BlackThread

Akshay Gupta, Evan Downing,
Ashwin Chintaluri, Yash Shah,
Liyan Wan, Deepti Kochar,
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

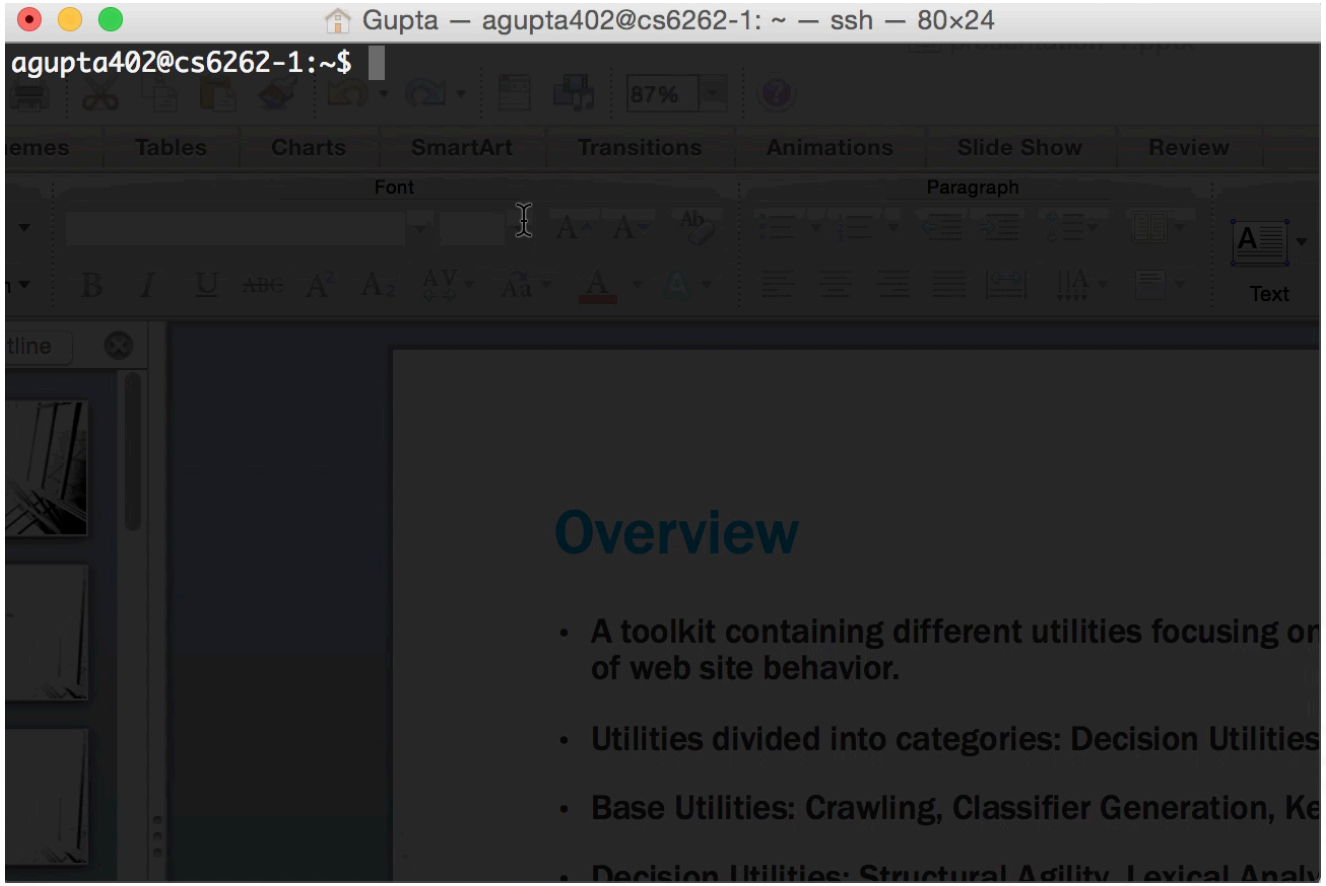
Road Map

