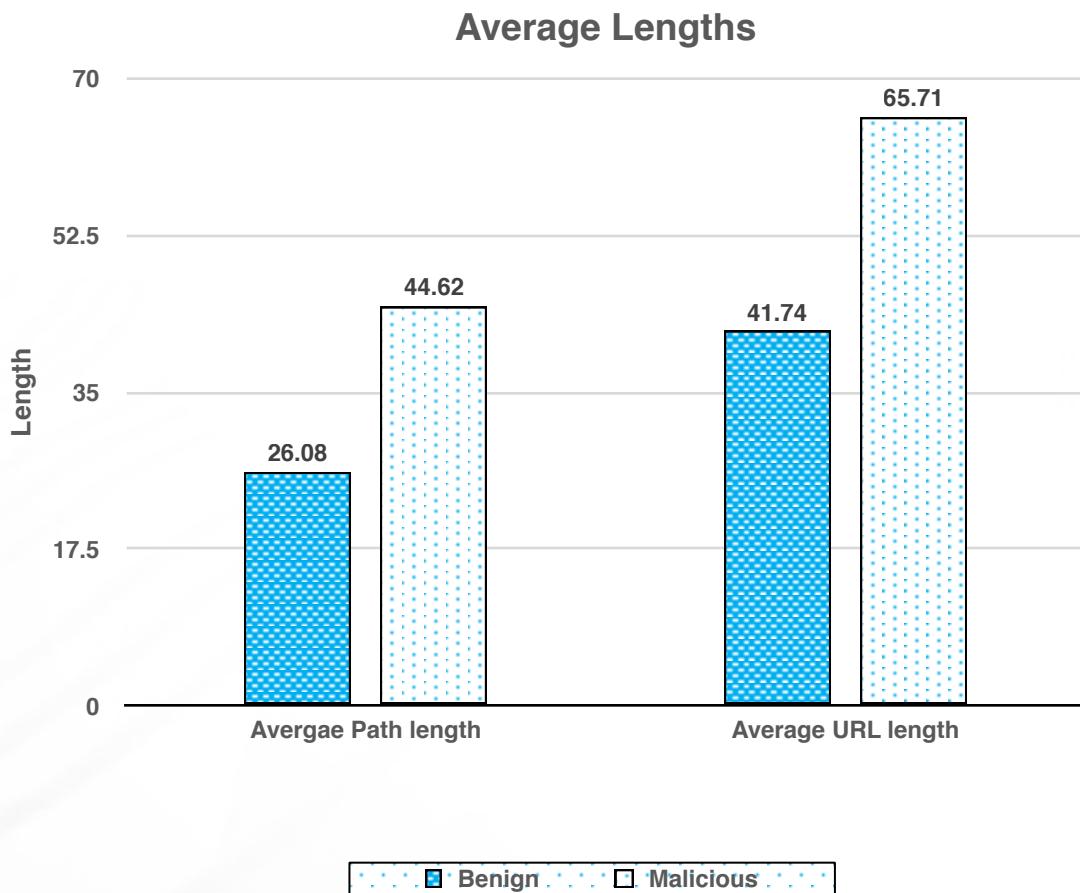
Lexical Analysis - Why

- Blacklisting – Inefficient to tackle the problems of frequently changing domains
- Need something more dynamic!
- URL's can be classified as malicious or benign based on their distinct characteristics.

