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: (a) quantify the extent of this parasitic URL naming, (b) develop a profile of phony web-sites, and (c) develop, ADS, an automated approach to detect phony sites. 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. The high level conclusions are good and bad at the same time: (a) parasitic URL naming is a wide-spread phenomenon, and (b) phony sites have several distinctive characteristics compared to the original sites. Using our empirically-derived profiles, we show that ADS can distinguish phone websites with 92% accuracy. We argue that ADS could be ultimately incorporated in a web-browser to warn users of potentially phony sites.**

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.

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, if at all. Several questions would be interesting to answer.

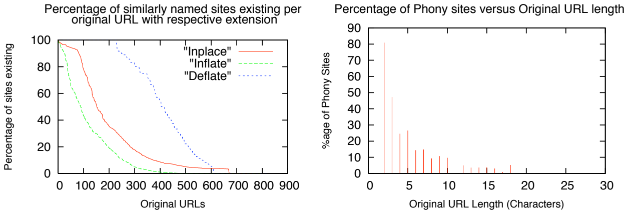
1) How prevalent is parasitic URL naming?
2) Are phony sites different than "legitimate" sites?
3) Can a web-browser detect phony sites automatically?

In this paper, we conduct a massive study whose goal is to: (a) quantify the extent of parasitic URL naming, (b) profile phony web-sites, and (c) develop an automated approach to detect phony sites, which could be incorporated in a web-browser. 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. We use a keyword based method to distinguish between phony and IS sites.

To the best of our knowledge, this work is the first to characterize traits of parasitic URL naming. Our main results can be summarized in the following points.

a. 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 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



(a) (b)

Fig. 1. (a) 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. (b) Percentage of phony sites existing versus length of original URL.

from parasitic URL naming.

b. Profiling phony web-sites. We develop profiles of phony websites regarding various aspects, such as content, network behavior, and network location.

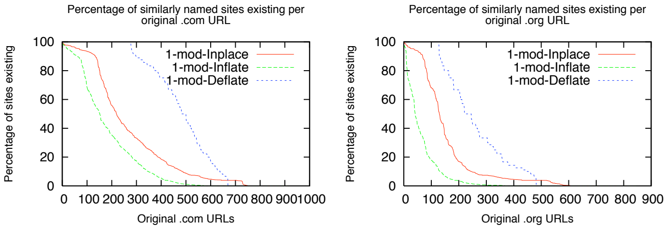
- **The size of the page of the original sites is larger than that of phony sites:** We find that for original sites, the average page size is 20 KB, while for phony sites, the average size is 9 KB.
- **Reaching a phony website involves a larger number of HTTP-redirections:** We find that the number of HTTP-redirections for connections which end up at phony sites are 6-9 times more than those that arrive at original websites.

c. Developing an automated detection approach. We develop a tool to detect phony sites based on the profiles that we have obtained. Namely, we transform the empirical observations into a set of criteria for detecting phoniness. We call our tool Automated Detection System or ADS for short and it could have an advisory role as a plug-in in a web-browser or even a mail application. We validate ADS in a controlled experiment of 100 sites, and we manually validate the legitimacy of each site. ADS identifies phony websites with 92% accuracy and 93% recall.

The next section presents related literature, while section III deals with quantifying the extent of URL poaching. Section IV presents a network-based analysis of parasitic URL naming, followed by section V which discusses the automated detection of phony sites.

II. Related Work

Efforts attempting to characterize the parasitic URL naming problem have been limited to identifying some ”typo squatters” [14] 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 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 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. Our work is different from the previous approaches since we conduct a detailed analysis of parasitic URL naming: which sites are effected, where are the fake sites hosted and what can be done to combat it. The subsequent section discusses how prevalent is parasitic URL naming in the Internet.



(a) (g)

Fig. 3. 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

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 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

