Cyber-Fraud is One Typo Away

Anirban Banerjee, Dhiman Barman, Michalis Faloutsos and Laxmi N. Bhuyan
Department of Computer Science and Engineering
University of California, Riverside
CA 92521
Email: anirban, dhiman, michalis, bhuyan@cs.ucr.edu

***Abstract*— Spelling errors when typing a URL can be exploited by website-squatters: users are led to *phony* sites in a phenomenon we call parasitic URL naming. These phony sites imitate popular websites and try to extract personal information from unsuspecting users, or simply advertise and sell products to users. In this paper, we conduct a massive study in order to quantify the extent of this parasitic URL naming We start with a corpus of 900 popular websites, which we refer to as *original URLs*, and generate roughly 3 million URLs by varying the original names systematically and exhaustively. Over a period of 60 days, we analyze how many sites have URLs very similar to our original URLs. We find that parasitic URL naming is a wide-spread problem and quantify the extent of this issue. We believe that this work will provide the first step towards research and tools to combat web-fraud.**

I. Introduction

Impersonation and deception is rampant [1]–[11] in the Internet and is one of the primary modus operandi for phishers [5], [6], pharmers [12], [13] and DNS squatters [14] to engineer complex scams. These unsavory entities use a plethora of mechanisms to fool users by decking up their *phony* sites with pictures and text which closely resemble popular sites. Worst case scenarios can range from unsuspecting users entering email and social-network passwords to credit card and social security numbers [2], [6], [8], [12], [13].

Pharming or *parasitic URL naming* is a relatively new but serious phenomenon [12], [13]. Pharmers, which we also call *URL poachers*, register domain names similar to prominent websites and expect to take advantage of users' mistyping of the URL address. Once at these fake sites, URL poachers attempt to advertise and sell products or to glean personal information off unsuspecting users. We will call these websites **phony** and the whole process **URL poaching**. Clearly, using bookmarks eliminates the problem of typos, but typing still takes place in many cases, such as when visiting a new website or using borrowed or public computers. Part of the proof is the great extent of URL poaching, as we will see later. For example, samachar.com is a popular news portal, which when mistyped as samchar.com opens up an adult site. In an effort to address this problem, popular sites often buy URLs which are similar to their own URL For example, gogle.com leads to google.com. Unfortunately, this approach can exacerbate the problem since it indirectly encourages URL poaching and URL squatting: it motivates people to register URL names resembling popular URLs and hope that they will be bought.

This work was supported by NSF NETS 22063 and NSF IDM 0208950 and a CISCO URP grant.

URL poaching is an important and enabling component of the larger cyber-fraud problem, which it can be loosely defined to include a range of activities, from annoying behaviors, like pop-up windows and spam, to identity theft. Extensive studies performed by the Gartner Group in 2004 [7], put a cost of Internet-based ID theft around $2.4 billion per year in the US alone, and report that around 5% of adult American Internet users are successfully targeted by such attacks each year. In fact, a study by Garfinkel and Miller [10] indicates the (high) degree to which users are willing to ignore the presence or absence of the SSL lock icon when making a security-related decision; and how the name and context of the sender of an email in many cases matter more (to a recipient determining its validity) than the email address of the sender. This is a natural motivation to study the security issues which arise due to such browsing habits.

URL poaching has not been studied extensively.

To the best of our knowledge, this work is one of the first extensive studies of characteristics of parasitic URL naming. Previous research has focused on quantifying the impact of cyber-fraud [4], [5], [7], [34] on end-users and on some aspects of parasitic URL naming problem. We shall discuss related research in the next section. In this paper, we conduct a massive study whose goal is to quantify the extent of parasitic URL naming.

We start with a corpus of 900 popular websites, which we refer to as *original URLs*, and generate roughly 3 million URLs by varying the original website names systematically and exhaustively. Each of these similarly named sites is either a phony site, or a legitimate site that happens to be similar, which we refer to as *incidentally similar (IS)* website.

Our main results can be summarized in the following points.

Quantifying the extent of parasitic URL naming.

Based on our measurements, we observe the following interesting characteristics of this problem.