- 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

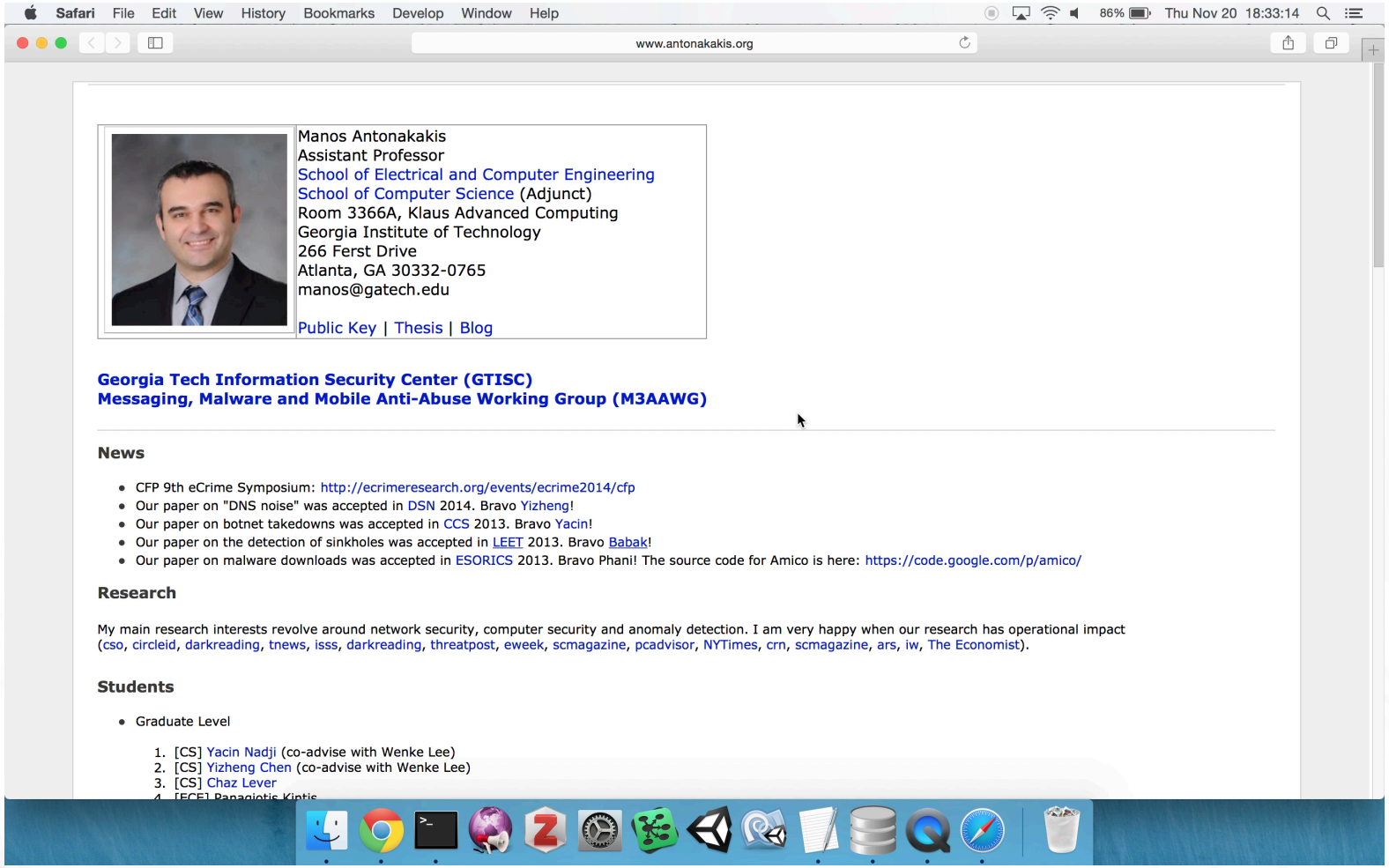
- Crawling - automated visiting of websites
- Scrapy (Python library)
- Our spiders: Tags, Content, iFrames/JS, URLs