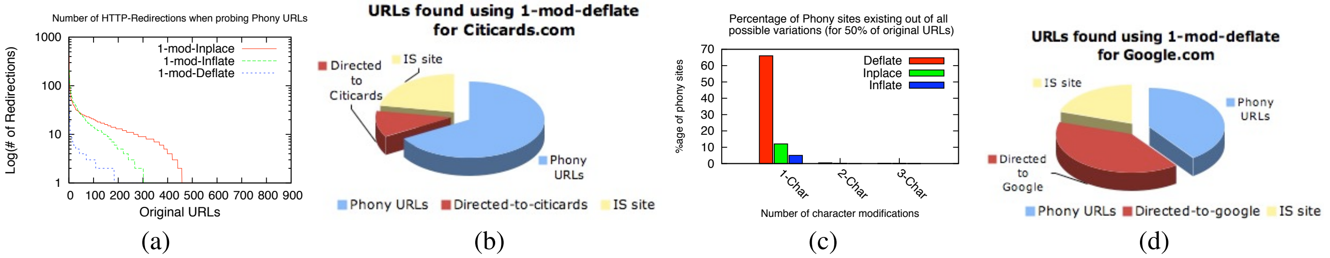


Fig. 2. (a) Number of HTTP redirections observed when probing for phony URLs obtained by applying various schemes to each original URL. (b) Phony, IS and re-directed URLs found for Citicards.com. (c) Percentage of phony sites existing (out of all possible permutations) for at least 50% of original URLs from corpus (d) Phony, IS and re-directed URLs found for Google.com.

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. Further, we modify the extension of the original URLs to find which particular extensions (say .com) are targeted aggressively by URL poachers.

3) **Analyzing HTML Content:** If a site (with a modified URL) is located, the content is downloaded for automated comparison with the original site to find the degree of similarity. This is done by using using keyword matching [23], [24]. We build a keyword set by stripping off words from HTML-headers manually from a subset of the corpus. At least 20% of keywords used are present in each original site. We are aware of other mechanisms to classify sites [25], [26], which use site structure and image content to calculate site similarity.

Since most of the sites probed during our experiments did not have significant image content, we chose to use this simpler keyword based matching mechanism. We define two broad approaches which we employ:

Method 1(M1): We compare the frequency keywords found in a legitimate page with the frequency in phony pages. For this we use a root mean square metric, which calculates the deviation in the number of occurrences of keywords by computing the Root Mean Square Error (RMSE) of the number of times a keyword appears in a phony site with respect to the number of times it appeared in the original content. This is a type of L2 metric represented by $\Sigma(x_i - o_i)^2$, where x_i represents frequency of occurrence of keyword i in a phony site and o_i represents frequency of occurrence in the original site. We are aware of the possibility that our results may depend on our choice of keywords. However, other methods such as website structure analysis and image comparison have limitations too. We have attempted to use words which can be easily related to particular categories, allowing us to observe semantic differences in the content. We built a second corpus of sites spanning different categories from which we picked out these keywords.

We find via manual inspection that for IS sites, RMSE is greater than 190. We use this method to distinguish incidentally similar (IS) from phony sites. The high value for RMSE is due to the fact that IS sites don't try to imitate original sites (content-wise) and so have high RMSE values. Additionally, we also analyze network specific features, such as HTTP redirections, IP-geolocation and DNS resolution entities.

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

Method 2(M2): We compare how *closely* phony sites imitate original websites. We use a bit-vector-based hamming distance metric to ascertain this. This is a type of L1 metric defined by $\Sigma|x_i - o_i|$, where x_i and o_i represents the ith bits in the two vectors. Via this method, we develop heuristics based on a bipartite graph mapping. Here we represent every keyword, K_n as a node in the bipartite graph and every site (phony or legitimate) as nodes S_n. We now observe which keywords appear on which sites. This simple formulation allows us to calculate the in-degree and the *weighted in-degree* of each phony site. The weighted in-degree, is simply the sum of the number of occurrences of the keywords which appear on a site. Further, we also analyze the similarity of phony sites with the original by using this graph representation to generate a bit vector for each site and compare the hamming distance to the bit vector of the original site. A 1 in the bit vector represents the presence of the specific keyword in the site.

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.(a) 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

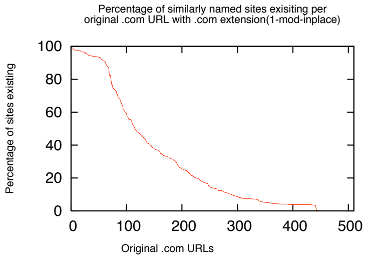


Fig. 4. Percentages of existing phony sites with URLs obtained from 1-mod-inplace scheme applied on original .com URLs.

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%.

Observation 2: **URL poachers prefer to register 1 character modifications of popular URLs**. Understandably, URL poachers do not expect Internet users to significantly mis-type 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. 2.(c), 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. 2.(d), 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. 2.(b) 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. 1.(b), 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 features are incorporated into ADS.

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. 3, 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. 3. 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.

