A Measurement Study of Web Redirections in the Internet

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 URLs that includes known legitimate and spam websites. We observed that URL redirections are widely used in today's Internet 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 websites. Furthermore, a very high percentage of spam redirections lead to an external domain. Apart from providing a quantitative view of URL redirections, we also provide a further classification of legitimate URL redirection based on the reason of redirection. We expect that our measurement results and classifications to provide a better understanding of the usage of URL redirection, which could help improving the spam filtering and other applications that rely on URLs as the web identifiers.

1. INTRODUCTION

The point of most email spam is to sell something to recipient or steal personal information. This means that the spam email must contain some way for the recipient to purchase what is being peddled or trick the user into providing their information. To this end, it is very common to see spam emails contain a URL that directs the user to a website[?]. In spam filters, it is common to have blacklist filters for URLs, but redirections make it hard for these filters to function correctly[?][?][?][?]. This means it is common to see spam messages that obscure their website URLs with redirection[?].

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, that 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 down types of redirections 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 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 [?] describe redirection as a spam hiding technique used by spammers to create doorway pages. Wu and Davison [?] 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 consider not only spam datasets but also a few common categories of

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

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. [?] discovered numerous doorway pages which rely on JavaScript redirection. Chellapilla and Maykov [?] 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. [?] studied URL redirections in the context that there are content providers which redirect the user to malicious sites. Niu et al. [?] conducted a study on forum spamming with context-based analysis using the 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 — Server-side redirection occurs when a client requests a resource and the server issues a directive in the form of HTTP status codes which makes the client request through a different URL. HTTP reply 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.

Redirections can also be performed by using publicly available services redirection services (such as tinyurl,shorturl). We have also observed that some other services are exploited by spammers to perform redirection even though they are not exactly meant to generate redirections. An example for this service is the "Google, I'm feeling lucky" which redirects the user to the first query result.

JavaScript redirections — JavaScript redirections are initiated at the client side through statements like ‵window.location=someurl′. 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. Unlike server-side redirections, a web page is actually loaded partially or completely into the client browser before the redirection occurs.

META redirections — Another client side redirection is based the META tags located in the HEAD section of the HTML page. 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. The redirection happens after a browser finishes parsing a HTML page, then the META refresh action is triggered to load content from the target URL.

3.2 Reasons for Using Redirection

As mentioned earlier, there are several reasons for URL redirection – both legitimate and illegitimate. In this section we look at some of the most common reasons for employing redirection. The order of appearance is based on the popularity among the classification results with most commonly observed reason listed first.

Virtual hosting (and DNS aliasing) — Virtual Hosting and DNS aliasing are where more than one sites (or just domain names) are mapped to a single IP address. This is commonly used by websites to register most commonly misspelled domains and redirect requests to these domains to the original server. Such organizations register the same domain name with different top-level domains. For example, requests to gooogle.com or google.net all redirect the user to http://www.google.com.

Load balancing — Most of the top websites host content on several servers. These servers either host specialized content or mirror each other. In cases of high volume of web traffic, requests are redirected to one of these servers; the criteria of which depends on the website. For example, popular websites like search engines host different mirror servers and requests are redirected based on the nature of resource requested.

Link tracking (via indirections) — Many websites use redirection for statistical and logging reasons. For example, websites log the advertisement clicks before it actually takes the user to the advertised webpage. For this to happen, advertisement clicks are taken to the originating website where the information is logged and then the request is redirected to the advertised webpage.

Resisting web spam (via indirections) — Many websites rewrite external links in their web pages by introducing a level of indirection through a server that is not indexed by search engines. For example, all user links posted at MySpace are disguised as a link with indirection from the domain name msplinks.com, not myspace.com. A web page linked from the former domain is much less valuable than a link from the latter one.

Providing Warnings (via indirections) — A level of indirection is also used to provide a warning to users when they are about to leave the current domain.

Security — Unauthenticated or unauthorized requests to a resource are usually redirected to a login page or an information page. For example, visiting a protected profile on MySpace redirects unauthenticated users to the login page. Similarly, transition from HTTP to HTTPS is often enabled through redirection. For example, a request to http://www.bankofamerica.com would be redirected to https://www.bankofamerica.com/index.jsp.

URL rewriting — Some websites rewrite long and script oriented URLs with short and user-friendly URLs and later redirect to the actual URL.

Other reasons — Redirection is also used for many other reasons, such as to keep bookmarks, to route requests in CDN, to redirect requests for invalid resources, and to provide personalized content based on client geolocations or user-agent types. For example, we observed that 15 out of

Alexa top 500 websites redirected differently when the user-agent field are changed from a popular browser (Firefox) and a crawler (Google-bot).