- **Parasitic URL naming is very prevalent.** We find that for nearly 57% of all original URLs in our corpus, more than 35% of all possible URL variations (for each original URL) exist in the Internet. Surprisingly, over 99% of these similarly named websites are phony.
- **One-character variations are most popular for phony URLs:** We find that 99% of phony URLs per original

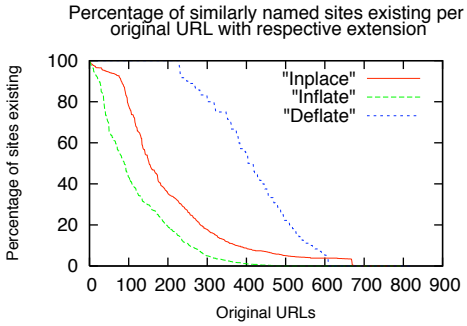


Fig. 1. Percentages of all similarly named sites existing, per original URL, with their respective extensions. Inplace represents percentage of all possible sites existing which have URLs different from original URLs by just 1 character. Similarly inflate represents percentage of all possible sites existing which have URL length greater by 1 character from original URLs and deflate represents percentage of all possible sites existing which have URL length less than 1 character from original URLs.

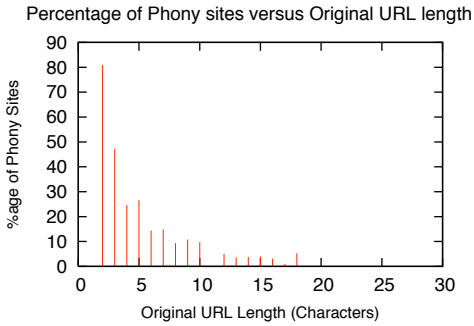


Fig. 2. Percentage of phony sites existing versus length of original URL.

URL, differ from the original ones by just one character, in length or in spelling. Further, URLs with less than 10 characters are more prone to being impersonated. Finally, URLs which belong to US and German banks suffer most from parasitic URL naming.

- **Effect of Original URL-extension on extent of URL Poaching:** Original URLs with a .com extension are impersonated by phony URLs with .biz, .net and .org extensions. Original URLs without a .com extension are impersonated by phony URLs with .com, .net and .org extensions.

The next section presents related literature, while section III deals with quantifying the extent of URL poaching followed by the conclusion.

II. Related Work

Efforts attempting to characterize the parasitic URL naming problem have been limited to identifying some "typo squatters" [14], [34] but have not been able to develop a profile. Experimental studies such as the one performed by Jagatic et al. [11], in which a social network was used for extracting information about social relationships, showed that more than

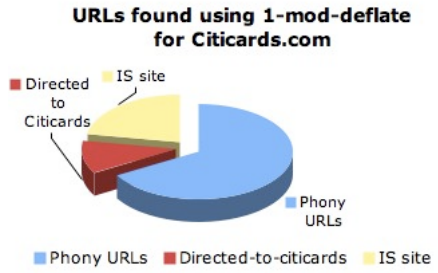


Fig. 3. Phony, IS and re-directed URLs found for Citicards.com.

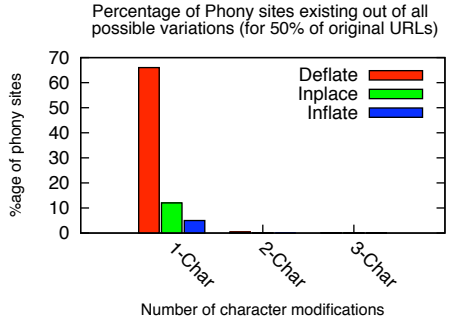


Fig. 4. Percentage of phony sites existing (out of all possible permutations) for at least 50% of original URLs from corpus.

