A Trend Micro Research Paper

# Soundsquatting

## Uncovering the Use of Homophones in Domain Squatting

Nick Nikiforakis
Department of Computer Science, Stony Brook University
Marco Balduzzi
Trend Micro Forward-Looking Threat Research Team
Lieven Desmet, Frank Piessens, and Wouter Joosen
DistriNet Research Group, KU Leuven

# CONTENTS

TREND MICRO LEGAL DISCLAIMER

The information provided herein is for general information and educational purposes only. It is not intended and should not be construed to constitute legal advice. The information contained herein may not be applicable to all situations and may not reflect the most current situation. Nothing contained herein should be relied on or acted upon without the benefit of legal advice based on the particular facts and circumstances presented and nothing herein should be construed otherwise. Trend Micro reserves the right to modify the contents of this document at any time without prior notice.

Translations of any material into other languages are intended solely as a convenience. Translation accuracy is not guaranteed nor implied. If any questions arise related to the accuracy of a translation, please refer to the original language official version of the document. Any discrepancies or differences created in the translation are not binding and have no legal effect for compliance or enforcement purposes.

Although Trend Micro uses reasonable efforts to include accurate and up-to-date information herein, Trend Micro makes no warranties or representations of any kind as to its accuracy, currency, or completeness. You agree that access to and use of and reliance on this document and the content thereof is at your own risk. Trend Micro disclaims all warranties of any kind, express or implied. Neither Trend Micro nor any party involved in creating, producing, or delivering this document shall be liable for any consequence, loss, or damage, including direct, indirect, special, consequential, loss of business profits, or special damages, whatsoever arising out of access to, use of, or inability to use, or in connection with the use of this document, or any errors or omissions in the content thereof. Use of this information constitutes acceptance for use in an “as is” condition.

# INTRODUCTION

Due to its critical position, Domain Name System (DNS) has, over the years, attracted many attacks targeting various parts of the protocol and the DNS infrastructure. These attacks can be grouped into the following target categories:

- Protocol weaknesses (e.g., DNS cache poisoning [14, 25])
- Vulnerable DNS server implementations (e.g., buffer overflows in BIND [20])
- User-DNS interactions

Among all of the aforementioned categories, attacks that target user-DNS interactions are the hardest to eliminate since they involve educating the entire current and future Internet population rather than technically correcting a protocol shortcoming or a software vulnerability.

One of the ways users interact with DNS is by typing domain names in their browsers' address bar. Attackers realized early on that users make spelling mistakes when typing the domain name of their desired destinations and started registering these "typo-including" domains in order to capitalize on potential incoming traffic. This practice was named "typosquatting" [19, 27] and typosquatters use these domains in a wide range of unethical and illegal ways, including showing competitors' paid ads [21] and exfiltrating user credentials through phishing [10]. In addition to typosquatting, other variations of domain squatting such as homograph attacks [11, 16] wherein attackers abuse the visual similarity of two characters from different character sets to construct domains that look like a popular authoritative domain's but lead to different destinations have been proposed over time.

This paper presents soundsquatting, a domain-squatting technique, that was uncovered while researching generic cybersquatting. Soundsquatting takes advantage of the similarity of words with regard to sound and user confusion on which word represents the desired concept. The attack is based on homophones (i.e., sets of words that are pronounced the same but are spelled differently such as *{ate, eight}).* Soundsquatting differs from typosquatting in that it does not rely on typing mistakes and that not all domains contain homophones and thus, not all domains can be soundsquatted.

To evaluate soundsquatting, an English homophone database was compiled and AutoSoundSquatter (AutoSS), a tool which, given a list of target domains, generates valid soundsquatted domains, was designed. For the Alexa top 10,000 websites, AutoSS was able to generate 8,476 soundsquatted domains, 1,823 (21.5%) of which were already registered. Through a series of automatic and manual experiments, these registered domains were categorized. Even though homophone-based domain squatting has not appeared in cybersquatting literature, its principles are known and practiced by cybersquatters, albeit less than typosquatting. Using data obtained through crawling, this paper shows that soundsquatting is being used for displaying ads on parked domains, stealing traffic from target domains, performing affiliate scams, conducting phishing attacks, and installing malicious software on unsuspecting visitors' systems.

In addition to studying the use of already-registered soundsquatted domains, 30 available ones were registered and the population of users that reached them were studied. A monthly average of 1,718 requests from real users originating from 123 countries was recorded. This shows that users are indeed susceptible to homophone confusion. Finally, six popular software screen readers were examined to show how they can all be abused to perform soundsquatting attacks against sound-dependent users who rely on text-to-speech software.

Overall, the findings show that soundsquatting can be abused in exactly the same way as typosquatting and thus should be taken into account by owners of large websites who want to protect their brand names and customers.

In sum, this paper:

- Uncovers a previously unreported domain-squatting attack type based on homophone confusion rather than on typographical mistakes, which has been dubbed "soundsquatting"
- Presents the architecture of a tool capable of automatically generating soundsquatted domains
- Presents the results of a systematic, large-scale analysis of existing soundsquatted domains targeting the Alexa top 10,000 sites, highlighting their abuse
- Actively measures the worldwide population of users who made homophone-related mistakes, confirming the validity and practicality of soundsquatting attacks
- Shows how soundsquatting can be used against sound-dependent users

# SOUNDSQUATTING

This section introduces all of the necessary terminologies for soundsquatting and describes the workings of AutoSS, a tool specially created to automatically generate soundsquatted domains, in detail. It also examines the soundsquatted domains that AutoSS generated for the Alexa top 10,000 sites.

## Terminology

Homophones are sets of words that have the same pronunciation. They can be spelled differently but have the same meaning such as *{guarantee, guaranty}* or spelled differently and have different meanings such as *{whether, weather}* and *{idle, idol, idyll}.*

Given the definition of homophones above, soundsquatting is defined as the practice of registering domain names that are homophones of authoritative ones. Soundsquatters, meanwhile, are individuals or organizations involved in soundsquatting. As in generic domain squatting, authoritative domains are those that soundsquatters target. These usually belong to high-traffic websites with millions of visitors. The more legitimate visitors a website has, the more visitors are likely to land on their soundsquatting counterparts. An authoritative domain targeted by a soundsquatting attack has been soundsquatted.

