A Study of URL Redirection Indicating Spam

Krishna Bhargrava Vangapandu
Deparment of Computer Science
University of Georgia
415 Boyd GSRC
Athens, GA
bhargav@uga.edu

Douglas Brewer
Deparment of Computer Science
University of Georgia
415 Boyd GSRC
Athens, GA
brewer@cs.uga.edu

Kang Li
Deparment of Computer Science
University of Georgia
415 Boyd GSRC
Athens, GA
kangli@cs.uga.edu

ABSTRACT

The use of URL redirections has been recently studied to filter spam as email and web spammers use redirection to camouflage their web pages. However, many web sites also employ redirection for legitimate reasons such as logging, localization, and load-balancing. While a majority of the studies on URL redirection focused on spam redirection, we provide a holistic view of the use of URL redirections in the Internet. We performed a redirection study on various sets of known legitimate URLs and spam URLs. We observed that URL redirections are widely used with more than 40% of legitimate URLs redirecting for various reasons. We also observed that server side redirection is prominent in both legitimate and spam redirection. Differing from legitimate URL redirection, JavaScript redirection is detected more often in spam URLs. Furthermore, a very high percentage of spam redirections lead to an external domain. We expect that our measurement results and classifications provide a better understanding of the usage of URL redirection, which could help improve spam filtering and other applications that rely on URLs.

1. INTRODUCTION

It is very common to see spam emails contain a URL that directs the user to a website that may try to sell something or steal user information[7]. To combat this, spam filters commonly have blacklist filters for URLs, but redirections make it hard for these filters to function correctly[1, 5, 10, 2]. This means it is common to see spam messages that obscure their website URLs with redirection[9].

Since redirection is a widely used technique, much research has focused on using URL redirections as an important factor in detecting spam. These previous works study URL redirections in spam URLs exclusively. However, many legitimate (non-spam) URLs also employ redirection for reasons such as load balancing, link tracking, and bookmark preservation. To build accurate and effective spam detection based on URL redirections, we need to understand the use of URL redirections for both spam and legitimate reasons.

Our study measures redirections in both legitimate URLs and spam email URLs. We observe that redirection is common in both spam and legitimate URLs. Spam URLs used redirection only slightly more than legitimate URLs at 43.63% and 40.97% respectively. This means that while spam URLs will use redirection, the fact that redirection occurs cannot be a means for determining whether an email should be considered spam or not.

To further our study, we broke redirections into multiple different types and examined whether the redirections were internal or external, whether the redirected domains were owned by the same organization. We found three types of redirections used most often Server-side, Javascript, and Meta. A clear difference between spam and legitimate URLs emerges when we categorize redirections into these types. We find that spam URLs use Javascript redirection much more often than do legitimate URLs 30% compared to 10%. Other types of redirection where used more often by legitimate URLs, but they were not disparate enough with the usage by spam URLs to be of use. Breaking down redirections into external or internal yields a result where external redirects are used more often by spam URLs and internal more often by legitimate URLs.

The Javascript and external redirection properties of spam URL redirection are expected and are a strong signal of spam. When a spammer tries to obscure their URL through redirection, it is preferrential to use a URL from a trusted site to do redirection. Spammers may or may not to able to modify a trusted site by editing files on the webserver. If they cannot edit the files, they may be able to inject Javascript code into the page; this code will redirect to external site where the content is in the spammer's control. If they can edit the files, the flexibility of Javascript redirection means its use is more likely letting a spammer hide the compromise from casual observation by the site admin. Finally, a site may offer a redirection service that can be exploited, leading to an external redirect.

The paper is organized as follows. Section 2 briefly describes the previous work on redirection. In section 3, the types of redirections, reasons for using redirection, how redirections are detected, and classification of redirection is presented. Section 4 describes the experiments conducted while section 5 provides a detailed analysis of the results observed.

2. RELATED WORK

Most of the previous work in the field of web redirections focused on spam redirections. In one of the early works on Web Spam classification, Gyongyi and H. Garcia- Molina [5] describe redirection as a spam hiding technique used by

CEAS 2009 - Sixth Conference on Email and Anti-Spam July 16-17, 2009, Mountain View, California USA

spammers to create doorway pages. Wu and Davison [10] conducted a preliminary study that contributes with a quantitative analysis of the presence of cloaking in the Internet. They look at redirections as one of the techniques to perform cloaking. Our study differs from these by considering not only spam datasets but also a few common categories of legitimate web sites.