80% of recipients followed a URL pointer that they believed a friend sent them, and over 70% of the recipients continued to enter credentials at the corresponding site. This is an indication of the gullible nature of most Internet users. Other studies regarding phishing, such as the one by Mailfrontier [9] and [6] provide credence to the fact that malicious impersonation in the Internet is a real threat. An important piece of work by Jakobsson et al. [5], [8] describes in detail how to set up a phishing experiment in order to measure how users might respond to such an unsafe environment. Articles and reports quoting various statistics lie testament to the problem we attempt to address [2]–[4], [13]. Unfortunately, all this body of work does not provide a comprehensive analysis of the parasitic URL naming problem. In fact this problem is so severe that heavyweights like The Coca-Cola Company, McDonalds Corporation, Pepsico, Inc., The Washington Post Company and others have all been forced to enter into litigation with entities which registered URLs closely resembling their official

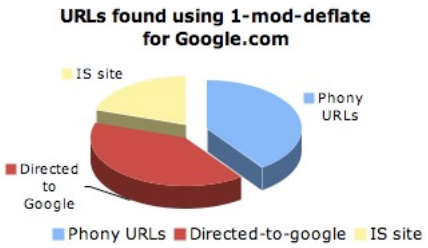
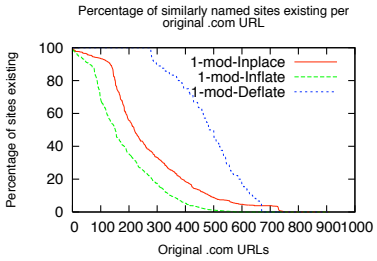
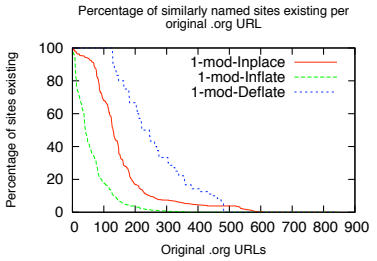


Fig. 5. Phony, IS and re-directed URLs found for Google.com.



(a)



(b)

Fig. 6. Percentage of sites existing with URLs similar to original corpus URLs and various extensions. (a) Displays percentage of sites which exist and have URLs similar to corpus URLs. Each phony URL has a .com extension. (b) Represents the same for .org

URLs [16]. In order to provide genuine websites with more clout when attempting to counter URL squatters, [15], the US passed the 'Anticybersquatting Consumer Protection Act of 1999'. Instead of all these legal mechanisms, this problem still exists. Another interesting piece of research is found at [33]. The authors develop an anti-phishing plug-in for web browsers called iTrust. Our work is different from this approach as iTrust links on to search results from Google to compute a threat score and warn the user. A study of typo-squatters has been conducted in [34]. Our study is more extensive than this effort, we probe approximately 3 million URLs. Further, the usage of third party redirections and the presence of cookies in a site does not always guarantee that a site is phony. Consider the site www.Bankofamerica.com is incorrectly flagged as unsafe (red) by the tool, strider, developed by the authors because of redirections to 3 domains. Moreover, we analyze the effect of the extension of a site (.com/.org etc) on its chances of being poached. The subsequent section discusses how prevalent is parasitic URL naming in the Internet.

III. How prevalent is Parasitic URL Naming

In this section, we quantify the extent of parasitic URL naming. We begin by explaining our measurement methodology.

A. Experimental Setup

1) **Building the Corpus:** We collected approximately 900 URLs among the most popular websites [17]–[22]. These original URLs were manually categorized in: brokerage firms (37 URLs), credit card firms (23), eCommerce sites (40), eMail providers (13), travel services (40), software vendors (32) and banking institutions which range from the US (253) to Canada (21), to Europe (220) and Asia (45) etc. The

TABLE I

Population Statistics for Different Extensions

Scheme with newly attached extension	Average	Variance	StDev.
1-mod-inplace (.com)	29	2045	1201
1-mod-inflate (.com)	19	1268	893
1-mod-deflate (.com)	52	4657	1869
1-mod-inplace (.org)	17.5	1190	880
1-mod-inflate (.org)	6.7	352	306
1-mod-deflate (.org)	30.5	2412	1481
1-mod-inplace (.biz)	11	669	550
1-mod-inflate (.biz)	2.7	119	112
1-mod-deflate (.biz)	14.5	966	755

average length of original URLs (without extension) was 8.9 characters, while the median was 8.