Web Structure Agility

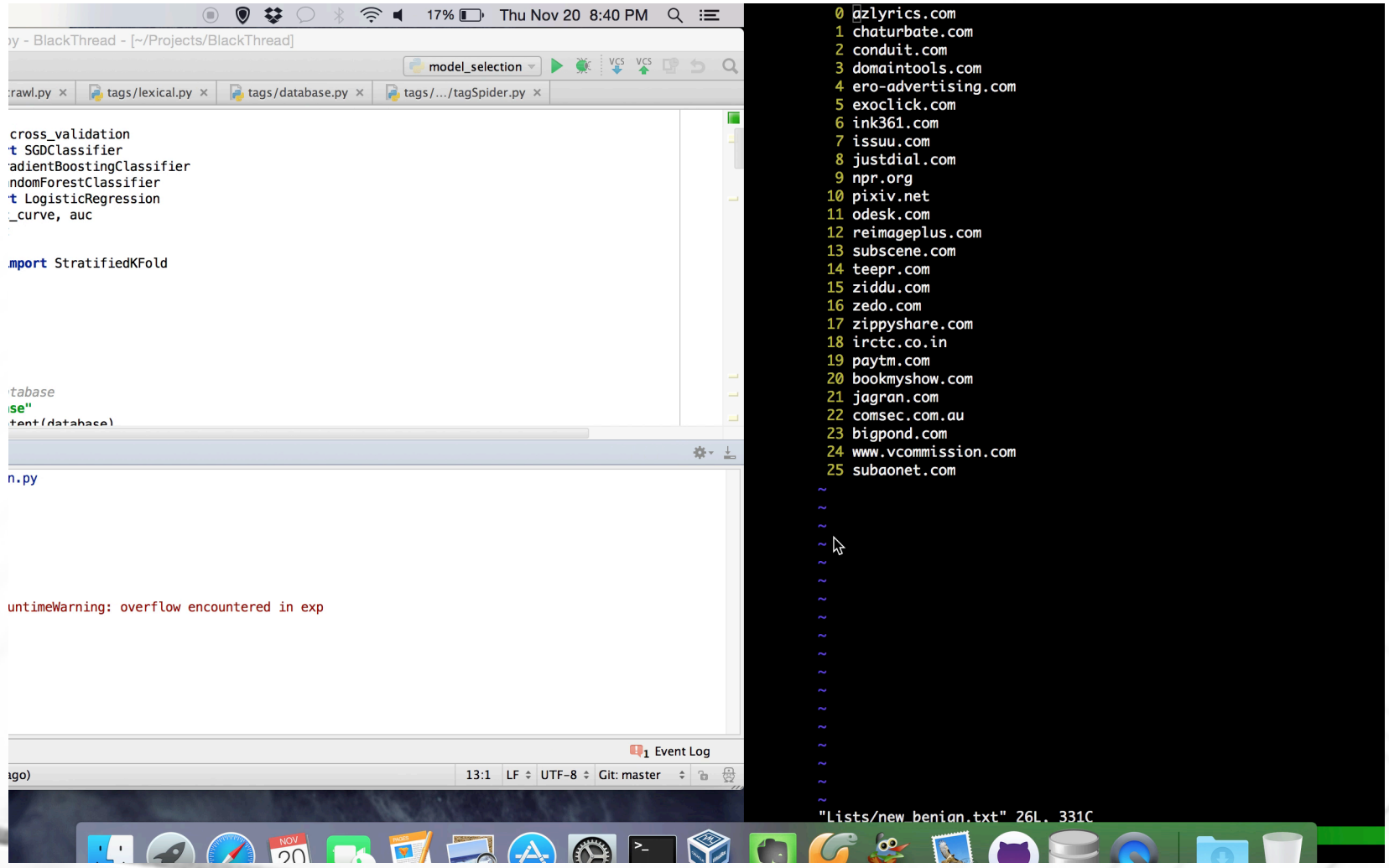
Web Structure Agility - Why

- 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



Demo: Machine Learning



Lexical Analysis

Lexical Analysis - Why

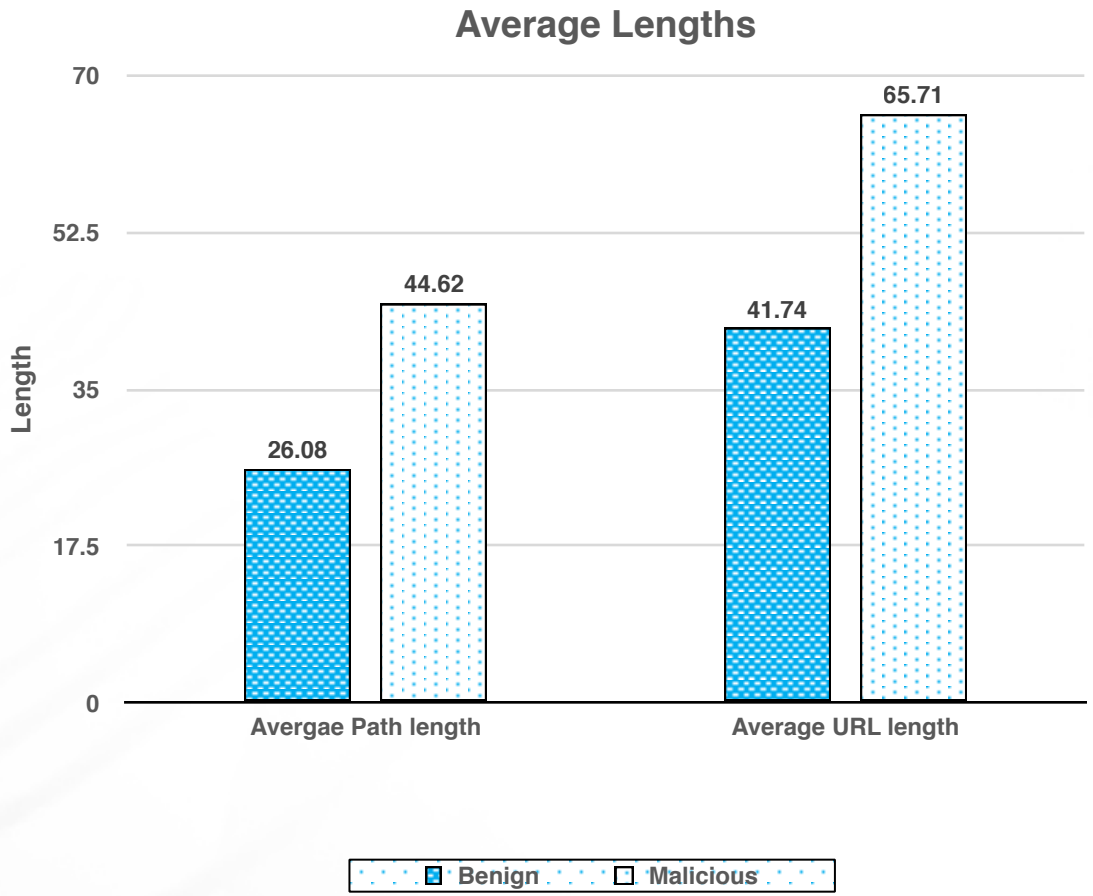
- Blacklisting – Inefficient to tackle the problems of frequently changing domains
- Need something more dynamic!
- URL's can be classified as malicious or benign based on their distinct characteristics.