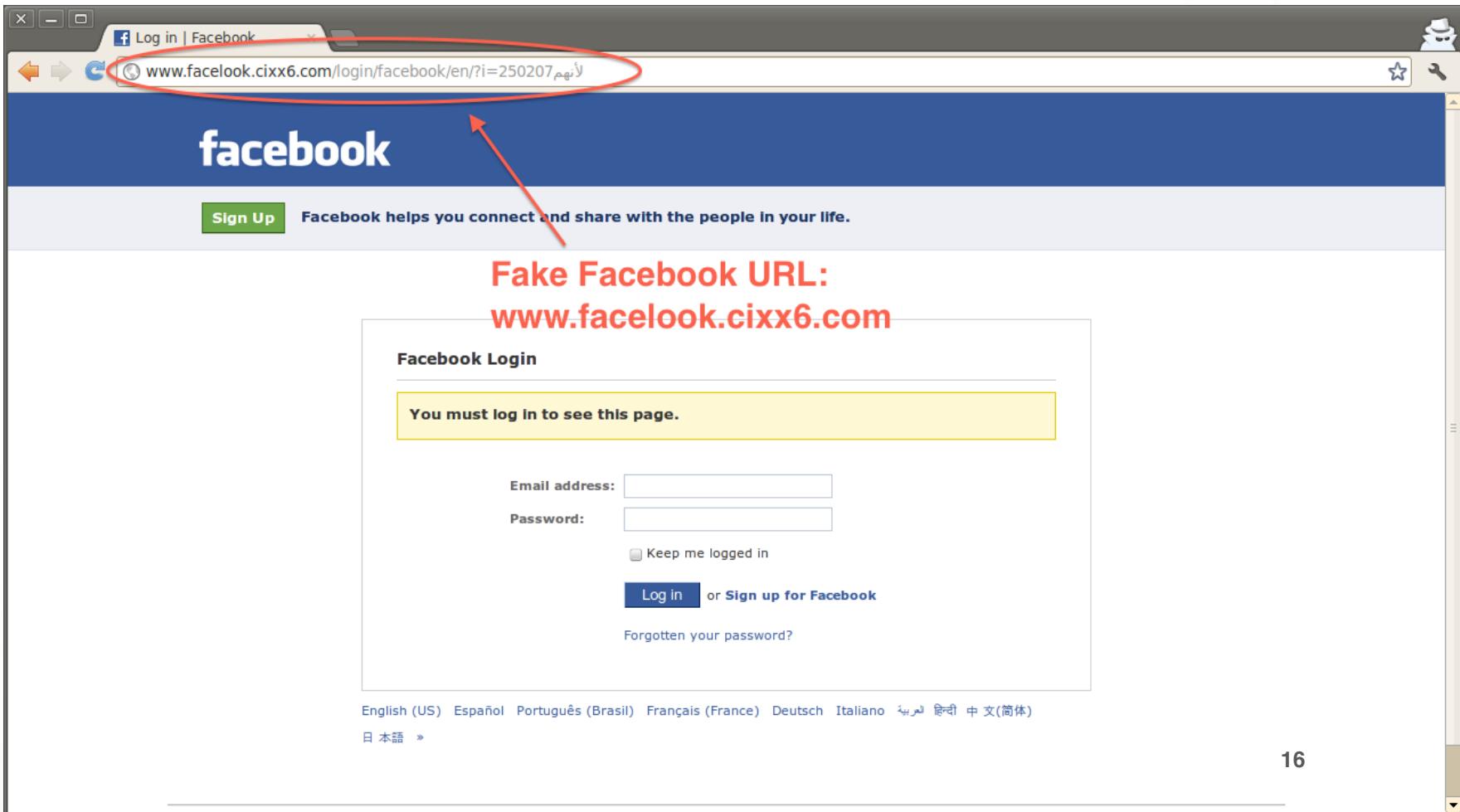
Features extracted from URL

Serial no.	Features analyzed for each URL
1	<i>Presence of brand name in the URL path</i>
2	<i>URL Length</i>
3	<i>Domain token count</i>
4	<i>Domain length</i>
5	<i>Average domain token length</i>
6	<i>Maximum domain token length</i>
7	<i>Path token count</i>
8	<i>Path length</i>
9	<i>Average path token length</i>
10	<i>Maximum path token length</i>
11	<i>Percentage of special characters</i>
12	<i>Character frequency</i>