JavaScript redirections have been shown to be used by spammers as a way to dupe users into viewing spam. Benczur et al. [2] discovered numerous doorway pages which rely on JavaScript redirection. Chellapilla and Maykov [3] look at JavaScript redirection explicitly with a focus on the techniques employed by the spammers. The Microsoft Strider team in their work on systematic discovery of spammers emphasized URL redirection as a common spam technique. They developed a tool, Strider URL tracer, which can be used to detect all the domains that a current web page connects to. With the aid of the Strider, Wang et al. [8] studied URL redirections in the context that there are content providers which redirect the user to malicious sites. Niu et al. [6] conducted a study on forum spamming with context-based analysis using Strider to identify doorway pages.

To our knowledge, most of the work mentioned above studied redirection in the context of spam. Our approach is different from the above work as we look at general web redirections on the whole. Our study involves detection of URL redirection, classification of detected redirections across multiple dimensions.

3. OVERVIEW OF REDIRECTION

A URL is said to be redirected, if a client requests a resource located at a specific URL, but the client's final destination at the end of the request is a different URL. This section describes the types of redirection techniques as well as the common reasons for redirections.

3.1 Types of Redirection

Based on the implementation techniques, URL redirections are classified into 1) Server-Side redirections, 2) JavaScript redirections, and 3) META redirections.

Server-side redirections — occur when a client requests a resource and the server issues an HTTP response which makes the client request through a different URL. HTTP reponse status codes of type 3xx as well as some 4xx with a location field in the header imply that the client has to redirect to a different URL. For example, request for http://www.google.net returns a status code 302 redirecting the request to http://www.google.com.

JavaScript redirections — are initiated at the client side through statements like
window.location="http://myspamlink.com".
These instructions are inserted into the script sections of the HTML page sent and when the client JavaScript engine executes these statements, it is redirected to the URL specified within the script. However, JavaScript redirection can be obfuscated to a complicated jumble and can make use of a browser's Document Object Model(DOM).

eval("\"moc.knilmapsym//:ptth\"=
noitacol.wodniw".split("").reverse().join(""));

This means that it is preferable to have a browser to detect this type of redirection.

META redirections — are based the META tags located in the HEAD section of the HTML page.

<META content="2;url=http://uga.edu"
http-equiv="refresh">

By setting the associated http-equiv attribute to refresh and the content attribute to a target URL, a redirection would be triggered at the client browser to this target URL.

3.2 Redirection Targets

The classification of redirections into external and internal ones is to validate a commonly held hypothesis. Spam URLs often have more external redirections than the redirections used in legitimate URLs.

The distinction between internal or external redirection is defined by the web domain ownership of the original and target URL. An external redirection is defined as a redirection between two URL domains which are not owned or managed by the same organization.

Detection of external redirections is not a straightforward task. Very often different domains are managed by the same organization. For example, a redirection from the msdn.com domain to the microsoft.com is not considered an external redirection since both these websites are managed by the same organization. Unfortunately, there are no systematic methods to check if two sites are owned by the same organization.

In addition to checking for identical domains, we adopt the following heuristics to differentiate redirections within the same organization and external redirections. For a redirection from the original URL domain X to a target domain Y, we check 1) if one is a sub-domain of the other, 2) if they share a common domain name server, 3) if they are two domains with a common Top-Level- Domain, and 4) if their IP address is in the same Class B range. If any of these checks return a positive result, we consider that X and Y belong to the same organization and to be internal redirects, all other redirections are considered external. Obviously, these heuristics are not always accurate. Therefore, we present the results of classification with and without each of these heuristics.

4. EXPERIMENTS

This section describes the system we used to detect redirections and the datasets used in the study.

4.1 System Description

The system we used to study redirections is comprised of a custom crawler used to collect the datasets, a redirection detector, an external redirection detector, and a heuristics based classifier.

The redirection detector uses SWT Browser Widget [4], a browser component that is commonly used in Javaenabled applications. For example, Eclipse, a popular Java IDE uses SWT Browser component in its in internal web browser. Though the SWT Browser is a GUI widget, our system does not use a graphical user interface and thus does not require user interaction. Because the system uses a real browser, it can accurately detect all known types of redirections.

The redirection detector parses each URL and redirections detected are classified as a server-side redirect or client-side redirect. The URLs containing client-side redirections are

Table 1: Summary of Redirection Detection

Dataset	URL	Total Redirections
Legitimate (Alexa)	107300	40.97%
Legitimate (UGA)	1953	39.07%
Legitimate (Blogs)	8878	25.03%
Spam Dataset	13451	43.63%