For instance, an authoritative weather site, *weatherportal.com,* can have a soundsquatted counterpart such as *whetherportal.com,* which can capture traffic to the authoritative domain should users mistakenly type “whether” instead of “weather.” Typing the wrong word and reaching the soundsquatted domain allows the soundsquatter, like generic domain squatters, to monetize visits in a wide range of unethical and illegal ways.

## Differences with Typosquatting

Before moving on to the discovery and study of soundsquatted domains, it is important to differentiate soundsquatting from typosquatting. As the term indicates, typosquatting involves “typos” (i.e., misspelling domain names, usually associated with typing mistakes). In 2006, Wang, et al., categorized the typos involved in typosquatting into five different categories [27]. Using the domain, *example.com,* and the intended URL, *www.example.com,* these are:

- **Missing-dot typos:** The dot following “www” is omitted (i.e., *wwwexample.com)*
- **Character-omission typos:** A character is omitted (e.g., *www.exmple.com)*
- **Character-permutation typos:** Consecutive characters are swapped (e.g., *www.examlpe.com)*
- **Character-replacement typos:** Characters are replaced by adjacent ones given a specific keyboard layout (e.g., *www.ezample.com,* where “x” was replaced by the QWERTY-adjacent “z”)
- **Character-insertion typos:** Characters are mistakenly typed twice (e.g., *www.exaample.com)*

Later research on typosquatting shows that in addition to the classes of typos above, domain squatters are also registering authoritative domains under different, less-popular top-level domains (TLDs) [4].

In all of the cases above, users intend to type a specific URL but accidentally mistype it, initiating a request for the wrong page before realizing they made a mistake. In contrast, in soundsquatting, users type exactly what they plan to even if their intended destination is different. The mistake occurs at the word level rather than at the character level and the substituted words are real dictionary words and not mistypes. Confusion between intended and typed words is further amplified when a domain contains a homophone that belongs to a set of same-sounding words with the same meaning. An example of this is *guarantybanking.com,* a banking website domain. As previously mentioned, "guarantee" is a homophone of "guaranty." As of this writing, *guaranteebanking.com* is parked and available for sale. In such a case, typing the "correct" domain involves memorizing a specific spelling rather than translating a concept into a word. It is also difficult to predict which spelling people who hear of "Guarantee Banking" for the first time will use.

## Generating Soundsquatted Domains

Any system built to discover domain-squatting activity requires at least the following two resources—a set of target authoritative domains and a list of rules and models to transform authoritative domains into possible squatted domains. In typosquatting's case, these rules may include domains that use the neighboring characters of every key on a specific keyboard layout and those that apply character omission, duplication, and replacement. In soundsquatting, the following resources are required:

- **Authoritative domain list:** Assuming that popular domains are targeted more than less popular ones, a list of the top 10,000 Internet websites according to Alexa was obtained. The number of unique domains contained in this list has been provided in a later section.

- **Dictionary:** Also called a "word list," this is required for extracting valid words from domain names. For instance, given a sufficiently large dictionary and the domain, *youtube.com,* an algorithm can straightforwardly search for the presence of all words in the domain, excluding the TLD, and conclude that it comprises the words "you" and "tube."

- **Transformation rules:** Apart from a dictionary, a database of English homophones is also required. A homophone database was compiled by scraping *homophone.com,* a website dedicated to homophones, along with Wikipedia's list of dialect-independent homophones [28]. The list of numbers from 1 to 100 along with their word forms (e.g., *{9, nine})* were also manually added to the homophone database. A few common idioms regularly used in Internet slang (e.g., *{you, u})* were added as well.