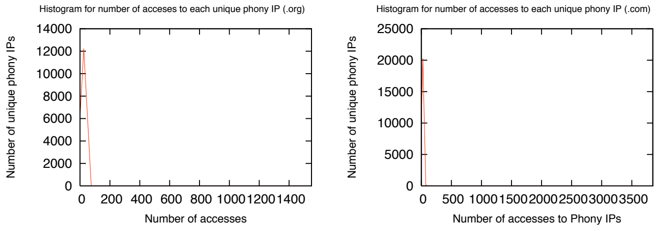
Observation 4: **Sites with .com extension have higher chances of being poached**. We observe that the population statistics for numbers of existing phony URLs with .com extensions are different from the other cases and thereby present a detailed analysis of the .com scenario. We present Fig. 4, which 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. 5 (a)-(c) which displays how .com sites are poached across the complete range of URL extensions. Fig. 5 (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.

3) Effect of URL category: We now study the effect of the original URL category on the percentage of phony sites discovered per original URL. We present Fig. 6 which depicts the percentage of phony sites per original URL in some categories. Due to space constraints we list results for a subset of the categories. We find that for all cases the 1-mod-deflate scheme contributes most to the number of phony URLs in comparison to other schemes. Further, *.com, .org and .net are the most aggressively poached domains by URL poachers*. Domains such as .mil, .edu, .gov and .biz see significantly lower levels of parasitic URL naming.

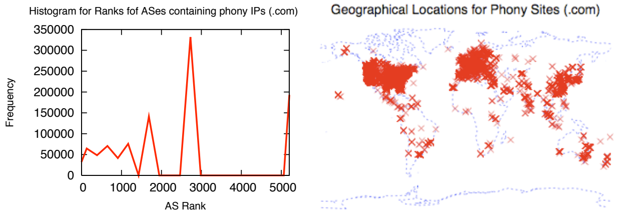
Observation 5: **URLs which belong to German banks suffer most from parasitic URL naming, followed by URLs which belong to US banks**. Other significantly poached categories are UK banks, software and technology companies and travel-related sites.

4) Defense employed by original sites: Here we attempt to understand the effect of strategies employed by original sites to fight URL poachers. The technique for making a web



(a) (b)

Fig. 7. Histogram for Frequency of observing unique phony IPs. (a) .org domain (b) .com domain



(a) (b)

Fig. 8. (a) Histogram for AS ranks of ASes which host URL poachers for the .com domain. (b) Geographical mapping for IPs discovered while spidering .com domain.

page available under many URLs is commonly termed as URL redirection and uses the "redirect" feature of the HTTP protocol. When a browser requests a URL from a web server, the server can return an HTTP redirection status code in the range 300-399, which indicates to the browser that it should look for the page at another URL. This other URL is specified in the "Location" HTTP header. Upon receiving a response with this status code, the browser finds the "Location" header and immediately issues a request for the URL specified in that field.

Popular sites often buy out domain names which are similar to their original sites. Accessing these misspelt sites leads to HTTP connections being directed to the home site.We observe that **less than 5% of original sites own misspelt URLs**. When probing for the existence of phony sites using misspelt URLs obtained from the 1-mod-deflate scheme, we see that HTTP connections terminate at about 250 unique sites, this is depicted in Fig. 2.(b). We find that for the same scheme, accessing misspelt URLs for 1800flowers.com leads to their home site. In fact, 22.7% of all possible misspelt URLs (obtained by 1-mod-deflate) are controlled by 1800flowers.com. For inflate and inplace schemes we observe that the the number of URLs controlled by original sites is lower. 1800flowers.com, still the most aggressive site to buy out misspelt URLs controls about 8.6% and 12.3% of misspelt URLs obtained via the inflate and inplace schemes respectively.

C. Page Size analysis

We begin by comparing the sizes of the HTML pages for phony sites versus legitimate sites. In Fig. 9. (c) and (d), we present the CDF plots for the sizes of HTML pages for legitimate and phony sites. These graphs clearly show the comparatively even spread of web page sizes for legitimate sites compared to the highly clustered sizes for phony sites. We find that for 90% of legitimate sites, **HTML page sizes are less than 56 KB, while for phony sites the 90% mark lies at 31 KB**. The most popular file sizes for phony sites are 305 bytes, 2.8 KB, 13 KB, and 74 KB. In fact all the 5032 sites with file size 305 bytes follow exactly the same HTML-page structure, employing one frame within the html body, with exactly the same tag indentations. This similarity hints towards the fact that these were mass-produced using HTML source generators. All these pages have exactly one hyperlink embedded inside them which points to the source from where the page dynamically loads the contents of the HTML frame. This frame contains the actual hyperlinks which advertise everything from mortgage services to bathtubs. Embedding hyperlinks which load the content of the HTML page dynamically allow the controller of all these pages to update HTML content effortlessly. Based on this information, one of the heuristics used by ADS to detect phony pages is page_size $\leq$ 31KB.