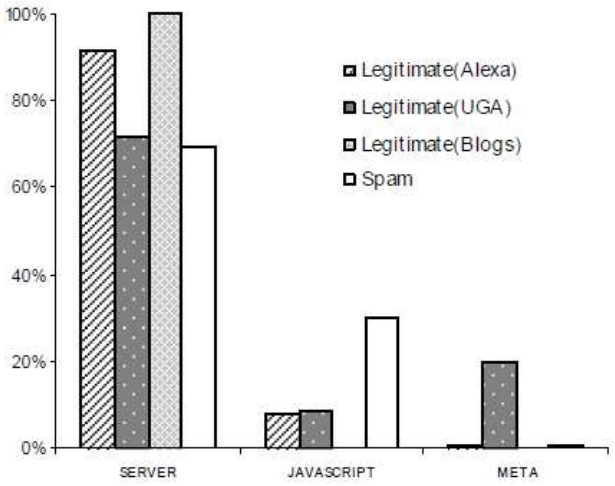

Figure 1: URL Redirection Comparison

further parsed and classified into JavaScript or META redirections. The system also contains a component that can be used to detect external redirections.

4.2 Datasets

In order to study the prominence of redirection, we have considered different types of datasets. These include a spam dataset whos URLs are taken from spam honey pots and three legitimate datasets of different natures.
1)legitimate(Alexa) is a dataset of the Alexa Top 500 website's URLs; 2) legitimate(UGA) include all the top level website's URLs in a class B IP range (128.192.*.*) owned by the University of Georgia, as a representative dataset for web sites in a university; and 3) legitimate(Blog) is a set of blog URLs collected by continuously following the "next blog" link on blogger.com, which randomly return popular blogs from the blogger.com website. Our legitimate datasets are not from emails because a ham corpus with URLs is hard to find.

The selection of the first dataset, legitimate(Alexa), as a representative dataset for legitimate URL redirection is based on the assumption that the Alexa TOP 500 web sites are spam free because they are the most popular Internet sites. We did not validate this assumption.

The selections of the other two datasets are validated by us manually. Manual classification of spam and legitimate URLs often require domain knowledge and the authors are familiar with the URLs on the university servers. Manual testing confirmed that the URLs in this dataset are legitimate.

5. RESULTS

This section describes the measurement results of URL redirections on the datasets described earlier.

5.1 Overview of Redirection Measurements

The overall number of URLs in each dataset as well as the total number of redirections (in percentage) are listed in table 1. Figure 1 provides a further breakdown of these results based on the techniques used to implement redirections.

Overall, we observed that URL redirection is common in all forms of URLs irrespective of the nature of the data set, and we found server-side redirections are predominant. These observations are true for popular Internet sites as indicated by the Alexa dataset, as well as the University and Blog dataset. About 25 to 40 percent of legitimate URLs actually involve redirection. Among them, the observed redirections are mostly server-side redirections. The study on Spam URLs presents a similar result: overall 43.63% of the URLs in the Spam dataset redirected to a different location. Among these, the dominant technology is server-side redirection (69.33%).

Although both legitimate and spam URLs heavily use server side redirection, there is a difference in their use of JavaScript redirections. Among all the datasets, Spam URLs tend to have a larger ratio (over 30%) of using JavaScript redirection than the ratio used by legitimate URLs (less than 10%). This indicates that JavaScript redirection should be more valuable when considering redirection behaviors as indications of spam. Many spam pages are hosted on exploited servers which do not allow server configurations or server-side scripts and JavaScript redirections are preferred over META for their flexibility.

The other significant difference in the results is actually the META redirections among legitimate data set. It turns out the university dataset has a relatively high percentage of META redirection (19.79%) while the other datasets all have very low percentage. META redirection is most likely used when the webmaster does not have control over the server which is common with university hosted web pages. Given detection of META redirects is pretty straight forward, it makes sense that spammers do not employ this technique often and only 0.46% of the spam dataset redirected using META techniques.

5.2 External vs. Internal Redirections

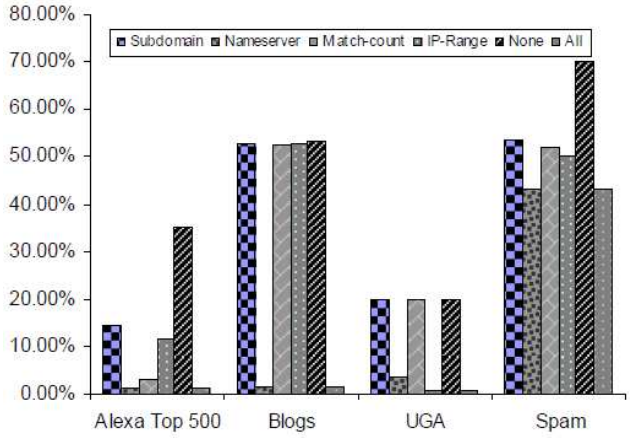

Figure 2: Comparison of Heuristics for External Redirection Detection

We further study the types of redirections based on the target of redirection (external or internal. Given the usage of four heuristics in determining whether a URL redirection goes to external domain or not, we first evaluate the effect of these heuristics. Each heuristic was individually applied and compared to the results where no heuristics were used ("None") and those were all four where used together ("All"). Figure 2 presents the comparison of external redirection detected with the use of each heuristic against the use of all heuristics and none in each dataset. The results indicate that these heuristics help identify actual external redirections more accurately. They confirm a general suspicion that spam web sites redirect externally much of time; as well, they confirm legitimate sites try to keep redirection to within the same domain.