Comparison of some URL characteristics



Phishing attack targeting Facebook



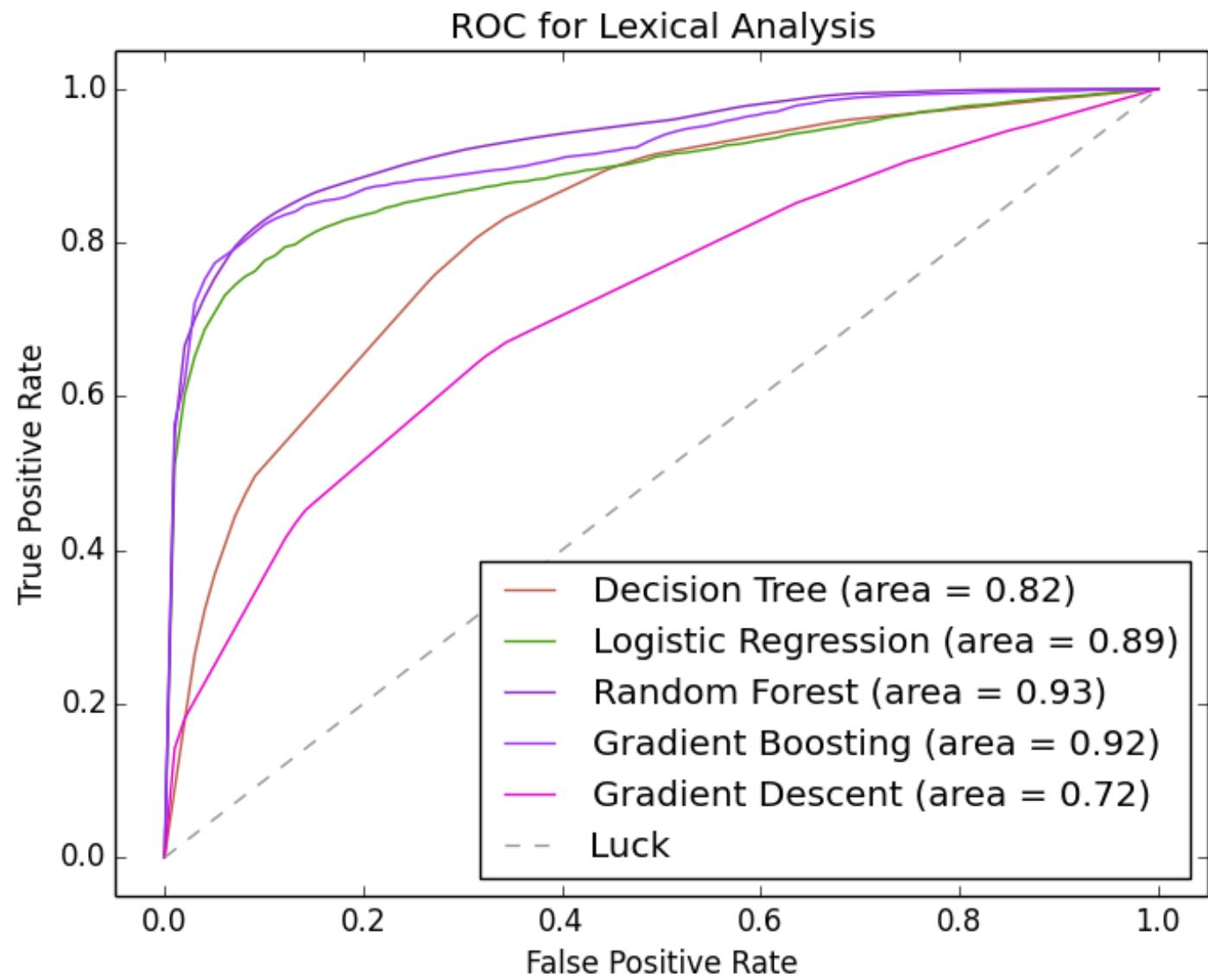
Data Collection and Training

- We considered a data set of malicious and benign URLs to train our machine learning classifier.
- Benign URL's- Crawled from Alexa top 500 websites.
- Malicious URLs- 10,000 verified phishing URLs from www.phishtank.com.
- We extracted the features and passed them to the machine learning classifier.
- We used Random Forest for classifying whether the URL is malicious or benign.