Excepting the legitimate usage, URL redirection is one of the most exploited techniques for spamming, which have been used for cloaking, domain forwarding, doorway pages. For details of the use of redirections in spam, please refer to early work on spam redirections [?][?][?]

3.3 Detecting and Classifying Redirection

This section describes our efforts of classifying URL redirections into several categories based on the implementation techniques (Server-side, Javascript, or META), the target of redirections (external vs internal), and the reasons of redirection.

3.3.1 Redirection Techniques

Classification based on the type of redirection techniques helps us identify the most commonly used techniques in various types of web sites. We briefly overview the required efforts to detect and classify redirections based on their implementation techniques. The detection of redirections can be achieved with or without a browser. Obviously, the detection of redirection techniques can be achieved by a browser. For example, the Microsoft Strider URL Tracer is useful in detecting redirections from a URL by loading the page into Internet Explorer. Browser-based approaches often have high overhead and thus non-browser based approaches have also been used. While detecting server-side and META redirections are relatively straight forward without a fullfeatured browser, JavaScript redirections are complicated to detect. Standalone Javascript interpreters, such as the Rhino JavaScript engine [?], have been used but they do not capture all intended JavaScript executions. We build our own detection and classification tool based on the SWT Browser Widget. We did not use other browserbased tool, such as Strider, because they do not provide a full classification function.

Detecting Server-side redirection — The detection of Server-side redirections can be easily achieved by monitoring the status code returned in HTTP responses. Redirection occurs when the status code returned is of type 3xx or 4xx with the location property in the response is set to a URL.

Detecting META redirection — META redirections can be detected by looking for META tags with the attribute http-equiv sets to refresh and the content attributes sets to a different URL. An example of such a META tag is shown below:

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

Detecting JavaScript Redirections — The detection of JavaScript redirections is not straightforward because of the possible level of obfuscation [?]. For example, shown below is a simple JavaScript statement that initiates redirection:

window.location="http://myspamlink.com"

Given the amount of flexibility that JavaScript offers to the web developers, the same statement can be obfuscated to many different versions that are hard to parse without executing the script. For example, the above mentioned script could be written as

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

For this reason, a detector should be capable of executing JavaScript and using a web browser is the best way in which one could detect all complex redirections.

3.3.2 Redirection Targets

The classification of redirections into external and internal ones is to validate a commonly held hypothesis, which considers that redirections at spam websites often have more external redirections than the redirections used in legitimate sites.

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 as 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. Obviously, these heuristics are not always accurate. Therefore, we present the results of classification with and without each of these heuristics. And in this study, if a redirection is not considered as external, it is considered as internal.

3.3.3 Reasons for Redirection

Classifying redirections based on reasons of redirection gives us an in depth understanding of the usage of URL redirections in various web sites. Since there is no precise information to determine the reason for using redirections, we use the following methods to infer and classifying them into categories. First, we manually inspect a small set of URL redirections in the legitimate dataset, and infer the motivations of redirections based on the content of original and target URL and web content, and the naming convention of the URLs. Second, we extract out keywords that we believe represent each category and build our heuristic classifier based on these keywords. For example, if keywords such as "login", "auth", and "https" appear in the destination URL, then the redirection is classified under Security category. Keywords such as "ad", "click", "overture" and "adtmt" classify the redirection into Advertisements (Link Tracking) category. We apply these heuristics with the classifier to the rest of the URL redirections and randomly sample the outcome for manual verification of the classification results. Throughout our study, the Keyword based URL heuristics are applied to identify redirections that are for link tracking, security, directory listings, etc.

Apart from the keyword heuristics, we also use other techniques to identify motivations such as load balancing and

Figure 1: URL Redirection Comparison

virtual hosting. These are determined by matching the domain names in the target and destination URL. For example, there is a close match between the domain www0.shopping.com and www1.shopping.com. The difference lies in the first part of the domain name with indications of a possible use of load balancing servers.

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 [**?**], 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 further parsed and classified into JavaScript or META redirections. The system also contains a component that can be used to detect external redirections. The internal and external redirections are further processed by a classifier component which uses URL heuristics to logically classify the redirection based on the motivation.

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 websites; 2) legitimate(UGA) include all the top level websites hosted 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 pages collected by continuously following the "next blog" link on blogger.com, which randomly return popular blogs from the blogger.com website.

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 websites often require domain knowledge and the authors are familiar with the websites hosted on the university servers. Manual testing confirmed that the URLs in this dataset are legitimate.