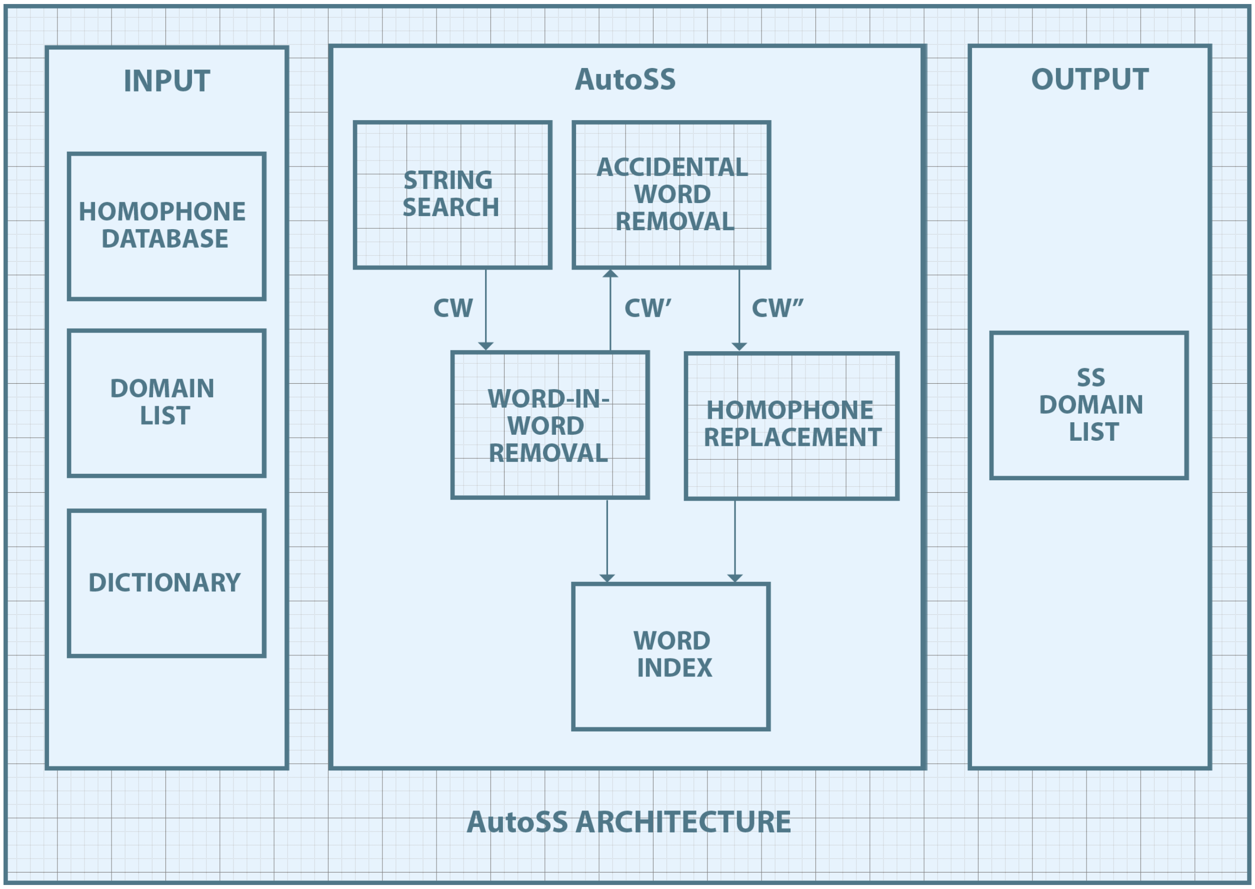


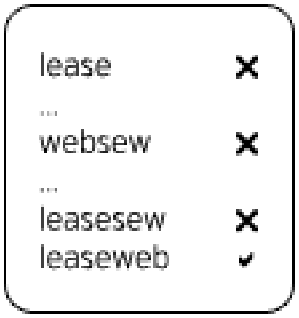
**Figure 1:** AutoSS's architecture; given a homophone database, a list of target domains, and a dictionary, AutoSS outputs a list of possible soundsquatted domains

To automatically generate soundsquatted domains, AutoSS, a tool that uses the resources above to generate valid soundsquatted domains, was created. AutoSS loads the homophone database and dictionary to memory. It then parses each entry in the Alexa list of websites to isolate the main domain from the domain extension and possible subdomains and paths. Dashes in resulting strings are perceived as indicators of word separation (e.g., *search-results.com* is split into "search" and "results" without the aid of a dictionary). Domains without dashes require performing a string search for the presence of every word in the dictionary. While this is a relatively fast process, the resulting set of candidate words (CWs) requires substantial processing mainly because of the presence of accidental words. This and other issues and the techniques used to automatically detect and resolve them are discussed in more detail below.

- **Word-in-word removal:** Consider the domain, *linkedin.com,* and the homophone set *{in, inn}.* Ideally, the tool should just discover homophones of "linked" and "in." However, a typical dictionary search will discover the words "in," "ink," "inked," "ked," "link," and "linked." The obvious next step would be to delete all words that are contained in others. The issue, however, is that while the words "in," "ink," "inked," "ked," and "link" are all contained in the word "linked," removing the word "in" from the list of CWs is wrong since it exists on its own after the word "linked." Doing so would also fail to generate soundsquatted versions such as *linkedinn.com.* To solve this problem, AutoSS was configured to work in the following manner:

- Whenever a pair of words *{a, b}* is found where *a* is included in *b, b* is replaced by another string of equal length in the domain name. Afterward, the domain name is searched again for the presence of *a.* If *a* is still found, then *a* is not deleted from the set of CWs. As such, in the example, the pair of words *{in, linked}* in *linkedin.com* is transformed to ______*in.com.* Since the word "in" is still found in the domain name, it is not removed from the list of CWs. Before proceeding, AutoSS also records the index of the word's location in the transformed domain in the Word Index component so that when words are replaced by their homophones later, the tool replaces the appropriate "in," avoiding results such as *linnkedinn.com,* which does not conform to the definition of soundsquatting since "linnked" is neither a valid dictionary word nor a homophone of any other word. At the end of this process, the list of CWs is limited to *{linked, in}* (i.e., CW' in Figure 1), which is the desired outcome.

- **Accidental word removal:** This module receives the possibly modified set of CWs from the Word-in-Word Removal module and attempts to identify and remove accidental words from the list. Consider the domain, *leaseweb.com,* which belongs to a Web-hosting service provider. The ideal word breakdown would be *{lease, web}.* Using the dictionary and selectively removing words in words, AutoSS discovers the words "lease," "sew," and "web." "Sew" is included since it is a dictionary word, which accidentally appears in the domain name, formed by the last two letters of the word "lease" and the first letter of the word "web." This problem was partially solved by attempting to exhaustively create permutations of CWs, including:



| | |
|---|---|
| lease | ✗ |
| ... | |
| websew | ✗ |
| ... | |
| leasesew | ✗ |
| leaseweb | ✓ |

This process continues until either the permutation perfectly matches the target domain name (i.e., CW" in Figure 1) or the computation times out due to the exponential nature of permutations. If time runs out before the module is finished, AutoSS falls back to the CW list after word-in-word removal.

- **Homophone replacement:** In this module, AutoSS uses the set of CWs discovered by previous modules and generates new domains by replacing one homophone with another. The module queries the homophone database for each CW. For each homophone discovered, the system generates a new soundsquatted domain by replacing the CW with a homophone. The module takes into account information found in the word index to replace the right words.

AutoSS also has a "Level" parameter that specifies the number of concurrent homophone replacements for domain names with more than one homophone discovered. Consider the case of *thepiratebay.se,* a popular Torrent tracker. AutoSS will discover the homophones *{the, thee}* and *{bay, bey}.* While these can be used to create the soundsquatted domains, *theepiratebay.se* and *thepiratebey.se,* a third domain can be generated by replacing both at the same time (i.e., *theepiratebey.se).* For this study, Level 2 was used to limit AutoSS to a maximum of two homophone replacements at a time even if a domain contains more than two homophones. While a higher level would significantly allow more combinations and generate more soundsquatted domains, three or more homophone mistakes in a single domain name are believed unlikely to occur.

- **AutoSS limitations:** Due to the flexibility of the English language and the freedom it affords with regard to wordplay, AutoSS's techniques for isolating words in domain names are necessarily heuristic based. A later section estimates the number of false positives AutoSS generates and briefly discusses possible ways to lower this number, which can be pursued in future research.

## Results

From the Alexa list of top 10,000 Internet websites, we extracted 9,926 Public Suffix + 1 domains. Given these domains and the homophone database, which contains 2,913 words with 1,337 homophone sets, AutoSS extracted a total of 6,418 homophones. Because the parameter was set to Level 2, AutoSS generated 8,476 soundsquatted domains. Interestingly, 67.3% of them did not have homophones.