We analyzed the results to see the percentage of redirections that are not within the same domain. It was observed that the amount of external redirection observed in the spam dataset is very high (46.81%) as opposed to that observed in the legitimate datasets (∼2%). Looking at Figure 2, you can see that the "Nameserver" heuristic dominates the classification for internal redirection. It should be acknowledged that it is not necessarily true that URLs in the same/different domain have the same/different domain name servers, but when looking at Top 500 and the Spam dataset, we can conclude that the introduced error is not too significant.

For the spam dataset, as expected, a high percentage of redirections leave to an external domain (46.81%) compared to to legitimate datasets at (0.39%-2.00%). Most of the client-side redirections are observed to be external. 87% of JavaScript redirections and 97% of META redirects leave the current domain and redirect the user to an external domain. On the other hand, only 43% of the Server-Side redirects actually leave the current domain, the majority of these server side redirections are from redirection services such as tinyurl.

6. CONCLUSION

In conducting our study on redirection, we found that the act of redirecting is not itself a good indicator of email spam. It turns out that URLs found in spam and legitimate URLs use redirection with about the same frequency, 43.63% and 40.97% respectively. Since redirection itself is not a good indicator, we studied the types of redirection and whether the redirections were external or internal. We found many good indicators of spam when looking at redirections in these ways.

Redirection when broken down by type, server-side, JavaScript, and meta, shows that spam is indicated by JavaScript redirects. Spam URLs have 30% of their redirections done with Javascript compared to 10% for legitimate URLs. Other types of redirection were used with similar frequency by both spam and legitimate URLs. We believe that Javascript redirection will remain accurate for spam detection because of the flexibility afforded in this type of redirection and possibility of injection into uncompromised hosts.

Redirection broken down into external and internal redirections gives us another good indicator of spam in external redirects. Spam URLs contain external redirection at 46.81% a much greater frequency than legitimate sites at 2%. Spammers will continue to have external redirections with a greater frequency because of inability to compromise the host of a trusted site or because of a redirection service at a trusted site.

We expect that the results of our study provide can benefit the applications that can potentially affected by URL redirections, such as spam filters and other content filters. We continue to refine our classification and detection tool, and we continue our monitoring for the detection of new redirections techniques and the trend of redirection deployments.

7. REFERENCES

[1] Apache Software Foundation and Contributors. *SpamAssassin URL Blacklist Configuration*. http://spamassassin.apache.org/full/3.2.x/doc/Mail_SpamAssassin_F

[2] Benczur, A., Csalogany, K., Sarlas, T., and Uher, M. Spamrank - fully automatic link spam detection. In *First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)* (2005).

[3] Chellapilla, K., and Maykov, A. A taxonomy of javascript redirection spam. In *AIRWeb '07: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web* (New York, NY, USA, 2007), ACM, pp. 81–88.

[4] Cornu, C. *Viewing HTML Pages with SWT Browser widget*. IBM OTI Labs, http://eclipse.org/articles/Article-SWT-browserwidget/browser.html.

[5] Gyongyi, Z., and Garcia-Molina, H. Web spam taxonomy. In *In First International Workshop on Adversarial Information Retrieval on the Web* (2005).

[6] Niu, Y., Wang, Y., Chen, H., Ma, M., and Hsu, F. A quantitative study of forum spamming using context based analysis. In *Proceedings of the 14th Annual Network and Distributed System Security Symposium (NDSS)* (2007).

[7] Pu, C., and Webb, S. Observed trends in spam construction techniques: a case study of spam evolution. In *Third Conference on Email and Anti-Spam, CEAS 2006* (2006).

[8] Wang, Y., Beck, D., Jiang, X., and Roussev, R. Automated web patrol with strider honeymonkeys: Finding web sites that exploit browser vulnerabilities. In *Network and Distributed System Security (NDSS) Symposium* (2006).

[9] Webb, S., Caverlee, J., and Pu, C. Introducing the webb spam corpus: Using email spam to identify web spam automatically. In *Proceedings of the 3rd Conference on Email and Anti-Spam (CEAS)*.

[10] Wu, B., and Davison, B. D. Cloaking and redirection: A preliminary study, 2005.