Summary: We now present a concise digest of our findings in Table II. Apart from the mentioned categories we find that Chinese banking institutions are lightly poached with the maximum URL poaching percentage close to 50% and for email providers, Gmail, Yahoo and Hotmail at least 60% of all possible phony URLs exist. Further, they are poached primarily in .com domain. German banks are most poached in the .net domain. Further, Japanese banks suffer from very low levels of parasitic naming. The average percentage of phony sites for highly poached categories such as German and US banks and travel related sites are displayed in Fig. 6.(a)-(c). In this section we have provided detailed information about URL characteristics of phony sites with respect to legitimate portals. ADS uses these features to segregate phony sites from original sites. In the subsequent sections we will focus on a network based analysis to understand where these URL poachers are located in the Internet.

TABLE II
HIGHLY POACHED WEBSITES IN VARIOUS CATEGORIES

Category	Highly Poached Entities
Brokerage Institutions	Merrill Lynch, Credit Suisse and Lehman Brothers
Canadian Banks	Banque Nationale Du Canada
Credit Card Companies	Chase, and jcbusa
eCommerce Retailers	Adidas, Gamestop, Walmart and Samsung
eMail Providers	Gmail, Yahoo and Hotmail
Indian Banks	UTI Bank
Social Networking Sites	Orkut and batchmates
Travel Sites	US air and Southwest
Software and Tech.	Adobe, Sun, and Lexmark

IV. Network centric analysis

We now present a network-based profile for these phony sites. We analyze the HTTP-redirections, IPs which host these phony sites, their AS-ranks by using an IP to AS lookup [27], and DNS-resolution IPs for each of them.

A. IP Geo location

Observation 6: **The majority of phony sites are consistently located within the mainland US and also in Europe**. We attempt to understand geographical characteristics for IPs

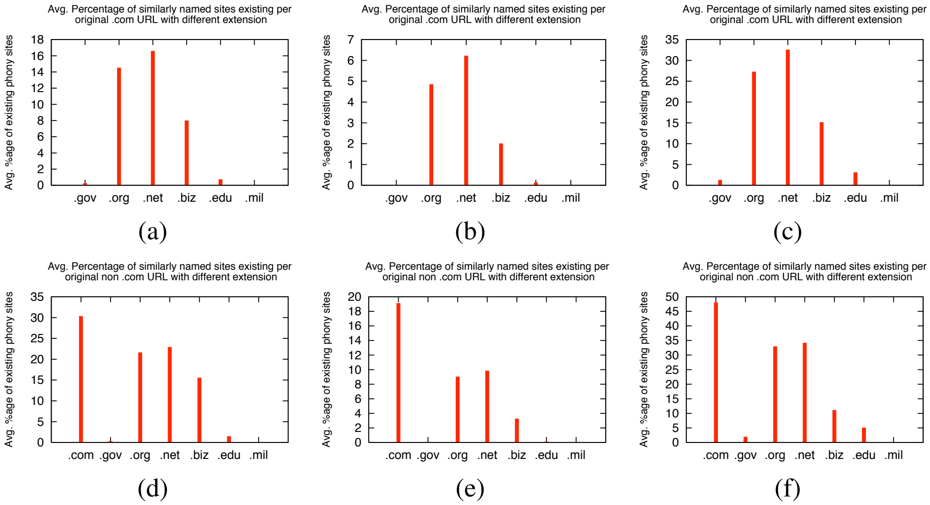


Fig. 5. Relationship between original extension of a URL and chance of being poached by phony sites with different URL extensions. (a)-(c) represents percentage of phony sites (with various extensions) which have URLs derived from original .com sites using 1-mod-inplace, 1-mod-inflate and 1-mod-deflate. (d)-(f) represent percentage of phony sites found which have URLs derived from various schemes applied on originally non-.com sites.