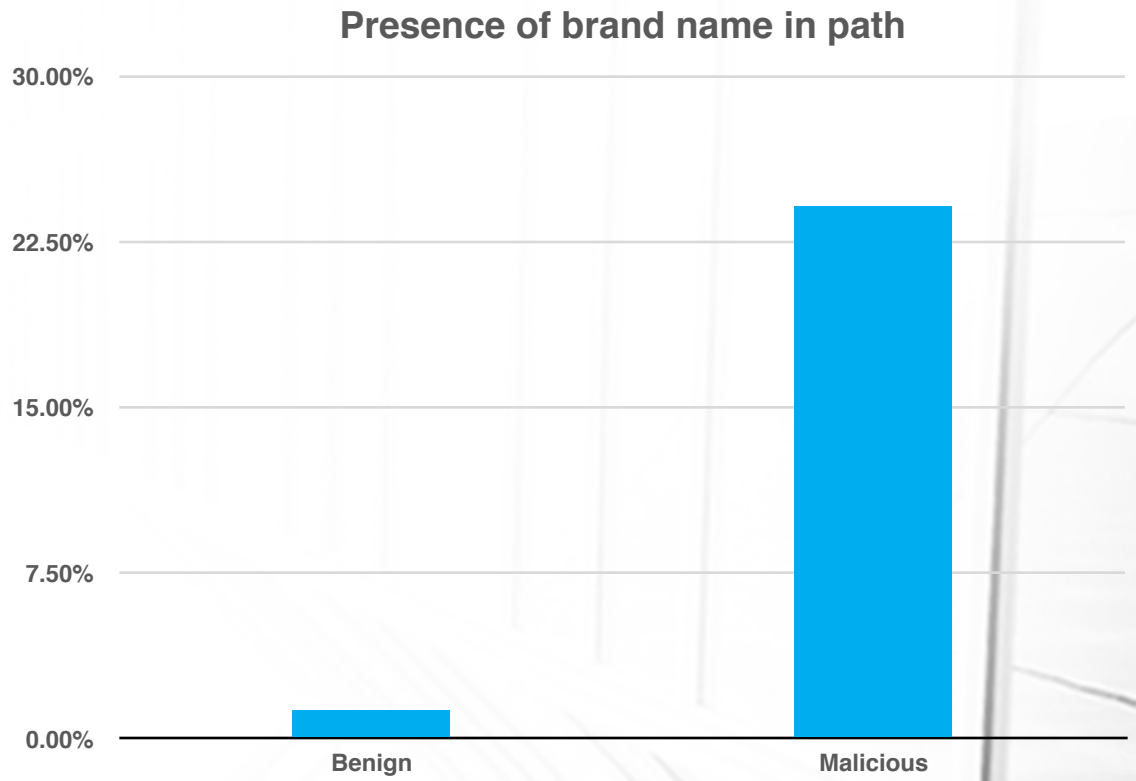
Features extracted from URL

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

http://thaibiznews.net/wp-includes/Text/Diff/Renderer/web/MicrosoftEmailUpgrade.html