2) **Obtaining URL name variations:** These original URLs were modified by either inserting/deleting one or more characters at a time and then probing to ascertain if a URL with the modified name was already registered. For example, consider that we intend to find how any similarly named sites exist for Google. We substitute the first character with all possible alphabet letters. We repeat this with each character in the URL to obtain all misspellings of the original with one character modification. We term this method **1-mod-inplace**, since it changes only 1 character in the original URL, without changing the length of the URL. We use other schemes too where we remove one character from the URL i.e. **1-mod-deflate** or increase the length of the URL by one character, i.e. **1-mod-inflate**. We also experiment with 2 and 3 character modifications for inplace, inflate and deflate schemes.

B. Profiling URL Name-space

1) Analyzing the effect of URL modification: We analyze the extent of URL poaching and to this effect modify original URLs to search for phony sites in the Internet.

Observation 1: **Significant numbers of phony URLs exist in the Internet**. We observe from Fig. 1 that modifying each of the original URLs by changing one character inplace, by inflating the length of the URL by one character and by decreasing the length of the original URL by one character leads to the discovery of significant numbers of similarly named sites. Employing the 1-mod-inplace scheme, for nearly 30% of corpus URLs, we find that about 30-90% of all possible similarly named URLs exist in the Internet. Using the 1-mod-inflate scheme for about 25% of original URLs, we observe the existence of about 20-90% of all possible URL permutations. Similarly, using the 1-mod-deflate scheme, for 57% of corpus URLs, we observe the existence of about 35-90% of all possible URL permutations. These figures indicate the widespread existence of sites with URLs closely resembling original URLs. We also experiment with schemes exploring multi-character spelling changes, URL inflation and deflation. We find that for 2 or 3 character schemes the percentage of phony URLs existing per legitimate URL is below 0.5%.

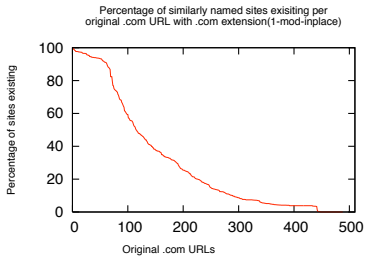


Fig. 7. Percentages of existing phony sites with URLs obtained from 1-mod-inplace scheme applied on original .com URLs.

Observation 2: **URL poachers prefer to register 1 character modifications of popular URLs**. Understandably, URL poachers do not expect Internet users to significantly mistype URLs and concentrate on registering domains with URLs which differ from legitimate ones by a small amount. From the obtained results it seems that *URL poachers expect most users to miss out on typing 1 character in URLs*. In Fig. 4, we see that for at least 50% of original URLs from our corpus, only 1 character modification schemes can uncover significant numbers of phony sites, hence in our work we concentrate primarily on 1 character modification schemes. Consider a popular site Google.com, depicted in Fig. 5, where we uncover all three types of sites (1) phony, (2) Similar sounding URL but pointing to main site (legit) e.g. gogle.com and (3) IS sites (goole.com). Another example presented in Fig. 3 depicts the case for Citicards.com. We observe that 66.6% of all possible misspelled URLs are phony. This is a significant number. Due to space constraints we defer from presenting graphs representing 2 and 3 character modification cases.

Observation 3: **Short original URLs suffer more from URL poaching**. As depicted in Fig. 2, for original sites which have URL length less than 10 characters, more than 10% of all possible phony URLs are registered in the Internet. This is an indication of URL poachers targeting sites with shorter names. This is somewhat expected as popular sites often have short names.

These observations clearly point to the strategies employed by URL poachers. In the next section we describe how these findings could be used to protect web-surfers against such threats.

2) URL extension analysis: Here we analyze how the extension of an original URL (.com etc) influences its chances of being poached. We generate phony URLs from original URLs using single and multi character inplace, inflate and deflate schemes. Subsequently, we attach an extension (.com etc) to these URL permutations to check whether the presence of a particular extension has an effect on the existence of fake sites in the Internet. We experiment with .com, .gov, .org, .net, .biz, .edu, and .mil. Some results for single character modifications are displayed in Fig. 6, which depicts percentage of all possible phony sites existing per original URL, when each phony URL has a .com or .org extension. Again, we reiterate that miniscule numbers of phony sites were found for multi-character schemes and due to space constraints we don't present results for them. We present Table I which describes the statistical characteristics of the data displayed in Fig. 6. Each scheme for the respective URL extensions, is listed in the first column while the next three describe the statistical features for the percentage of existing phony URLs in the Internet. We can clearly observe that for .com, .org and .biz extensions, the difference in the averages among the inplace and deflate schemes is higher than the inflate scheme. Further, the standard deviation is lesser for inflate schemes. This suggests that **numbers of phony URLs, obtained through the inflate schemes, discovered per original URL are generally much less than those discovered by inplace and deflate schemes**. This shows again that URL poachers expect users to either omit or misspell a character while typing a URL.