The highest-ranking domain that had homophones was *youtube.com,* for which AutoSS generated the soundsquatted domains, *yewtube.com, ewetube.com,* and *utube.com.* The domain with the highest number of homophones was *wearehairy.com,* ranked 5,663 in the Alexa list of websites. It had 12 different homophones, resulting in 32 different soundsquatted domains. From the 1,337 sets of homophones, 568 (42.48%) were used at least once to generate a soundsquatted domain. Table 1 shows the top 10 homophone sets used by AutoSS on the Alexa list of websites.

**Homophone Sets AutoSS Used Most on the Alexa Top 10,000 Websites**

| Homophone Set | Number of Times Used |
|---|---|
| *{2, two, to, too}* | 735 |
| *{1, one, won}* | 300 |
| *{ere, air, aire, are, ayr, ayre, err, eyre, heir}* | 278 |
| *{four, 4, for, fore}* | 250 |
| *{bi, buy, by, bye}* | 223 |

| Homophone Sets AutoSS Used Most on the Alexa Top 10,000 Websites | |
|---|---|
| **Homophone Set** | **Number of Times Used** |
| *{do, dew, due, doe, dough}* | 208 |
| *{whirled, whorled, world}* | 156 |
| *{yew, you, ewe, u}* | 150 |
| *{cite, sight, site}* | 134 |
| *{0, zero, -xero}* | 134 |

Figure 2 correlates a website's ranking with the number of homophones found in its domain name. The scatter plot reveals that there is no significant relationship between the two, which means that, on average, low-ranking websites are just as vulnerable to soundsquatting than high-ranking ones.

This was an experimental validation of what was intuitively expected—the number of soundsquatted domains an authoritative domain has depends more on its owner's choice of words and has nothing to do with its popularity, at least among the top 10,000 Alexa websites.

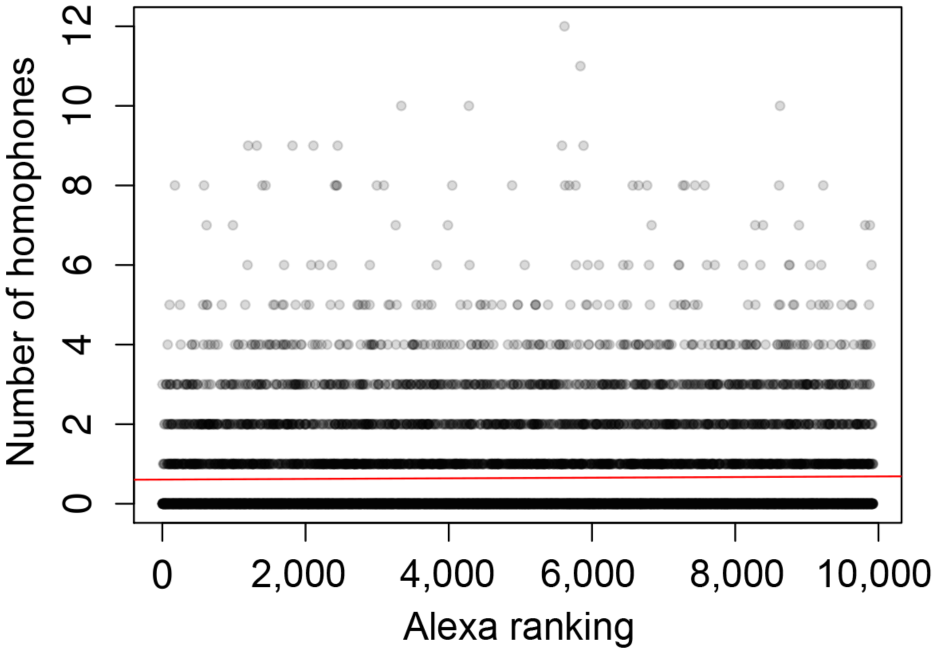


**Figure 2:** Scatter plot that shows the lack of significant correlation between a website's popularity and the number of homophones found in its domain name *(r = 0.019)*

# SOUNDSQUATTING EVALUATION

This section analyzes existing (i.e., already registered) soundsquatted domains obtained through a series of automated and manual experiments. It also categorizes them according to purpose.

## Categorization Method

As previously mentioned, AutoSS was able to generate 8,476 soundsquatted domains based on the Alexa top 10,000 websites. To find out if domain squatters are already aware of homophones and the principles of soundsquatting, a two-step process was applied to identify already-registered soundsquatted domains. First, all of the domains were tested if they would resolve to IP addresses. A domain that successfully resolves is obviously registered. Although one that does not resolve, it may still be registered but has not been assigned a valid IP address. Whois lookups were performed on the set of domains that did not resolve to IP addresses. Attempts to register them with a popular domain name registrar were also made. At the end of this process, 1,823 domains (i.e., 21.5% of the total number of domains generated) turned out to already be registered.

To classify the registered domains, a crawler based on PhantomJS [15] was used to visit each domain, waited for 10 seconds (i.e., to allow remote content to load), and took a screenshot of the page as well as recorded the HTML and final URL for later processing. The final URL was used to detect redirections from soundsquatted to different domains.

A semiautomatic approach was used to categorize each site. The screenshots of all of the pages were manually skimmed and images that looked alike were grouped together. Most of these were parked pages (i.e., pages that show ads) that were somewhat relevant to the domain names and usually advertised that the domains were for sale. Other groups comprised pages with little content, stating that the sites were "under construction." These may be placeholder pages owned by popular registrars informing their clients how to set up a website on a registered domain. Accessing some pages led to generic errors such as a 404 error. The corresponding HTML of a few domains in each group were examined and generic HTML and JavaScript signatures that could automatically categorize the remaining pages in each group were also created. Through this approach, the page-characterizing scripts eventually automatically classified 77.2% of all of the domains crawled. The remaining 417 unclassified domains were manually classified by visiting each website and carefully inspecting its source code, available Whois information, and any similarity (e.g., visual, content, and audience) with their authoritative counterparts.