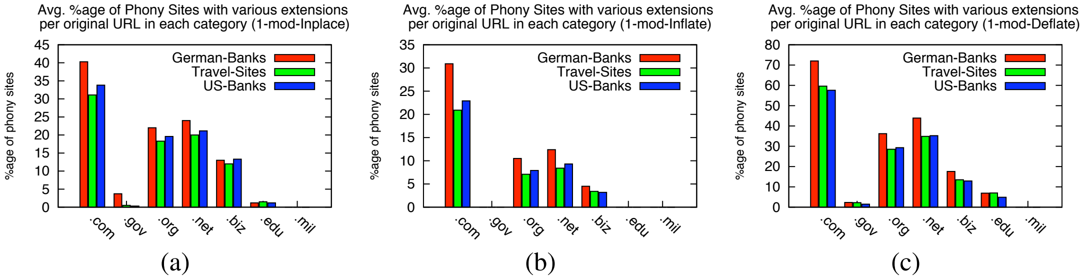
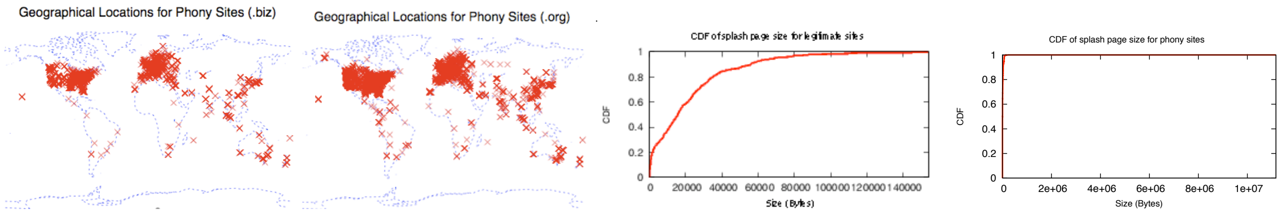


Fig. 6. Average percentage of phony URLs existing per legitimate URL, in each category. (a)-(c) represent 1-mod-inplace, 1-mod-inflate and 1-mod-deflate schemes for German banks, travel related sites and US banks.

of phony sites. We utilize APIs for Geo-mapping services provided by [30]. We observe in Fig. 8.(b), that IPs relating to phony sites in the .com domain are extensively spread throughout the mainland US as well as in Europe and Japan. For the .biz domain, the trend is similar even though the percentage of locations mapped in Asia are lesser. The case for IPs in the .org domain is similar according to the density of locations plotted in the mainland US. For IPs in the .edu domain, we observe sparse location mapping in Asia, with the hotspots primarily restricted to the US and Europe. For IPs in both .gov and .mil domains we observe a complete lack of locations outside the mainland US. While for IPs in .net domain we observe a much more dense concentration of locations inside the US, Europe, Japan, Australia and South America. Clearly, we can observe that the majority of phony sites are consistently located within the mainland US and also in Europe. For .org, .com, .biz and .net domains Japan seems to be a hotspot from where phony sites are hosted. This is displayed in Fig. 9.

B. AS centric analysis

We find that the number of distinct AS numbers (ASNs) for ASes which host phony sites are again maximum for the .com domain. For .com we observe 3324 unique ASNs from which phony sites are hosted. The respective numbers for .org, .net, .gov, .biz, .edu, and .mil are 2465, 2566, 111, 1264, 414 and 13. These numbers are further indication of the fact that *URL poachers are more widespread within the .com, .org and .net domains*. We now present information about the degree based ranks of ASes. Using the latest CAIDA [29] AS-Rank dataset and mechanisms employed in [27] , we are able to observe the relative ranks of ASes by mapping ASNs to ranks. We present Fig. 8.(a) which depicts the AS ranks of the various unique IPs recorded for the .com domain. Clearly, we observe four clusters: **ASes which have ranks in the range (a) 1-900, (b) 1600-1800, (c) 2600-2800 and (d) 5100 and beyond host most of the phony sites**. This is true for all other domains too. We present Table III, which lists the top 5 ASes for each domain. Each entry in the table lists a tuple separated by a colon. The first part represents the ASN while the next represents the percentage of total phony sites found in that AS. Interestingly, *the top 5 ASNs active in .gov, .edu and .mil domains are significantly different from those in .com, .org, .net and .biz domains*. ASNs 19318 (AS degree=6), 26496 (AS degree=4, Go Daddy Inc.), 33626 and 8560 (AS degree=11, Schlund & Partners AG) are some of the top common ASes which host these phony sites.



(a) (b) (c) (d)

Fig. 9. (a)-(b) Geographical mapping for IPs discovered while spidering different domains. (c)-(d) represent CDF of HTML page sizes for legitimate and phony sites.