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%

Table 2: Impact of Popularity on Redirection

Dataset	% Redirections
Alexa Top 100	18.00%
Alexa Top 101-200	17.00%
Alexa Top 201-300	22.00%
Alexa Top 301-400	23.00%
Alexa Top 401-500	17.00%

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 websites 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. Study on Spam sites presents a similar result: overall 43.63% of the URL 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 sites heavily use server side redirection, there is a difference in their use of JavaScript redirections. Among all the datasets, Spam sites tend to have a larger ratio (over 30%) of using JavaScript redirection than the ratio used by legitimate sites (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 Impact of Popularity and Depth

Table 3: Impact of Depth on Redireciton in Alexa Top 100

Alexa Top 100 Level	% Redirections
Level 1	18.00%
Level 2	38.05%
Level 3	41.47%

Figure 2: Comparison of Heuristics for External Redirection Detection

While the overall results demonstrate the wide use of redirection in legitimate sites, it is still interesting to see whether the use of redirections have any direct connections with the web site popularities and the depth of the URL on a web site.

To study the impact and depth, we collected top level (Level-1) URLs from the Alexa Top 100 to Alexa Top 500 websites, and we crawl to different level of Alexa Top 100 sites. It has been studied [?] that crawling websites up to 3-5 levels is sufficient to cover over 90% of the pages linked. Therefore, for Alexa Top 100 websites, the first three levels of pages and those linked to them are crawled.

The breakdown of percentage of redirections observed across the Alexa Top 500 websites (Level-1 pages) is shown in the Table 2. From the results, it can be observed that the popularity of a website has no significant impact on the amount of redirection observed. In spite of 2-3% variation, each sector is close the average amount of redirection, 19.40%

We further conducted experiments on the first three levels of Alexa Top 100 Websites and the amount of redirection observed is shown in the Table 3. From the results obtained, it appears that as one goes deeper into the websites, the amount of redirection increases. This result meets our expectation as the deeper a web site goes, the more possible reasons of redirections apply (see Section 3.2).

5.3 Study on External Redirections

We further study the types of redirections based on the target of redirection (external or internal) and the reasons for the redirection. 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%). The complete classification of the redirections into internal and external is shown in Table 4. Looking at the table you can see that the "Nameserver" heuristic dominates the classification for internal redirection. It should be acknowledged that it is not necessarily true that websites 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.

The amount of external redirection in the legitimate URL (0.39% - 2.00%) is observed to be less compared to the spam dataset (46.81%). Advertisement generated revenue is significant in the Internet and it is very common for the top websites to host advertisements. This is consistent with our measured result, where 53.10% of external redirections were to track advertisement clicks to help keep track of advertisement revenue.

For the spam dataset, as expected, a high percentage of redirections leave to an external domain (46.81%). 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.

Table 4: External vs. Internal Redirects

Type	SPAM	LEGITIMATE (Alexa)		
External Redirection	46.81%	2%	Link Tracking (Advertisements)	53.10%
			Indirection (external links)	37.32%
			Indirection (anti-spam)	2.95%
			Security	0.64%
			Others	5.83%
Internal Redirection	53.19%	98%	Virtual Hosting	59.86%
			Load Balancing	34.07%
			URL Rewriting	1.3%
			Security	1.28%
			Others	4.07%

5.4 Study on Internal Redirections

We further classified the internal redirections of the Alexa dataset based on the reasons for redirection. The dominant internal redirections are primarily virtual hosting and load balancing. We observed that Virtual hosting redirections account for 59.86% of the total observed internal redirections and load balancing accounts for 34.07%, followed by small percentage of URL redirection for the purpose of rewriting and security.

This result matches with our expectation. Since Alexa top sites are popular sites based on the number of user accesses, all of these sites would try their best to attract users. Naturally, these sites would have multiple names to attract users, and to handle high volume of web access, multiple servers are likely to be used with load balancing applied.

6. CONCLUSION

In conducting our study on redirection, we found that the act of redirecting is not itself a good indicator of a spam URL. It turns out that 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 URLs have a tendency to use more Javascript redirects 30% compared to 10% for legitimate URLs. Other types of redirection were used with similar frequency by both spam and legitimate URLs. This means that for detecting spam URLs, methods should only rely and must able to execute Javascript when looking at just redirection.

Redirection broken down into external and internal redirections gives us a good separation of spam URLs and legitimate URLs. Spam URLs contain external redirection, 46.81%, with a much greater frequency than legitimate sites, 2%. As such, a detection method using the heuristics presented in this paper can be fairly sure a external redirection is a mark of spam.

We expect that the results of our study provide can benefit the applications that can potentially affected by URL redirections, such as search engine and content filter. We are refining our classification and detection tool and continuing our monitoring for the detection of new redirections techniques and the trend of redirection deployments.