Domain Name Classification via Web Search Results Mining

Tiffany Cauthen and Roberto Perdisci

The University of Georgia

Goals

- Learn the differences between malicious and legitimate domain names.
- Create a tool to classify the legitimacy of unknown domain names.
- Help unsuspecting victims detect malicious domain names and prevent infections.

The Problem

- Domain names are mnemonic addresses to different Internet sites.
- Malicious domains are often used to spread malicious software or remotely control infected machines.
- Legitimate domain names do not host harmful content and provide useful services.
- We want to create technologies that can help us automatically distinguish between malicious and legitimate domain names.

Conclusion, Next Steps

- Accurate identification of malicious domain names by leveraging “Internet sentiment” about domains seems possible.
- Next steps: Collect more data, combine the three classifiers into one, and improve overall accuracy.

Our Approach

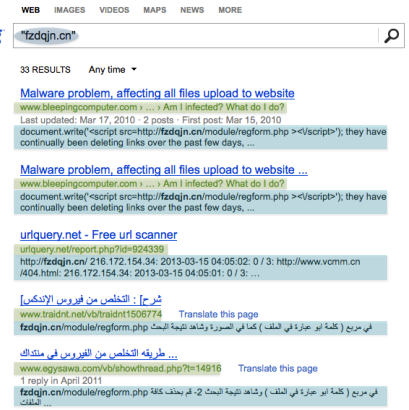
For each domain, we query the domain name string on public search engines to learn the “Internet sentiment.”

We consider the following features:

Domain Results
The hypertext links in the search results

Web Content
We visit each link and grab the related web page content

Domain Name
Syntax-based analysis of queried domain (e.g., length, letter frequency, etc.)



Preliminary Results

Classifier	Correctly Classified Instances	Incorrectly Classified Instances	Weighted Accuracy Average	Confusion Matrix a b < - Classified As w x a = Legitimate y z b = Malicious	
Domain Results	1071 **92.0103 %**	93 7.9897 %	0.92	351	85
				8	720
Web Content	1013 **87.0275 %**	151 12.9725 %	0.87	291	145
				6	722
Domain Names	1148 **97.2058 %**	33 2.7942 %	0.972	726	19
				14	422

{cauthen,perdisci}@cs.uga.edu
University of Georgia
Department of Computer Science

This material is based upon work supported by the National Science Foundation under Grant No. CNS-1149051.