## Categorization Results

Combining the results of automatic classification and manual investigation resulted in the following categories of registered soundsquatted domains:

- **Authoritative-owned domains:** Out of the 1,823 domains studied, 155 soundsquatted domains that belonged to the owners of their authoritative counterparts were identified. In a vast majority of cases, users are automatically redirected to the correct authoritative domains without warning or the appearance of additional dialogs. Redirection almost always happens through a 301/302 HTTP response status code although users can occasionally be redirected to 1–2 intermediate hosts, which in turn redirected them to the appropriate domains. In such cases, the intermediate hosts belonged to brand-protecting companies that most likely registered the domains so users who made mistakes when accessing their correct domains were redirected to their appropriate destinations.

In two instances, the owners of the authoritative domains attempted to educate their users about homophone confusion. *Myfreepaysight.com,* a soundsquatted domain for the adult site, *myfreepaysite.com,* for instance, greets visitors with a message pointing out the difference between the two domains when the latter is visited.

- **Parked/Advertising/For-sale domains:** Parked domains have been identified by prior research as the preferred means to monetize domain squatters [21, 27]. As previously mentioned, these domains do not contain real content, except ads that are constructed on demand usually by a domain-parking agency based on the words included in their names and owners' preferences. This category also includes domains that showed ads even if they are not affiliated with large domain-parking agencies (e.g., *net0.net,* a soundsquatted version of *netzero.net)* and those listed as "for sale." In sum, ad-driven domains comprise the largest chunk of existing soundsquatted domains (i.e., 954 cases or 52.3%).

- **Affiliate-abusing domains:** An examination of the soundsquatted domains that redirected users to the appropriate ones revealed that 32 abused affiliate programs. Affiliate programs promised domains small commissions for every new customer visit.

  In affiliate abuse, attackers take advantage of legitimate sites' affiliate programs by appending their own identifiers to those of unsuspecting visitors. Consider the domain, *mybrowsercache.com,* a soundsquatted version of *mybrowsercash.com.* As of this writing, every time users visit *mybrowsercache.com,* they are automatically redirected to *http://www.mybrowsercash.com/index.php?refid=312044.* Notice that a specific referrer identifier is added to the URL. This allows attackers who registered *mybrowsercache.com* to earn a commission every time users confuse "cache" for "cash." The owners of *mybrowsercash.com,* meanwhile, lose their commission.

- **Hit-stealing domains:** Analysis revealed 22 cases where attackers used soundsquatting to capture legitimate website traffic to feed to their own "business-related" domains. In a majority of cases, the authoritative and soundsquatted domains had similar content even if they had different owners. Experiments revealed that most of the soundsquatting targets were adult, online shopping, and travel websites such as:

  - *Ashemailtube.com* is a soundsquatted version of *ashemaletube.com,* a transvestite-oriented porn website. Visiting the soundsquatted domain redirects users to *trannydates.com,* a dating website that specifically caters to transvestites.

- *Video-1.com,* a soundsquatted version of the adult video portal, *video-one.com,* currently hosts an online sex shop.

- *Todomains.ru* provides domain-registration services and is a soundsquatted version of *2domains.ru,* a large Russian domain registrar.

- *Gamefive.com* is a soundsquatted version of *game5.com,* an online gaming site. The soundsquatted domain was tagged “for sale” for three years before it was turned into an online gaming site.

- *Textsail.ru* is a soundsquatted version of *textsale.ru.* Both websites sell articles and stories on a wide range of topics.

This category also includes soundsquatted domains that profit from the trustworthiness associated with their well-known and popular authoritative counterparts. In such cases, it is not necessary for the content of the soundsquatted domains to match that of their authoritative counterparts. The owners of *freemale.hu,* for instance, is probably exploiting the popularity of well-known Hungarian email service provider, *freemail.hu,* to promote their Web page in the same way that *tvto.no* abuses the popularity of the website of Norwegian channel, TV2, *tv2.no.* The soundsquatted domain redirects users to an online casino website.

- **Scam-related domains:** Soundsquatted domains can also be used for scams. Sixteen cases where soundsquatted domains were used for various scams (e.g., fake lotteries and surveys) were identified. For instance, *vhone.com,* a soundsquatted version of *vh1.com,* redirects users to a survey website that promises an opportunity to win high-end electronics in exchange for their participation. Users are then trapped in a series of redirections that constantly promise more and more prizes in exchange for divulging more and more personal information such as their names, email addresses, and mobile phone numbers.

- **Domains that promote related domains:** This category includes seven soundsquatted domains that promote materials related to the content their authoritative counterparts. *Teambeechbody.com* is a soundsquatted version of *teambeachbody.com,* an online fitness club where people can subscribe as “fitness coaches” and gain commission for successfully coaching users. As of this writing, visiting the soundsquatted domain redirects users to the pages of specific coaches in *teambeachbody.com,* giving the coaches better chances of getting selected over others on the website. In another case, the soundsquatted domain, *rednovel.com,* redirects users to *http://www.lvse.com/site/readnovel-com-3550.html, a readnovel.com* (i.e., the authoritative domain) page that contains a safety score, user comments, and a list of similar websites.

- **Other domains:** Analysis revealed that six soundsquatted domains were used for malicious purposes (e.g., to install malware and acquire personal information). *Movreel.com,* a free-of-charge moving-streaming service provider is being soundsquatted by *movreal.com.* At a first glance, *movreal.com* appears to be another movie-streaming service provider, as it asks users to download a browser plug-in (i.e., *AVS_Media_Player.exe)* in oder to watch videos. The plug-in is, however, malicious and detected by most security vendors as a Solimba variant (i.e., an installer of other malicious software and adware). Similarly, *utube.com,* a soundsquatted version of *youtube.com,* uses videos to social-engineer users into first divulging personal information then, depending on their browsers, installs a browser extension. Mozilla® Firefox® users then see unwanted search results and pop-up messages, apart from running the risk of becoming part of statistics gathering.

Two domains that likely acquire private user information, particularly email credentials, were found. One of these is *innbox.lv,* a soundsquatted version of the well-known Latvian service provider's domain, *inbox.lv.* Both websites offer free email accounts. Two soundsquatted domains were also involved in phishing campaigns against e-commerce and business-related websites.

Overall, 1,037 (56.88%) of the 1,823 registered soundsquatted domains were tagged "malicious." Out of the remaining domains, 155 belonged to their authoritative counterparts' owners; 300 were owned by different legitimate organizations; and 331 were offline, showed HTTP errors, or were under construction when visited.

## User Characterization

In previous sections, the registered soundsquatted domains were categorized according to purpose. Let us now look at users who, due to homophone confusion, landed on soundsquatted domains.