TABLE III
LIST OF TOP 5 ASES HOSTING PHONY SITES PER DOMAIN

Domain	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
.com	19318:8.1	26496:5.3	33626:4.4	13768:4.4	8560:3.3
.org	19318:6.1	26496:4.8	8560:4.3	33626:4.2	6245:2.5
.net	11486:9.9	19318:4.6	8560:3.9	26496:3.8	3561:2.4
.biz	8560:13	26496:8.7	33070:8.3	33626:5.0	6245:2.8
.gov	2152:5.3	7018:4.4	2714:3.9	3136:3.3	15130:3.2
.edu	16966:2.9	8560:2.3	6325:2.1	4323:2.0	31822:1.8
.mil	721:18.5	2152:14.8	27046:11.1	724:7.4	701:7.3

C. DNS analysis

We analyze who handles the Authoritative-DNS-resolution for phony sites. This information is useful to understand whether IPs of phony sites are recorded widely among popular DNS servers or not. We find 9627 unique IPs which resolve DNS requests for phony sites in the .com domain. The respective numbers for .org, .net, .biz, .edu, .gov and .mil are 9026, 8593, 3826, 793, 208 and 19. This again is a clear indication of the fact that parasitic URL naming is much more active in .com, .org and .net domains compared to the other domains. We find that for the .org domain, out of the top 5 DNS IPs, two belong to Level3 communications (L3c) while the rest belong to InterNap services (INS). For the .net domain we identify NTT America Data Centers, L3c and INS host the top 5 DNS resolution servers. Similarly, for the .mil domain, California State University Network (CSUN), US Army TAC and L3c own DNS servers. For the .gov domain, CSUN, L3c and Sprint perform DNS resolutions. For the .edu domain Sprint, Qwest and AT&T own the most referenced DNS servers. For the .com and .biz domains we identify NTT America Mid Atlantic, INS and Schlund & Partners AG, INS respectively. This information clearly points to the fact that most referenced locations for DNS resolutions regarding phony sites are very restricted. In fact the top 5 DNS servers for original IPs were found to belong to Internet Media Network, Verisign, UltraDNS, Salomon Inc. and Deutsche Telekom AG. None of these figure in the top 5 DNS server IPs for phony sites. This information could potentially be a simple first check to warn the user that they might be visiting a phony site.

D. HTTP redirection analysis

Here, we study how HTTP connections to misspelt sites finally end up at phony sites. HTTP redirections are often used to host dynamically changing content on different webpages or to make available a webpage under different names. *Observation* 7: **the number of URL redirections for phony URLs obtained from inflation and inplace modification schemes is much higher than ones obtained by deletion schemes**. We present Fig. 2.(a), which shows that the number of URL redirections for phony URLs. Phony URLs obtained by using the 1 character deletion scheme display significantly lesser redirections when compared to inflation and inplace modification schemes. Since URL redirection is often employed to direct users to updated site content [28], this metric indicates the dynamic nature of these phony URLs and the associated fake-content. We observe that the average number of URL redirections for 1-mod-inplace and 1-mod-inflate schemes were about 9 and 6 redirections. Surprisingly, average number of redirections for 1-mod-deflate was only 1. This could imply that phony URLs obtained by inplace and inflate schemes could point to sites which have phony URLs obtained by deletion schemes. In fact, this provides credence to our previous deduction that URL poachers expect users to miss out on typing one character from prominent URLs and hence seem to register domains with such misspellings more aggressively. For 2 and 3 character schemes, we discover very few phony URLs and the number of redirections for deletion schemes is less than for inplace and inflate methods, the difference is smaller though. We use these heuristics within the ADS framework.

E. Phony sites: Behavior analysis

We find that all the sites with a 305 byte HTML page, attempted to open pop ups linking to http://b.casalemedia.com. Further, most of them obtained the contents for the main page frame from http://www.searchnut.com, while the domain that is advertised on the main HTML page is usually http://mortgages-rates.com. This single common entity is the most aggressive URL poacher we could identify. We discover that this entity owns 1380 (via inplace),1549 (via deflate) and 295 (via inflate) sites within the .com domain. Similarly it is responsible for 479 (inplace), 401 (deflate) and 82 (inflate) sites within .org and 414 (inplace), 350(deflate) and 82 (inflate) sites in .net domain. Clearly, **the most aggressive URL poacher concentrates on sites with .com, .org and .net extensions**. Most URLs registered by this entity are discovered by applying inplace and deflate schemes. This is in keeping with our previous finding that URL poachers expect users to miss out, or misspell characters while entering URL addresses. We also categorize these results based on the kind of URLs that they attempt to poach. We present Fig. 10.(a) -(c), which depicts the numbers of phony URLs controlled by this single entity across the three most significant domains. We can clearly observe the common trend in the three graphs, wherein all three peak in the same categories. We find that the numbers for phony URLs peak for the following categories: credit-card companies and eCommerce retailers (categories 4 and 5), German banks (category 8), social networking sites and software companies (categories 14 and 15) and US banks (category 20). The second most prolific URL poacher is sedoparking.com/sedo.com which declares clearly on its pages that its business model is based on registering sites with phony URLs in order to display ads to unsuspecting users. All this information can be used to develop simple filters integrated with browsers to warn users that they might be visiting phony sites.