Comparison of some URL characteristics





Phishing attack targeting Facebook

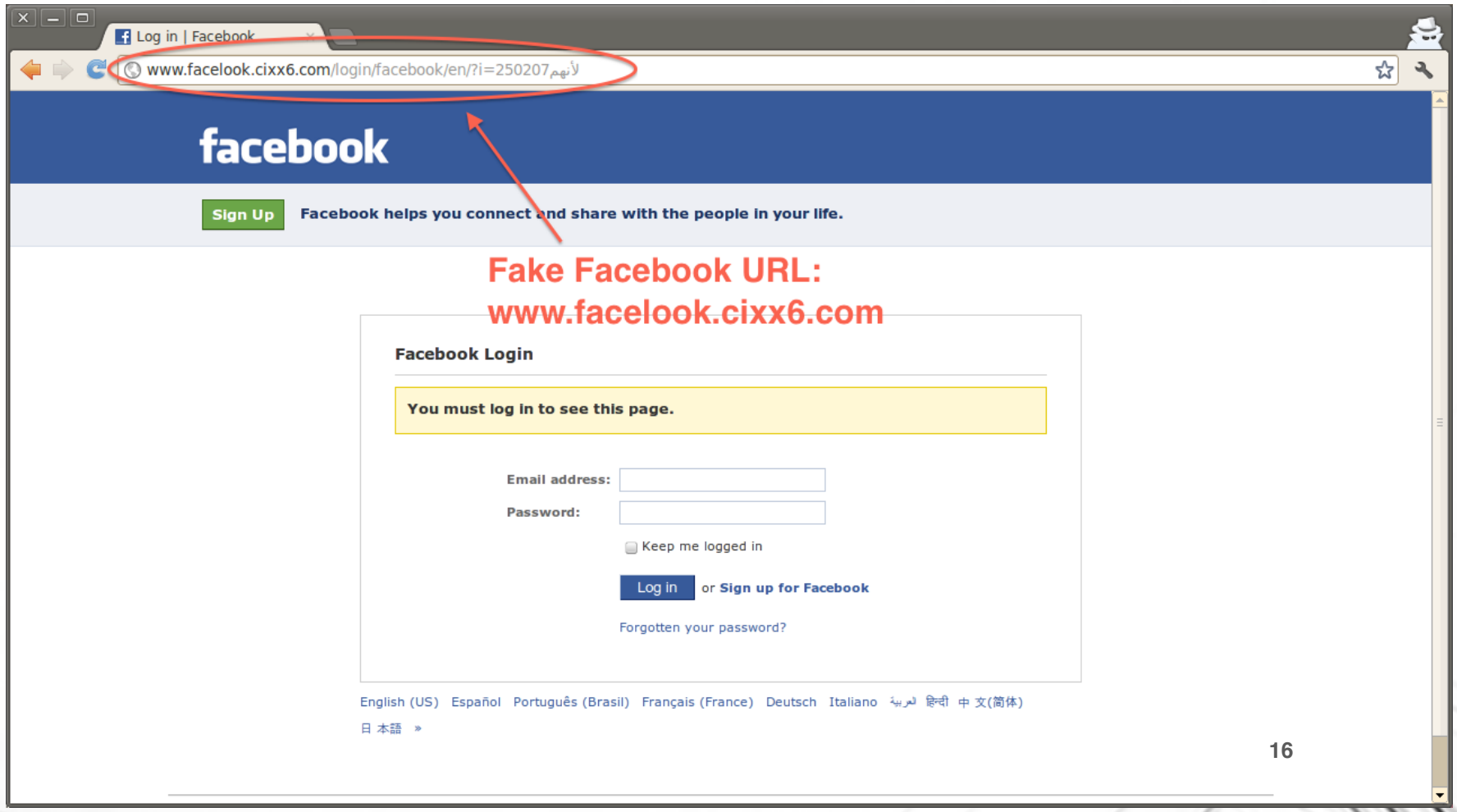


Image from www.eff.org

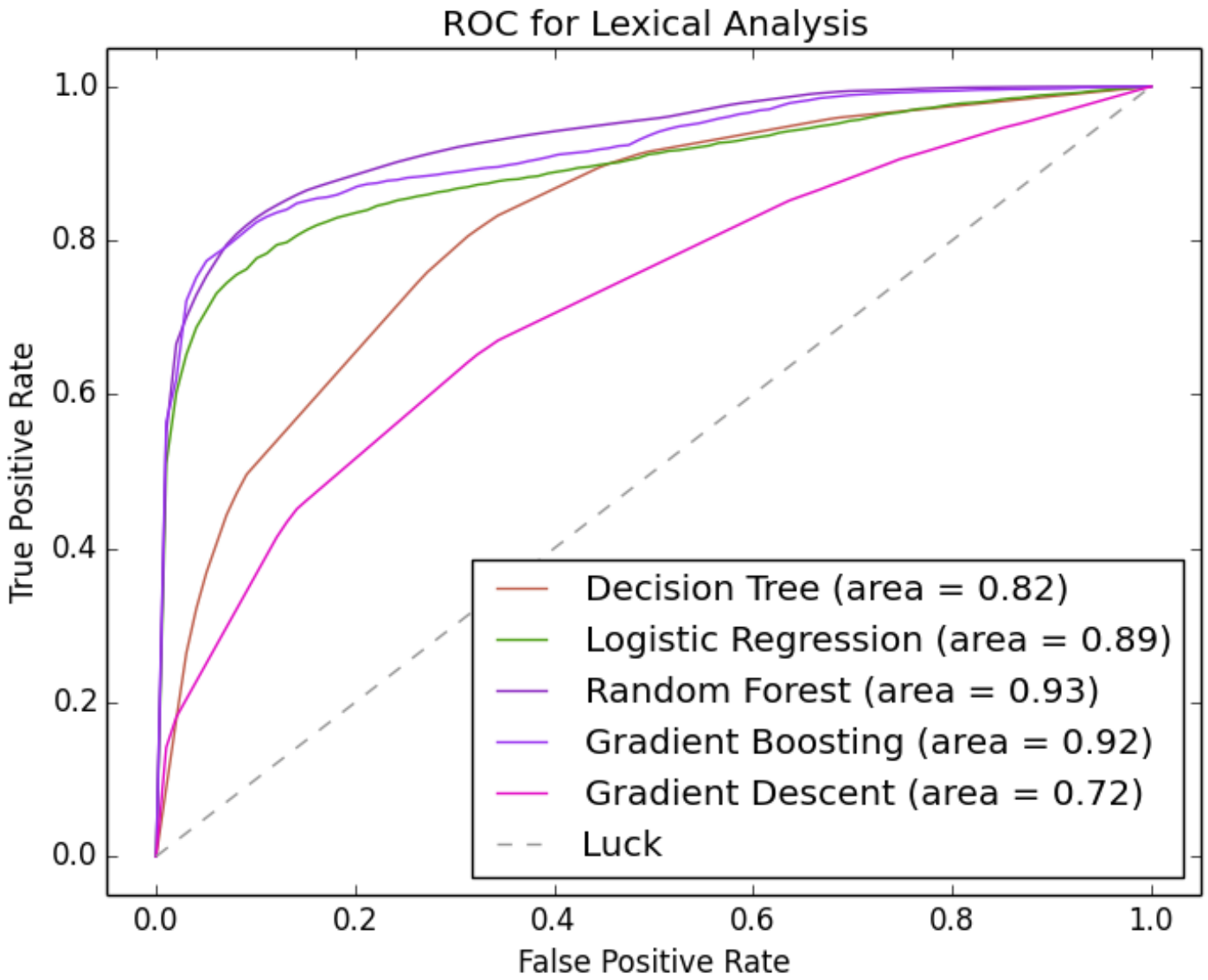
Data Collection and Training

- 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.