Model Selection: Evaluation Statistics

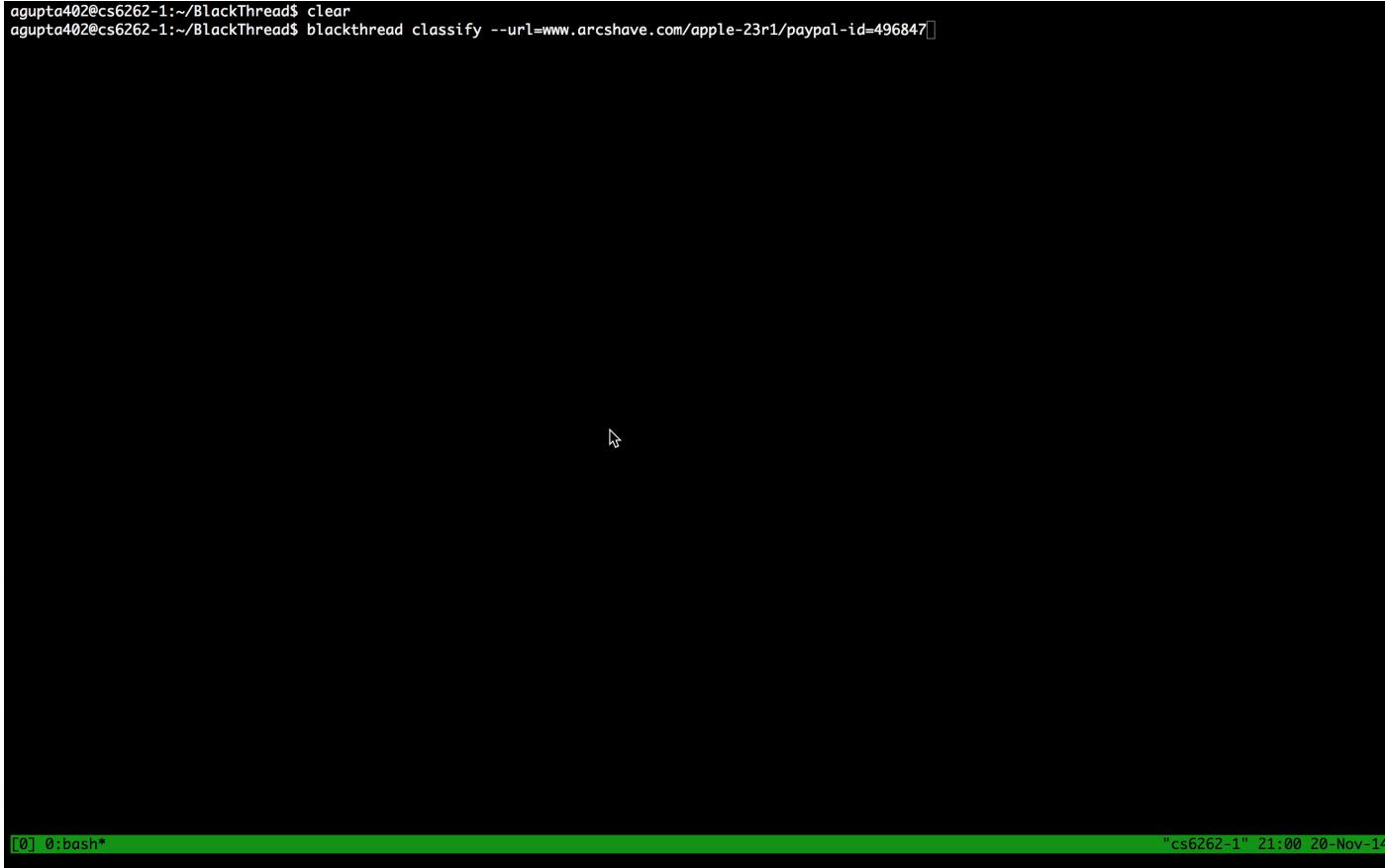
	Accuracy	Precision	Recall	ROC Area
Support Vector Machine	0.79 (+/- 0.16)	1.00 (+/- 0.00)	0.69 (+/- 0.22)	0.84 (+/- 0.11)
Logistic Regression	0.83 (+/- 0.10)	0.85 (+/- 0.14)	0.85 (+/- 0.03)	0.89 (+/- 0.09)
Random Forest	0.86 (+/- 0.11)	0.85 (+/- 0.14)	0.94 (+/- 0.02)	0.94 (+/- 0.09)
Decision Tree	0.82 (+/- 0.10)	0.80 (+/- 0.13)	0.92 (+/- 0.03)	0.82 (+/- 0.10)
Stochastic Gradient Descent	0.76 (+/- 0.14)	0.76 (+/- 0.12)	0.72 (+/- 0.16)	0.85 (+/- 0.11)
Gradient Tree Boosting	0.84 (+/- 0.10)	0.83 (+/- 0.14)	0.91 (+/- 0.02)	0.92 (+/- 0.11)