F. Profiling IPs used by URL-poachers

We begin by profiling the IPs of various phony sites. We term such IPs as phony IPs.

Observation 8: **Most phony IPs host less than 100 phony sites from itself**. We present Fig. 7.(a)-(b), depicting the histogram for the number of times each unique phony IP is observed for .org and .com domains, while probing for phony URLs. We can clearly see that only miniscule numbers of phony IPs host large numbers of phony sites. In all cases we observe that most phony IPs host less than 100 phony sites. This is a possible indication of the fact that parasitic URL naming is a widespread problem, not controlled by a select group of entities. In fact most phony IPs, irrespective of the domain attract less than 100 hits to themselves. We conduct a more granular ananlysis. We count the number of unique phony IPs which host phony sites, which have a .com extension and son on. For .com, we find 20448 unique phony IPs. The respective numbers for .org, .net, .gov, .biz, .edu, and .mil are 12283, 13154, 173, 4383, 703 and 20. In the subsequent section we discuss how we develop ADS.

V. ADS: Detecting URL poachers

In this section, we describe ADS which automatically detects phony sites based on a set of empirically derived criteria.

TABLE IV

KEYWORD BASED ANALYSIS OF LEGITIMATE VS PHONY SITE-CONTENT.

	M1		M2							
Site	RMSE avg	RMSE SD	Orig ID	PID mode	PID avg	PID SD	WPID avg	WPID SD	HD avg	HD SD
Blockbuster	13.28	12.11	10	17	15.44	9.67	57.33	31.37	17.22	7.95
DeutscheBank	16.69	14.92	25	25	18.39	7.33	101.56	57.6	14.39	11.84
Batchmates	26.27	34.73	19	17	15.64	6.59	60.5	25.66	17.21	9.54
Abercrombie	44.24	72.81	13	17	19.45	13.76	124.2	150.44	25.55	10.95
Adidas	60.26	80.87	11	17	14.1	8.1	62.02	55.87	17.07	4.92
GoldmanSachs	76.07	82.06	16	17	14.89	8.39	75.53	68.4	18.49	7.33
AmericanExpress	77.82	62.09	17	17	14.71	8.19	66.01	55.65	17.43	3.99
Amazon	81.73	96.67	36	17	16.08	10.62	80.93	77.89	28.37	7.82
Delta	89.74	107.16	17	17	12.63	7.81	55.34	52.32	18.86	6.06
Apple	93.34	87.86	22	17	14.19	8.6	65.28	60.39	19.91	2.87
Costco	94.89	164.5	36	17	14.6	10.14	72.1	85.02	33.78	3.8
BestBuy	98.05	96.46	29	17	14.28	9.31	65.71	69.6	26.39	2.35
Adobe	101.73	110.02	17	17	14.1	8.43	64.11	57.99	19.34	4.84
Dell	101.78	106.41	17	17	13.89	8.41	62.52	60.73	19.04	6.41
BankOfAmerica	155.87	287.9	32	17	8.14	8.37	67.65	65.45	23.54	5.04
AMD	162.23	108.24	14	17	12.58	8.46	57.13	63.42	20.3	5.79

A. The ADS Tool

1) Detection criteria: The ADS tool uses a set of criteria to decide if a site is phony or not. All the criteria must be met to decide if a site is phony. Next, we describe each criterion in detail and present an overview in Table V.

A site is classified as phony only if it satisfies all criteria. We use the following heuristics:

(a) URL length and edit distance: As discussed in previous sections, URLs which are less than 10 characters long, and which differ from original URLs by one character are nearly guaranteed to be phony. As seen in observations 2 and 3.

(b) No. of HTTP-redirections: 6-9 HTTP redirections are a good indicator of whether a site is phony. Most original websites did not have any HTTP redirections. As described by observation 7.

(c) URL extension: If trying to open a site with .com extension leads to opening of a .biz/.net/.org site, the final site is possibly phony. As discussed in observation 4.