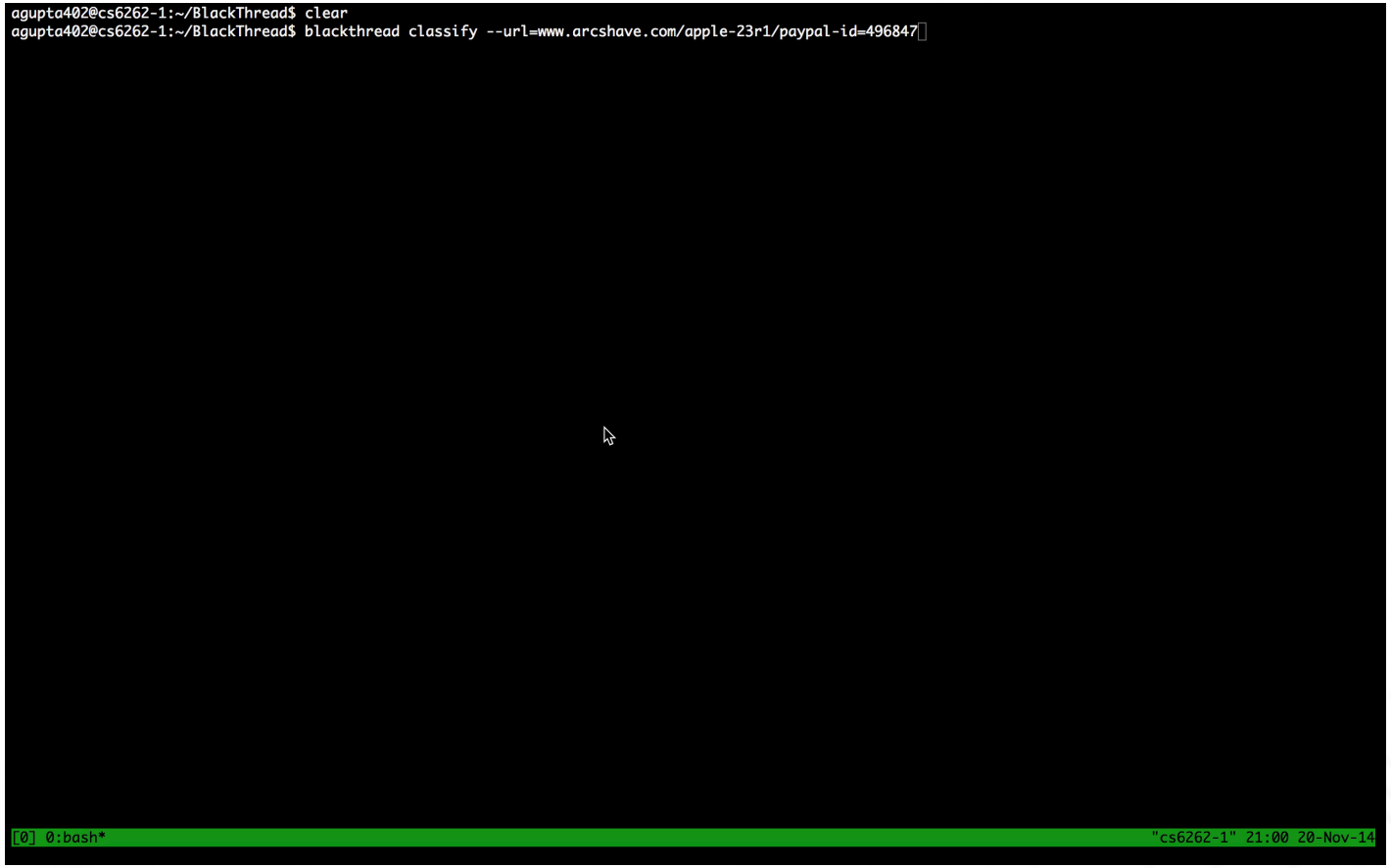
Model Selection: Evaluation Statistics

	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



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

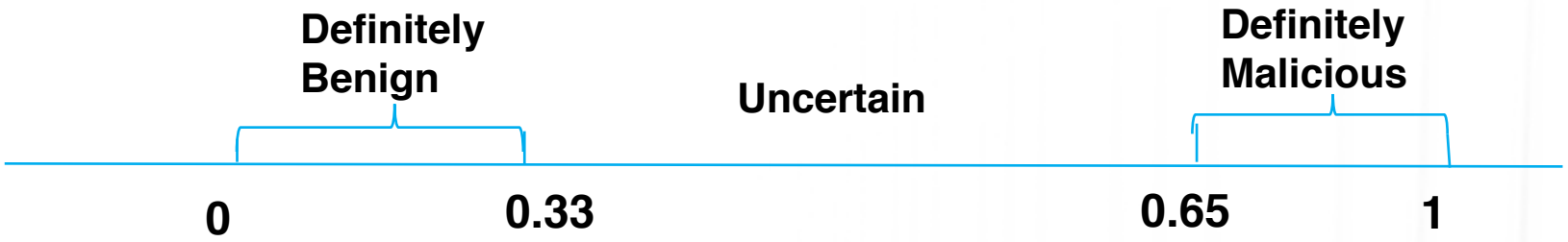
Four Statistical Features

- 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
Benign	*0.70*	*0*	*0.73*	*3.71*	*Max 0.33*
Malicious	*4.14*	*0*	*2.90*	*5.14*	*Min 0.65*
Threshold	*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