Model Selection: Receiver Operating Characteristics



Demo: Online Classification

```
agupta402@cs6262-1:~/BlackThread$ clear  
agupta402@cs6262-1:~/BlackThread$ blackthread classify --url=www.arcshave.com/apple-23r1/paypal-id=496847
```



Challenges

- URL based Lexical Analysis cannot be performed on URLs which employ URL shortening services.
- Eg:<http://tinyurl.com/urlwiki>
- Large number of variations possible for a given set of URLs (malicious or benign) leading to false positives.
- Thus cannot be used a stand alone method of detection.
- However, it can be used as a recommendation system.

Frame and Script Analysis

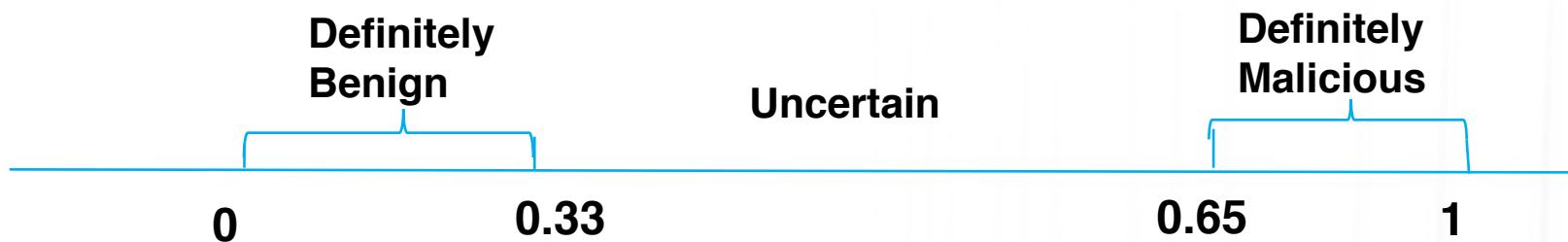
Four Statistical Features

- iFrame Tag Ratio (iTR)
- iFrame Zero-Size Counter
 - "height='0'","width='0'","display:none","opacity:0","visibility:hidden"
- iFrame JS Ratio (iJSR)
 - Crawler can not perform redirect
- iFrame JS function calls
 - "eval","setTimeout","link","unescape","exec","unbound","escape"

Analysis: Feature

	Frame Ratio*10000	JS Ratio	Zero Iframe	Js calls	Total
<i>Benign</i>	0.70	0	0.73	3.71	Max 0.33
<i>Malicious</i>	4.14	0	2.90	5.14	Min 0.65
<i>Threshold</i>	1.43	0	1.82	4.43	0.49
	25%	0%	40%	35%	

Analysis



- Weight for different Features
- 2000 malicious and benign known URL for testing
- True Positive: 93%, False Positive: 7%