As previously mentioned, AutoSS generated 8,476 soundsquatted versions of the Alexa top 10,000 websites. Among them, 1,823 (21.5%) were already registered, leaving 6,653 unregistered. To actively measure the global user population and assess the viability of soundsquatting attacks, we registered our own soundsquatted domains and monitored the requests they received. Due to the lack of prior soundsquatting research, there was no objective or historical way to assess which among the unregistered domains would attract more users than others. As such, the list of available soundsquatted domains were manually examined. A total of 30 domains covering a wide range of soundsquatting techniques were chosen for further study.

**Soundsquatted Domains Studied to Determine User Characteristics**

| Authoritative Domain | Homophone Pair | Soundsquatted Domain | Number of Human Requests per Month |
|---|---|---|---|
| *thefreedictionary.com* | *{the, thee}* | *theefreedictionary.com* | 283 (39.86%) |
| *fc2.com* | *{2, too}* | *fctoo.com* | 165 (44.84%) |
| *jimdo.com* | *{do, doe}* | *jimdoe.com* | 150 (38.27%) |

| Soundsquatted Domains Studied to Determine User Characteristics | | | |
|---|---|---|---|
| Authoritative Domain | Homophone Pair | Soundsquatted Domain | Number of Human Requests per Month |
| *turbobit.net* | *{bit, bitt}* | *turbobitt.net* | 132 (36.07%) |
| *leboncoin.fr* | *{coin, quoin}* | *lebonquoin.fr* | 110 (74.32%) |
| *adserverplus.com* | *{ad, add}* | *addserverplus.com* | 98 (60.49%) |
| *profitclicking.com* | *{profit, prophet}* | *prophetclicking.com* | 56 (48.28%) |
| *hostgator.com* | *{gator, gaiter}* | *hostgaiter.com* | 45 (45.92%) |
| *sitesell.com* | *{sell, cel}* | *sitecel.com* | 44 (40.00%) |
| *discuz.net* | *{disc, disk}* | *diskuz.net* | 43 (40.19%) |
| *tube8.com* | *{8, ait}* | *tubeait.com* | 42 (43.30%) |
| *clixsense.com* | *{sense, scents}* | *clixscents.com* | 40 (44.44%) |
| *a8.net* | *{8, eight}* | *aeight.net* | 48 (43.24%) |
| *newegg.com* | *{new, gnu}* | *gnuegg.com* | 37 (36.63%) |
| *redtubelive.com* | *{red, read}* | *readtubelive.com* | 44 (51.76%) |
| *fiverr.com* | *{err, air}* | *fivair.com* | 33 (37.93%) |
| *exoclick.com* | *{click, clique}* | *exoclique.com* | 32 (45.71%) |
| *theglobeandmail.com* | *{mail, male}* | *theglobeandmale.com* | 35 (38.46%) |
| *pastebin.com* | *{bin, been}* | *pastebeen.com* | 35 (39.77%) |
| *ku6.com* | *{6, sics}* | *kusics.com* | 28 (33.33%) |
| Total | | | 1,718 |

The first three columns of the table above show 20 of the 30 authoritative domains studied, the homophone pairs used, and their soundsquatted versions. While three of target domains above could be associated with typosquatting (e.g., *theefreedictionary.com,* the rest radically differ from domains that researchers have, over the years, associated with typosquatting (e.g., *prophetclicking.com).* Most of the domains were registered in December 2012 while others were registered in March 2013. To present a uniform view of traffic, the monthly average number of requests received by each domain was obtained until December 11, 2013.

All domains, subdomains, and requests for specific file paths resolved to a single blank page while recording each request's details in a set of Apache log files. Users were not automatically redirected to the authoritative domains they sought to avoid reinforcing the behavior of typing the wrong domains. They were instead made aware of their mistake. (Ethical considerations regarding the experiment are discussed in the Appendix).

The last column shows the monthly average number of human requests received during the period of monitoring, along with the percentage of human requests among all requests. To assess soundsquatting's impact on human behavior, bot visits had to be separated from human visits. There is no single, generic technique that can perfectly separate bot from human visits. If such a technique exists, attackers would already be using it to perfectly evade security researchers by detecting all high-interaction honeypots and never presenting them with malicious code.

In this paper, requests that had nonstandard user agents were identified during the preliminary manual inspection. Using keywords extracted from these requests, we assembled a set of nine generic identifiers such as "spider," "bot," and "crawl" that many bots have in common. In addition to these generic identifiers, 707-specific bot signatures from *useragentstring.com* were scraped. As a result, if a user agent contains any of the 716 bot signatures in the predetermined set, a request was classified as a "bot request." To account for bots that do not identify themselves, each requester's IP address was also queried based on the blacklist provided by *stopforumspam.com,* a database with hundreds of thousands of IP addresses that belong to known forum-spamming bots. Finally, each address was queried based on a list of IP addresses used by well-known search engine spiders [1].

Results show that the 30 soundsquatted domains monitored received an average of 1,718 human requests per month. The total monthly number of requests was 4,150. The domain that received the highest number of hits, *theefreedictionary.com,* can also be considered a typosquatting candidate and so naturally attracted more traffic than the domains that were just soundsquatted. Apart from requests for each website's main page, many requests for subdomains within each domain were also recorded. Let us consider *jimdo.com,* a Web application that allows users to create their own websites and host them on its subdomains. The *jimdoe.com* logs contained requests for 176 subdomains associated with personal websites such as *awesomegrizzlybears.jimdoe.com, karatedojo-oppeln.jimdoe.com,* and *armaniwoe.jimdoe.com,* all valid subdomains under *jimdo.com.* These visits show that even though people can accurately type relatively long and obscure subdomains, they can still confuse homophones.

Geolocating the IP addresses of all requests showed that, while users from 42 countries crawled the chosen domains, human requests originated from 123 different countries. This shows that users from all countries are prone to homophone confusion and thus vulnerable to soundsquatting attacks.

In general, each soundsquatted domain received between two and 283 human requests per month. While these numbers are not incredibly large and probably smaller than those obtained by popular typosquatted domains, soundsquatting and typosquatting are not competing techniques. They instead complement each other in domain squatters' arsenal. Since this is the first soundsquatting study, domains with homophone replacements ranging from more likely to less likely were registered. Careful attackers, however, can target domains better and thus acquire more visitors at less cost.