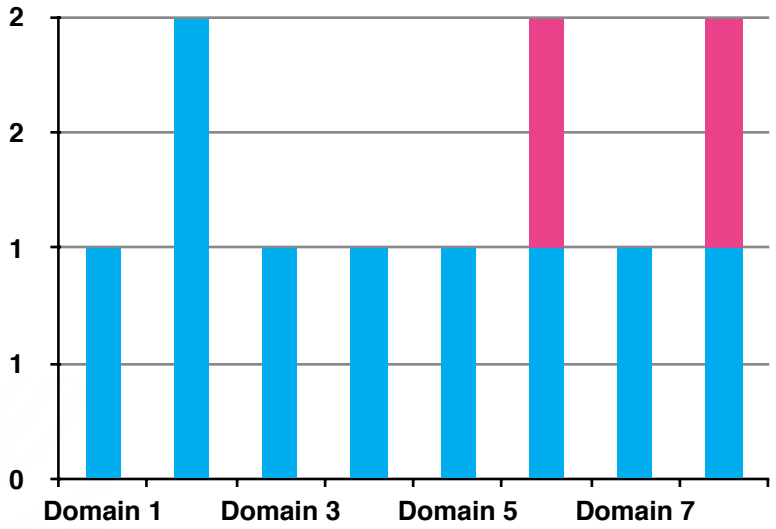
DNS based Agility Analysis

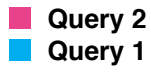
- The metric for agility utilized is Time and Location
- Database includes:
  - Set of selective domains falling under the categories 1) Famous Domains 2) Common Domains 3) CDN's 4) Malicious domains
  - Set of RDNS servers
- Each domain is queried against a RDNS for IP resolution

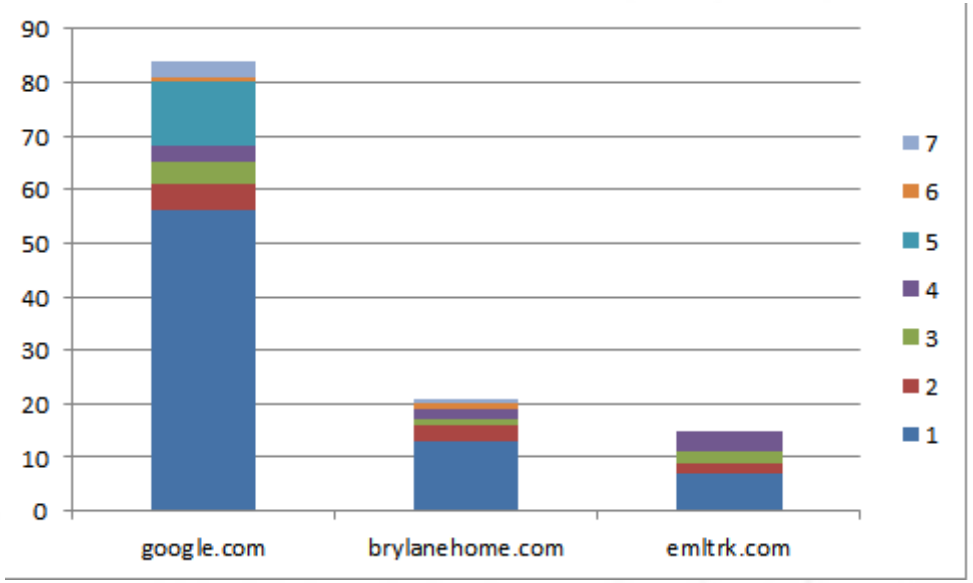
Extracted Data

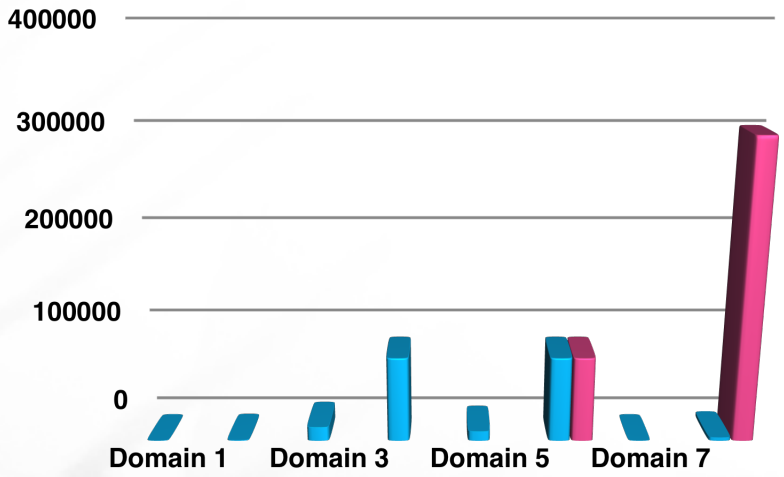
- Two main data categories are retrieved
  - The resolved IP for each Domain-DNS pair
  - TTL associated with each IP obtained
- The obtained data is used to calculate the following metrics
  - Number of distinct resolved IP addresses
  - Average Time-to-Live for the domain