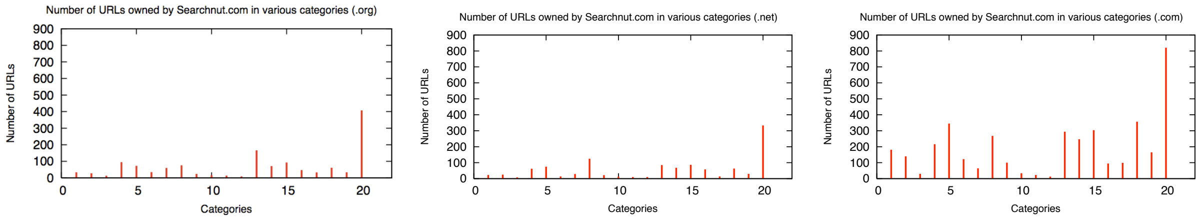
(d) Page size: 90% of phony sites have web page size less than 31 KB, while 90% of original sites have size less than 56 KB. Phony sites have HTML pages which are much smaller than original sites. We use this fact in ADS. A detailed description was presented in Sec. III.C.

(e) Pop-up: Phony sites often try to open pop-ups, most targeted towards well known URL poachers. This is an indication that a site is phony. This was discussed in Sec. IV.E.

(f) Keyword analysis: From initial experiments we find that websites hosted by URL poachers satisfy the following ranges: 12<PID<19, 60<WPID<100 and 0<HDavg<30. After analyzing a subset of sites, we found that those sites which displayed RMSE>190 were IS sites. We use this information and all sites which had RMSE >190 were marked as IS. As seen in Sec.III.A.

We envision ADS as a open-source plug-in to popular open-source browsers. The plug-in manages a list of popular legitimate URLs (indexed according to categories) and compares URLs being typed by the user in real-time with this list. This can be implemented by [32]. Then it analyzes HTTP connection characteristics followed by a comparison of the HTML content of the suspect page with the HTML content stored in local caches. This kind of web-object caching is already implemented by browsers. Next, we present several case studies from our corpus to highlight the appositeness of these methods.

2) Validation: We selected a set of 100 sites, which consisted of both original sites and sites with misspelt URLs. ADS was run on this set to identify phony sites. The results were validated by manual inspection. We refer to sites in the test set with misspelt URLs as *suspect* sites. A summary of our findings follows in Table IV where we observe that searchnut.com, represented by the entry 17 in PID mode column, is extremely aggressive at hosting phony sites. Here, RMSE represents Root Mean Square Error, avg: average, SD: Standard Deviation, ID: In-Degree, PID: Phony-site In-Degree, WPID: Weighted PID and HD: Hamming Distance. We find credence for our earlier deduction that German banks are heavily poached from the data displayed in the Deutsche Bank row in Table IV. In fact URL poachers also seem to try and make their phony sites resemble Deutsche Banks page as can be observed from the low RMSE values. A similar case can be observed in the Blockbuster and Batchmates rows, where low RMSE values appear. Surprisingly, we observe some cases where URL poachers do not attempt to deck up their sites with similar keywords which are present in the original site's content. This is seen for AMD, Bank of America, BestBuy, Costco and Adobe. We find that URL poachers attempt to insert a large variety of keywords advertising products and/or



(a) (b) (c)

Fig. 10. Number of URLs operated by Searchnut.com within .com domain, sub-classified according to URL categories. We observe clear peaks for specific categories and this trend is reflected across other domains too.

TABLE V
CRITERIA USED BY ADS TO DETECT PHONY SITES

Feature-set	Primary Criterion
URL name	suspect-site URL length $\leq$10 & suspect URL differs from original URL by 1 char.
HTTP-redirections	No. of HTTP-redirections for when accessing suspect sites is between 6-9 and suspect_URL_length$\geq$original_URL_length
URL extn.	Accessing a suspect site with .com extension finally opens a site with .biz/.net/.org extn.
AS-Rank	If HTML content of suspect site hosted by ASes with ranks, 2600-2800/5100
HTML Page Size	HTML Page size of suspect site is $\leq$ 31KB/305B
Behavior	Accessing suspect site leads to opening a popup linking to casalemedia/sedoparking.com
Keyword analysis	For suspect site HTML content 12<PID<19, 60<WPID<100 & 0<HDavg<30

services on these phony sites which leads to large PID values. Due to space constraints we refrain from displaying the complete list of sites and the associated M1 and M2 metrics. These metrics are useful in identifying phony sites. We find that using these and other heuristics described earlier, ADS is 92% accurate for identifying phony sites from the test set with 93% recall.

VI. 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. Interestingly, URLs which belong to US and German banks suffer most from parasitic URL naming followed by software and technology companies and travel-related sites. To combat this, we develop ADS which uses a plethora of meaningful features ranging from URL names and HTTP redirections to criterion obtained from keyword analysis of sites. We find that ADS can successfully identify phony sites with 92% accuracy and 93% recall rate.

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/survey/quiztest.html.
[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