Observation 4: **Sites with .com extension have higher chances of being poached**. We observe that the population statistics for numbers of existing phony URLs with .com extensions are different from the other cases and thereby present a detailed analysis of the .com scenario. We present Fig. 7, which depicts the percentage of existing phony URLs with reference to all possible permutations obtained by inplace modifications of original URLs which ended with .com. We find that for 23% of all original .com URLs, about 50-90% of all possible phony sites exist. *This clearly indicates that a URL with .com extension has a high chance of large numbers of phony sites poaching it*. Further, we analyze how sites with .com extensions are poached across different domains, namely .org, .gov, .biz, .net, .edu and .mil. We present Fig. 8 (a)-(c) which displays how .com sites are poached across the complete range of URL extensions. Fig. 8 (d)-(f) displays the case for corpus URLs without a .com extension. We observe that:

1) Original URLs with a .com extension are impersonated primarily in .biz, .net and .org domains.
2) Original URLs without a .com extension are impersonated primarily in .com, .net and .org domains.

IV. Discussion and Future Work

We have observed interesting characteristics of the parasitic URL naming problem. We envision the use of these features as the building blocks for an online warning system which would be able to determine if a site is phony or legitimate and provide an appropriate warning message. We are aware of services which provide malware related warnings, such as McAfee Site Advisor [31]. These services however are targeted more towards filtering sites which have been known to spread malicious code by exploiting vulnerabilities in browsers and operating systems. It would be interesting to observe the efficacy of the features described in the previous section on a set of sites and compare them with existing commercial services such as McAfee Site Advisor.

We are in the process of extending this work. Specifically, (a) developing a more complete profile of phony sites and (b) designing an automated tool to detect whether a website is phony.

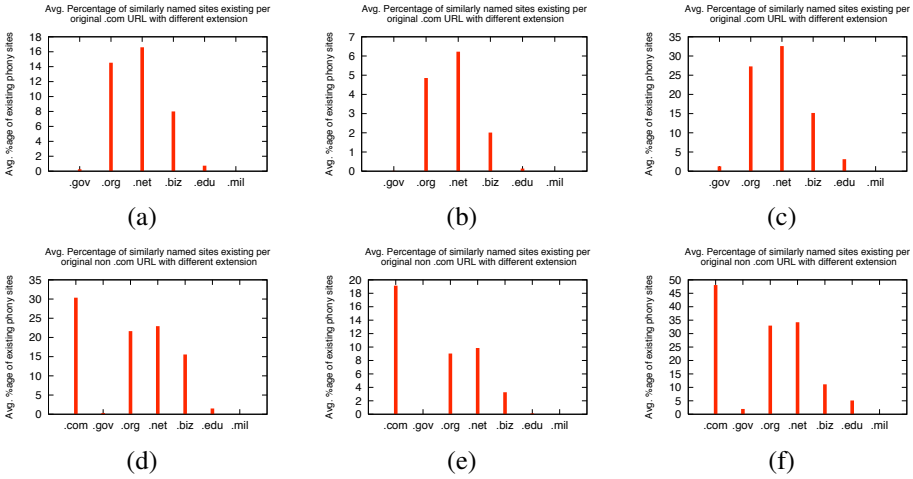


Fig. 8. Relationship between original extension of a URL and chance of being poached by phony sites with different URL extensions. (a)-(c) represents percentage of phony sites (with various extensions) which have URLs derived from original .com sites using 1-mod-inplace, 1-mod-inflate and 1-mod-deflate. (d)-(f) represent percentage of phony sites found which have URLs derived from various schemes applied on originally non-.com sites.