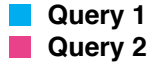
Observed Data

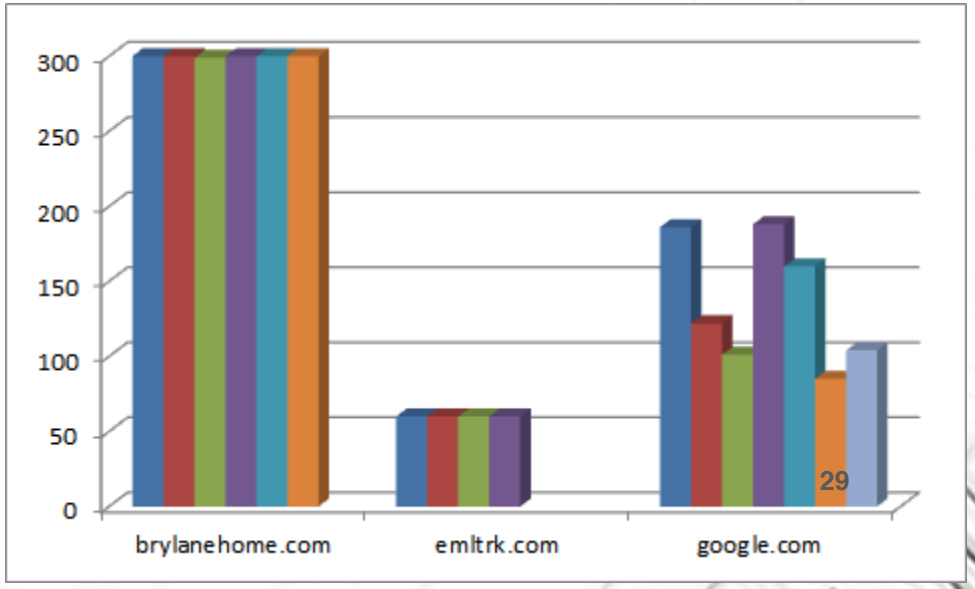












Analysis

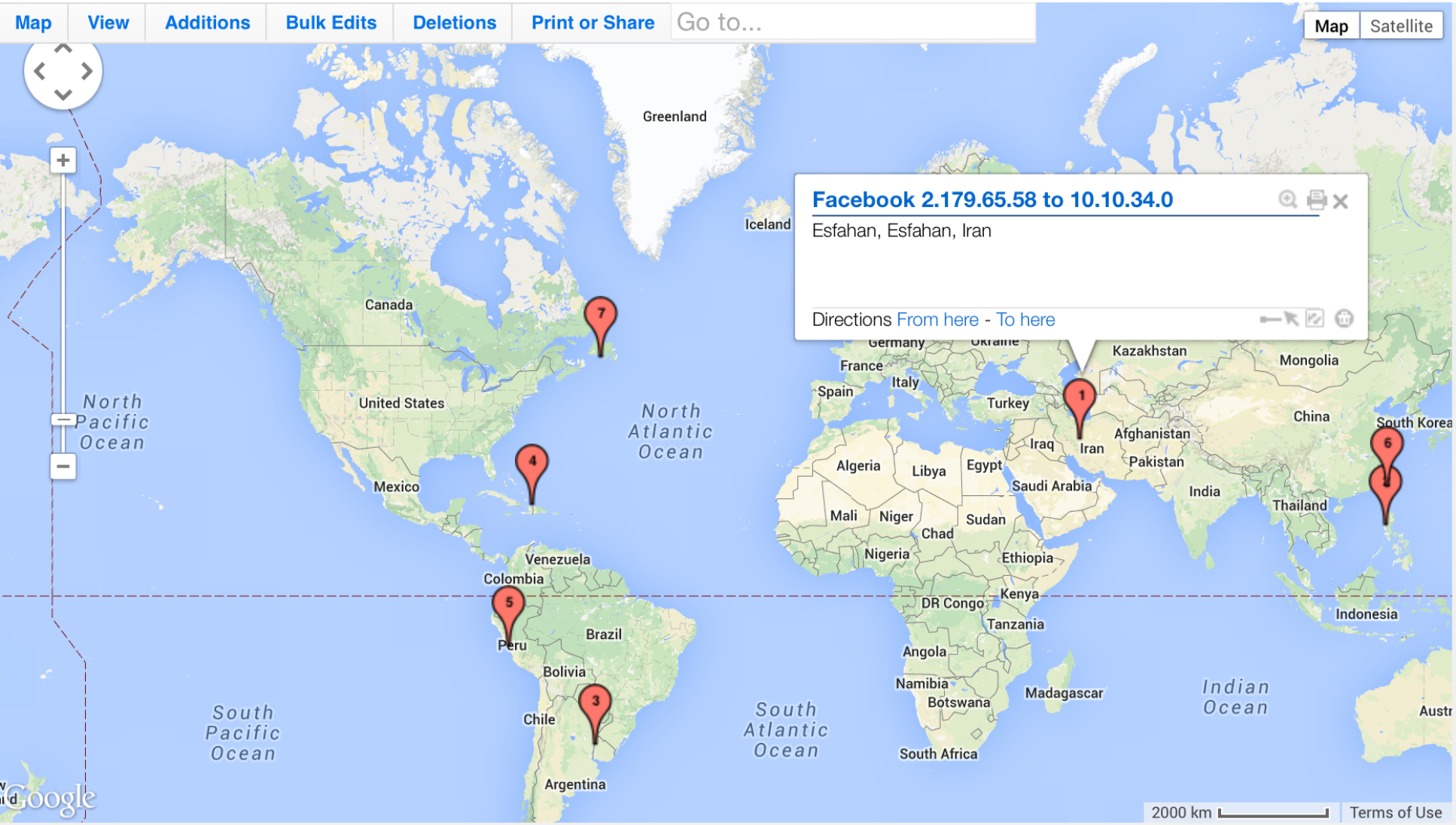
```
> dig @200.88.127.22 google.com

;; 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
```



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

Page 1 of 1

Conclusion

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

Questions