DNS Agility

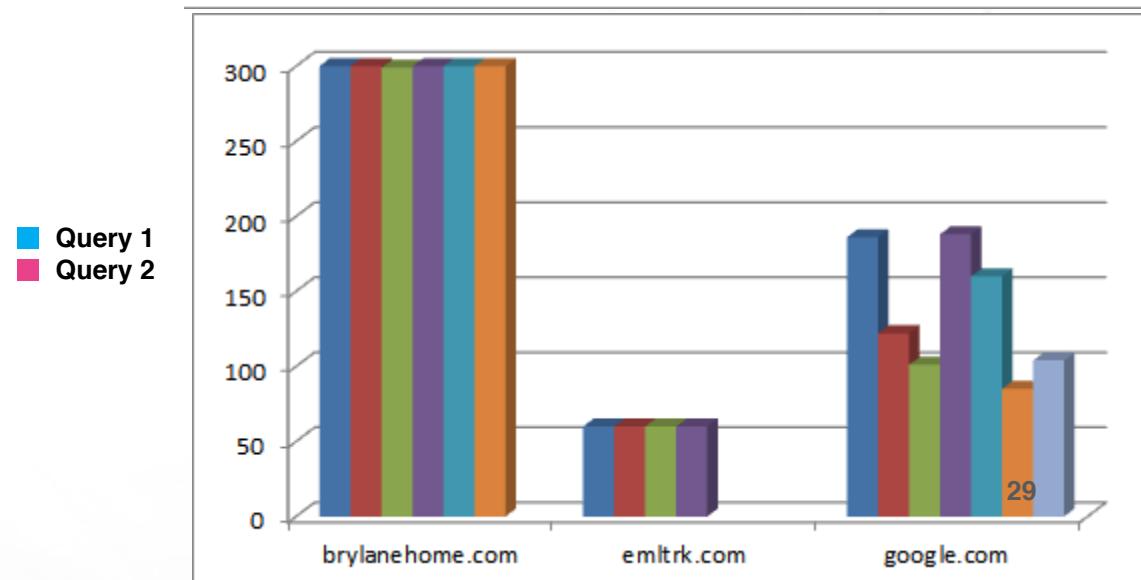
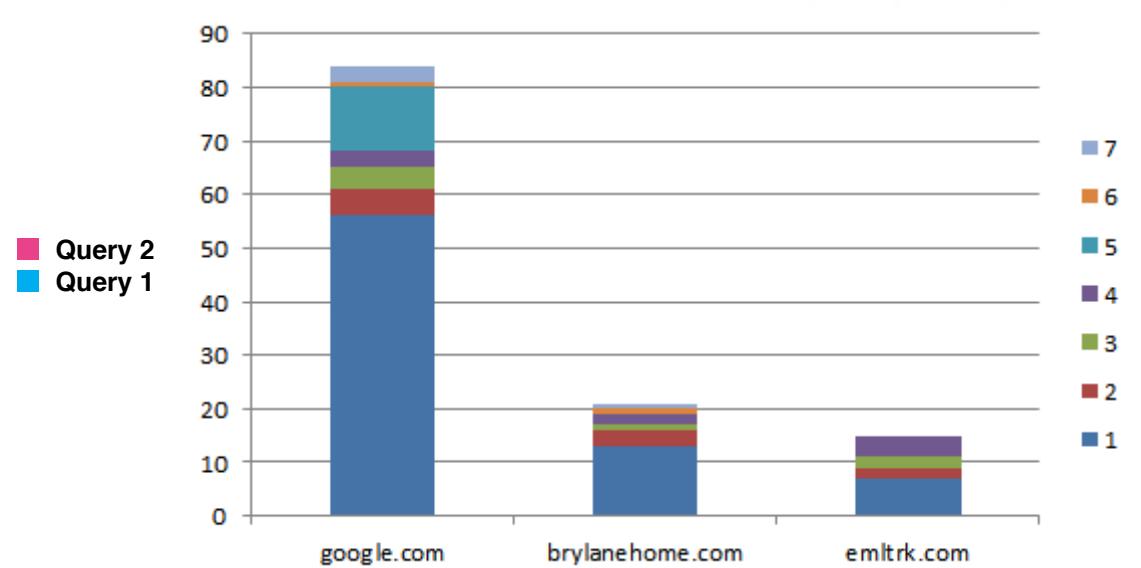
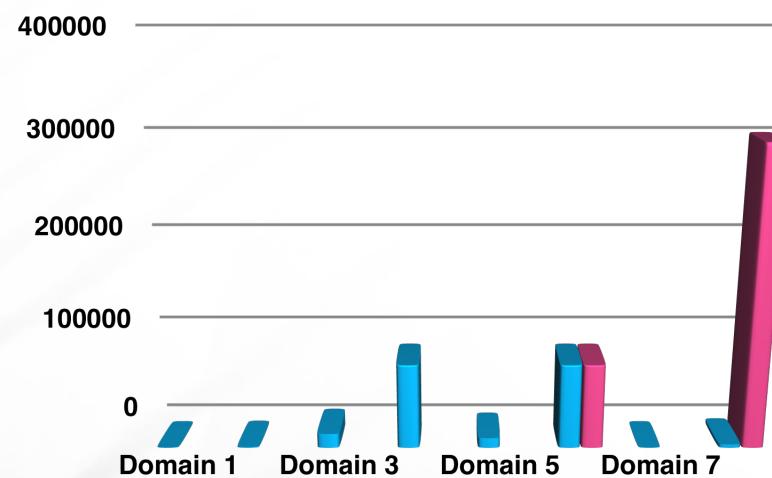
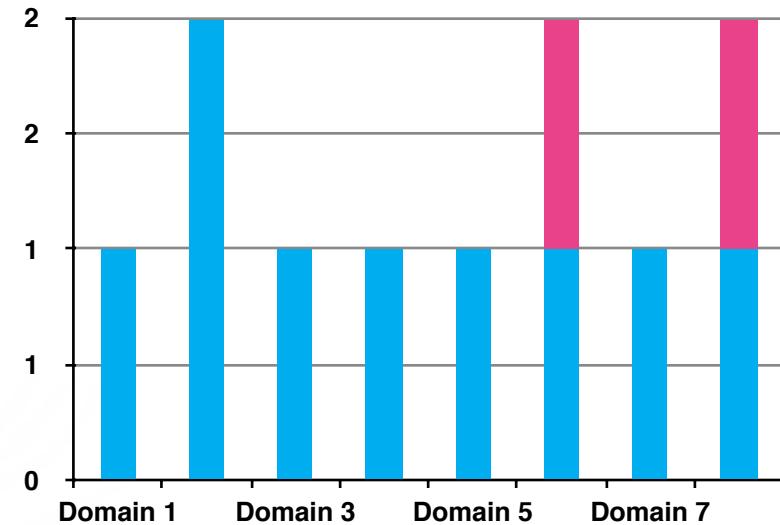
DNS based Agility Analysis