Finally, a significant number of emails (e.g., social networking invitations, product shipment notifications, email-account-creation credential notifications, mobile phone service bills, etc.) was sent to the soundsquatted domains monitored. It was evident in all cases that the emails were meant to be sent to accounts that belonged to the legitimate domains that were soundsquatted but were missent due to homophone confusion. Receipt of these emails further shows that businesses and users are indeed vulnerable to soundsquatting attacks.

# SOUND-DEPENDENT USERS

This section describes a soundsquatting attack that can victimize people who rely on sound when using computers.

According to the Word Health Organization, the world currently has 285 million visually impaired people, 39 million of whom are blind [2]. Severely visually impaired people cannot properly interact with computers without the help of assistive technologies. The two most popular assistive technologies for the visually impaired are Braille displays and screen readers [9]. Both assistive technologies convert content otherwise-consumed by sight into something that can be consumed by touch or sound instead. Considering the definition of homophones and their relation to soundsquatting, a new attack type can clearly be seen.

Users that depend on screen readers to consume content in emails, Web pages, social media messages, or instant messages are vulnerable to accessing links that point to soundsquatted domains. Soundsquatted domains will be “read” near-identical with authoritative domains, giving the visually impaired no reason not to access the link offered. While Braille displays are not vulnerable to this attack, the fact that around 90% of the visually impaired live in developing countries combined with the high cost of Braille displays suggest that due to limited resources and possible portability issues, screen readers are used more than Braille devices. Apart from the visually impaired, hundreds of thousands of smartphone users use personal assistant software such as Apple’s Siri®, which has text-to-speech capabilities when they engage in other activities (e.g., driving or running) that make it hard to operate their smartphones.

To test this theory, an email with two links, one pointing to *youtube.com* and another to *yewtube.com* was sent. Five popular free screen readers (i.e., built-in screen readers of Windows® XP, Windows 7, and Mac OS X; Linux-based, open source ORCA [23]; and Thunder screen reader [26]) were used. A text message with the same information was also sent to an Android™ smartphone with Skyvi [5], a popular Siri-like application used by more than 260,000 people.

In all six cases, the two links sounded identical to each other, which means that a sound-dependent person would have no means to tell a legitimate link from a malicious one. To further exacerbate the issue, soundsquatting attacks can also work with pseudohomophones (i.e., combinations of characters that are not real dictionary words but are purposefully constructed to sound like real words such as *{joke, joak}* [24]). Pseudosoundsquatted domains can be crafted even for target domains that do not contain homophones such as *phacebook.com* and *phaceboocc.com).*

Due to the potentially large number of domain variations and the specificity of this attack type, the responsibility of protecting sound-dependent users lies in the hands of text-to-speech software developers. One way of protecting against this threat is for text-to-speech software to switch to “spelling mode” whenever a link is encountered so users know they are accessing the right links and can avoid visiting malicious websites.

# LIMITATIONS AND FUTURE RESEARCH

While AutoSS accounts for many corner cases when attempting to identify words comprising domain names, there is, unfortunately, still room for false positives (i.e., domains that do not conform to the definition of soundsquatting and the concept behind it). For instance, there are many domains in the Alexa top 10,000 websites that do not have English words such as *laredoute.fr,* a French e-shop. AutoSS uses an English dictionary and will identify "lare," "do," and "ute" from the domain name. Its accidental word removal module will successfully combine these words to form "laredoute" and use them in the homophone replacement database, resulting in improbable domains such as *laredewute.fr.* Already-available typosquatting systems do not suffer from such a problem since they operate at the character level [21, 27] unlike soundsquatting tools such as AutoSS, which operate at the word level.

To estimate the number of false positives, 424 (5%) of the soundsquatted domains generated were randomly sampled and each homophone replacement was manually examined to ensure that none of the domains are false positives. At the end of this process, 80 false positives out of the 424 domains investigated (i.e., 18.9% with a margin of sampling error ±4.75%) were identified. While the number of false positives is not negligible, the study's main purpose was to investigate a previously unreported domain-squatting technique and evaluate its practicality and adoption for the Web.

Lack of punctuation in domain names makes identifying the language they are written in challenging. One way around this problem is to actually inspect a site's main page, characterize its language, and assume that its domain name contains words in the same language. The researchers will leave the exploration of this and other techniques to reduce false positives to future work.

# RELATED WORK

To the best of our knowledge, this paper is the first to uncover the use of homophones to perform domain squatting and systematically study its adoption as well as users' susceptibility to attacks.

Domain squatting is the first form of cybersquatting that involves registering domains with trademarks that belong to other people or companies before their rightful owners have the chance to do so [6, 8, 13]. Domain squatting later evolved into typosquatting [8, 21, 27] or the act of registering domains that are mistypes of popular authoritative domains. Its beginnings can be traced back to 1999 through the Anti-Cybersquatting Consumer Protection Act (ACPA), which mentioned URLs that were "sufficiently similar to a trademark of a person or entity."[3]

Apart from typosquatting, other less popular types of domain squatting (e.g., domains that abuse the visual similarity of characters in different character sets [11, 16] and capture traffic originating from erroneous bit-flips in user devices [7, 22]) also exist.

# CONCLUSION

This paper uncovered a new type of domain squatting using similar-sounding words rather than relying on typographical mistakes. Dubbed “soundsquatting,” it described a system that automatically generates soundsquatted domains and showed that attackers are already familiar with the concept of soundsquatting, abusing domains in ways similar to known types of domain squatting. Registering our own soundsquatted domains allowed us to show that it is possible for well-selected soundsquatted domains to attract hundreds of human visitors every month. The relationship between text-to-speech software and soundsquatting was also briefly examined. This paper also showed that attackers could abuse text-to-speech software to trick sound-dependent users into visiting malicious soundsquatting and pseudosoundsquatted domains. Overall, the paper’s findings verify the practicality of soundsquatting and show that homophone confusion should be accounted for by website owners and registrars as well as in cybersquatting countermeasures.

Acknowledgment: The researchers would like to thank anonymous reviewers for their valuable comments. This research was financially aided by the Prevention Against Crime Programme of the European Union (B-CCENTRE); the Research Fund KU Leuven; and the EU FP7 Projects, NESSoS and STREWS.