V. Conclusion

Our research has focused on quantifying the parasitic URL naming phenomenon. We conduct an extensive measurement based analysis probing more than 3 million sites obtained by modifying URLs from a corpus of 900 popular sites. We uncover that most phony URLs differ from legitimate ones by just 1 character, in length or in spelling. We find that URL poaching is a widespread problem, 99% of all misspelt sites discovered by us were found to be phony. We find that for nearly 57% of all original URLs in our corpus, more than 35% of all possible URL variations (for each original URL) exist in the Internet. This proves the massive scale of this problem. Interestingly, short URLs, less than 8 characters in length are more susceptible to poaching and sites with .com extension have a significant chance of being poached. We expect these results to be helpful in developing a phony website warning system to protect web surfers.

References

[1] http://www.antiphishing.org
[2] http://www.crime-research.org
[3] http://www.csmonitor.com/2005/0505/p13s01-stin.html
[4] http://www.cs.cmu.edu/ help/security/hoaxes_scams.html
[5] M. Jakobsson and J. Ratkiewicz, "Designing Ethical Phishing Experiments: A study of (ROT13) rOnl auction query features.", WWW 2006.
[6] Rachna Dhamija and J. Doug Tygar. The battle against phishing: Dynamic security skins. In Proc. ACM Symposium on Usable Security and Privacy (SOUPS 2005), pages 7788, 2005.
[7] Gartner Inc. Gartner study finds significant increase in e-mail phishing attacks. http://www.gartner.com (April 2004).
[8] M. Jakobsson and S. Myers. Phishing and Counter-Measures. John Wiley and Sons Inc, 2006.
[9] Mailfrontier phishing IQ test. http://survey.mailfrontier.com/
[10] Garfinkel, S., and Miller, R. Johnny 2: A user test of key continuity management with S/MIME and Outlook Express. Symposium on Usable Privacy and Security.
[11] T. Jagatic, N. Johnson, M. J., and Menczer, F. Social phishing. 2006.
[12] http://www.ngssoftware.com/papers/ThePharmingGuide.pdf
[13] http://www.drive-bypharming.com/
[14] http://www.caida.org/ nevil/Bojan_Zdrnja_CompSci780_Project.pdf
[15] http://www.uspto.gov/web/offices/dcom/olia/tmcybpiracy/repcongress.pdf
[16] http://www.nysd.uscourts.gov/courtweb/pdf/D08MNXC/02-08168.PDF
[17] http://www.pharming.org
[18] http://www.alexa.com
[19] http://www.forbes.com
[20] http://www.netvalley.com
[21] http://www.wired.com
[22] http://www.consumersearch.com
[23] www.cs.auckland.ac.nz/ trebor/papers/CHEN02.pdf
[24] T. Honda, M. Yamamoto and A. Ohuchi, Automatic Classification of Websites based on Keyword Extraction of Nouns, Information and Communication Technologies in Tourism 2006, Springer Vienna, '07.
[25] S. Roy, S. Joshi and R. Krishnapuram, Automatic categorization of web sites based on source types, Procs. of the fifteenth ACM conference on Hypertext and hypermedia, pages 38–39, 2004.
[26] G. Kening, Y. Leiming, Z. Bin, C. Qiaozi and M. Anxiang, Automatic Classification of Web Information Based on Site Structure, Procs. of Intl. Conf. on Cyberworlds, 2005.
[27] A. Banerjee, A. Mitra and M. Faloutsos, Dude where's my Peer, Procs. of Globecom, ISET, 2006.
[28] J. A. Kunze, Towards Electronic Persistence Using ARK Identifiers, ARK motivation and overview. July 2003.
[29] http://www.caida.org
[30] http://api.hostip.info
[31] http://www.siteadvisor.com/sites/sedoparking.com
[32] http://www.google.com/press/zeitgeist.html
[33] T. Ronda, S. Saroiu, and A. Wolman, iTrustPage: Pretty Good Phishing Protection, Tech. Report CSRG-545, Dept. of Comp. Sc., Univ. of Toronto, Dec. 2006.
[34] Yi-Min Wang, Doug Beck, Jeffrey Wang, Chad Verbowski, and Brad Daniels, Strider Typo-Patrol: Discovery and Analysis of Systematic Typo-Squatting, in Proc. Usenix SRUTI Workshop, July 2006.