- The metric for agility utilized is Time and Location
- Database includes:
 - Set of selective domains falling under the categories 1) Famous Domains
2) Common Domains 3) CDN's 4) Malicious domains
 - Set of RDNS servers
- Each domain is queried against a RDNS for IP resolution

Extracted Data

- Two main data categories are retrieved
 - The resolved IP for each Domain-DNS pair
 - TTL associated with each IP obtained
- The obtained data is used to calculate the following metrics
 - Number of distinct resolved IP addresses
 - Average Time-to-Live for the domain

Observed Data



Analysis

```
> dig @200.88.127.22 google.com

;; QUESTION SECTION:
;google.com.          IN      A

;; ANSWER SECTION:
google.com.        233    IN      A      190.167.241.183
google.com.        233    IN      A      190.167.241.148
google.com.        233    IN      A      190.167.241.177
google.com.        233    IN      A      190.167.241.152
google.com.        233    IN      A      190.167.241.168

;; Query time: 77 msec
;; SERVER: 200.88.127.22#53(200.88.127.22)
;; WHEN: Thu Nov 20 18:51:40 2014
```

[Querying v4.whois.cymru.com]

[v4.whois.cymru.com]

AS	IP	AS Name
6400	190.167.241.183	CompaÃ±Ãa Dominicana de TelÃ©fonos, C. por A. - CODETEL,DO

Map View Additions Bulk Edits Deletions Print or Share Go to... Map Satellite

Facebook 2.179.65.58 to 10.10.34.0

Esfahan, Esfahan, Iran

Directions From here - To here

1 2 3 4 5 6 7

Name
1 Facebook 2.179.65.58 to 10.10.34.0
2 Google 110.54.254.150 to 188.43.64.0
3 Google 200.40.230.36 to 200.40.0.0
4 Google 200.88.127.22 to 190.166.41.0
5 Google 205.211.206.141 to 186.32.236.0
6 Google 61.31.233.1 to 60.199.175.0
7 Google 70.36.0.5 to 24.222.46.0

2000 km Terms of Use

Page 1 of 1

Conclusion

- Take Away
- Github: <https://github.gatech.edu/agupta402/BlackThread>
- Future Works

Questions