# REFERENCES

1. Daniel Kramer. (2011). *IP Addresses of Search Engine Spiders.* Last accessed September 24, 2014, http://iplists.com/.

2. WHO. (2014). *World Health Organization.* "Visual Impairment and Blindness." Last accessed September 24, 2014, http://www.who.int/mediacentre/factsheets/fs282/en/.

3. Cybertelecom. (March 5, 2014). *Cyber Telecom.* "Anti-Cybersquatting Consumer Protection Act." Last accessed October 8, 2014, http://www.cybertelecom.org/dns/acpa.htm.

4. A. Banerjee, D. Barman, M. Faloutsos, and L. N. Bhuyan. (2008). *Proceedings of IEEE INFOCOM.* "Cyberfraud Is One Typo Away."

5. BlueTornado Inc. (2012). *Skyvi.* Last accessed September 24, 2014, http://www.skyviapp.com.

6. S. E. Coull, A. M. White, T. F. Yen, F. Monrose, and M. K. Reiter. (2010). *IFIP SEC '10.* "Understanding Domain Registration Abuses."

7. A. Dinaburg. (July 2011). *Proceedings of BlackHat Security.* "Bitsquatting: DNS Hijacking Without Exploitation."

8. B. Edelman. (2003). "Large-Scale Registration of Domains with Typographical Errors."

9. Even Grounds Inc. (2007–2013). *Even Grounds.* "How Do Blind People Use the Computer." Last accessed September 24, 2014, http://www.evengrounds.com/blog/how-do-blind-people-use-the-computer.

10. Rik Ferguson. (May 21, 2009). *Countermeasures.* "Tvviter Typosquatting Phishing Site." Last accessed September 24, 2014, http://countermeasures.trendmicro.eu/tvviter-typosquatting-phishing-site/.

11. E. Gabrilovich and A. Gontmakher. (February 2002). *Communications of the ACM, 45 (2):128.* "The Homograph Attack."

12. G. Gee and P. Kim. (September 2011). "Doppelganger Domains." Last accessed September 24, 2014, http://www.wired.com/images_blogs/threatlevel/2011/09/Doppelganger.Domains.pdf.

13. J. Golinveaux. (1998–1999). *University of San Francisco Law Review 33 U.S.F. L. Rev.* "What's in a Domain Name: Is Cybersquatting Trademark Dilution?"

14. A. Herzberg and H. Shulman. (2013). *CNS '13.* "Fragmentation Considered Poisonous, or: One-domain-to-rule-them-all.org."

15. A. Hidayat. "PhantomJS: Headless WebKit with JavaScript API."

16. T. Holgers, D. E. Watson, and S. D. Gribble. (2006). *Proceedings of USENIX ATC.* "Cutting Through the Confusion: A Measurement Study of Homograph Attacks."

17. M. Jakobsson, P. Finn, and N. Johnson. (March–April 2008). *Security & Privacy, IEEE, 6 (2):66–68.* "Why and How to Perform Fraud Experiments."

18. M. Jakobsson and J. Ratkiewicz. (2006). *WWW '06.* "Designing Ethical Phishing Experiments: A Study of (ROT13) rOnl Query Features."

19. D. Kesmodel. (2008). "The Domain Game: How People Get Rich from Internet Domain Names."

20. R. McMahon. (2000). "BIND 8.2 NXT Remote Buffer Overflow Exploit."

21. T. Moore and B. Edelman. (2010). *Financial Cryptography and Data Security, 175–191.* "Measuring the Perpetrators and Funders of

Typosquatting."

22. N. Nikiforakis, S. V. Acker, W. Meert, L. Desmet, F. Piessens, and W. Joosen. (2013). *WWW '13, 989–998.* "Bitsquatting: Exploiting Bit-Flips for Fun, or Profit?"

23. Orca: A Free, Open Source, Flexible, and Extensible Screen Reader.

24. M. S. Seidenberg, A. Petersen, M. C. MacDonald, and D. C. Plaut. (1996). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22 (48–62).* "Pseudohomophone Effects and Models of Word Recognition."

25. J. Stewart. (2003). "DNS Cache Poisoning—The Next Generation."

26. ScreenReader.net: Freedom for Blind and Visually Impaired People.

27. Y. M. Wang, D. Beck, J. Wang, C. Verbowski, and B. Daniels. (2006). *SRUTI '06.* "Strider Typo-Patrol: Discovery and Analysis of Systematic Typosquatting."

28. *Wiktionary.* (May 24, 2014). "List of Dialect-Independent Homophones." Last accessed September 24, 2014, http://en.wiktionary.org/wiki/Appendix:List_of_dialect-independent_homophones.

# APPENDIX

## Ethical Considerations

Registering soundsquatted domains and receiving user traffic to them may raise ethical concerns. However, analogous to the real-world experiments conducted by Jakobsson, et al. [17, 18], we believe that conducting realistic experiments is the only way to reliably estimate the success rate of attacks in the real world. Moreover, we believe that our findings will help websites protect their brands and customers.

The data collected for the experiments includes each request's time stamp; the IP address of the host performing the request; domain, path, and GET parameters; and user agents provided by the Apache Web server. This data is collected by every Web server in standard server logs and many Web developers even share this information with third parties such as Google Analytics™ for the purpose of gathering usage statistics. The server logs were only accessible to the authors of this paper. Similarly, the emails were all collected in a single, password-protected email account of one of the authors. We did not attempt to extract any information from these emails nor trace their senders. Gee and Kim performed a similar experiment in 2011, capturing emails through typosquatting domains and released statistics to the research community as a demonstration of the dangers of typosquatting [12].

Trend Micro Incorporated, a global leader in security software, strives to make the world safe for exchanging digital information. Our innovative solutions for consumers, businesses and governments provide layered content security to protect information on mobile devices, endpoints, gateways, servers and the cloud. All of our solutions are powered by cloud-based global threat intelligence, the Trend Micro™ Smart Protection Network™, and are supported by over 1,200 threat experts around the globe. For more information, visit www.trendmicro.com.

TREND MICRO™

Securing Your Journey
to the Cloud

225 E. John Carpenter Freeway, Suite 1500
Irving, Texas 75062 U.S.A.

Phone: +1.817.569,8900