**域名抢注攻击调研**

# 1背景及原理

域名系统( Domain Name System，DNS) 是互联网中重要的基础设施，由于 IP 地址的数字特征不便于人们识记，几乎所有的互联网应用都需要通过DNS 完成域名到 IP 地址的转换，因此域名在互联网中的地位非常重要。

在域名劫持攻击中，域名抢注是一个很多人研究的领域。在网络成立之初，人们会注册与知名公司、知名产品、知名商标相关的域名，然后以更高的价格将域名卖回其合法所有者来牟利，这种做法被称为域名抢注（Cybersquatting）。

当互联网变得越来越流行并且大公司已经为他们的业务购买了适当的域名时域名抢注攻击又有了新的发展，分别为错拼域名抢注（Typosquatting）、比特域名抢注（Bitsquatting）、同音域名抢注（Soundsquatting）、组合域名抢注（Combosquatting）。

# **1.1 Typosquatting原理**

Typosquatting是指互联网用户在浏览器中输入网址时出现拼写错误被网络犯罪分子误导至恶意网站，而不是用户实际想访问的站点。如用户输入paypap.com而不是paypal.com。

# **1.2 Bitsquatting原理**

Bitsquatting是指网络犯罪分子会对一些因为计算机内存、CPU缓存由于环境，或者制造缺陷，在传输过程中域名的二进制编码存在比特位翻转而导致实际访问的域名并非正常输入的域名进行抢注，如mic2osoft.com因为二进制比特的变化导致不是microsoft.com。

# **1.3 Soundsquatting原理**

Soundsquatting是指网络犯罪分子利用了单词的发音相似性抢注了一些知名域名的同音域名，用户很可能因为发音上的相似性而被误导至这些恶意网站，如用户会键入 whetherportal.com而不是weatherportal.com。

# **1.4 Combosquatting原理**

Combosquatting是指网络犯罪分子把知名域名的名字原封不动借用过来，不作任何改动，然后在其前面或者后面加上一些字符，如youtube-login.com而不是youtube.com,，如baidu-jabber.com而不是baidu.com。

# 2现状

研究人员发现，网络犯罪分子通常会抢注一些过期的旧域名，以便利用这些域名的剩余信任度，如某些恶意软件使用金融机构所托管的旧域名[1]，僵尸网络将可信任的旧域名伪装成其C＆C控制服务器用于与其他主机之间的通信[2]，甚至劫持整个自治系统[3,4]。

除了利用过期的旧域名，在Typosquatting中，攻击者会利用各种非法不道德的方法来使用这些域名，如展示竞争对手的广告[5]、网络钓鱼[6]、隐藏在弹出广告中的恶意诈骗[25-26]。Nikiforakis等[7]表明Typosquatting也会发生在远程脚本中，开发人员输入了错误的域名，从而导致其网站容易受到恶意脚本的注入。Wang等[18]研究表明大多数Typosquatting域名会将访问其的流量指向或重定向到其域名托管机构。

在Bitsquatting中，Nikiforakis等[8]发现比特域名主要依赖与硬件故障而不是人为因素，根据Verisign的统计[27]，每1M-10M个DNS请求就会出现一个比特反转导致错误的请求。

在Soundsquatting中，Nikiforakis等[9]发现同音域名抢注已被用于在网页的广告中，企图从目标域名中获取用户流量，进行大规模诈骗以及进行网络钓鱼攻击。

在Combosquatting中，Kintis等[24]发现这样的抢注方式已被用于网络钓鱼，恶意软件分发，商标滥用，APT攻击中。

# 3方法

# **3.1 Typosquatting**

在Typosquatting的研究中，大部分研究人员为了获取Typosquatting域名集以便后续研究，通常选取 Alexa网站排名中topN的网站的二级域名作为研究对象，对这些域名分别进行指定编辑距离的字符替换、重复、删除、互换顺序等操作主动生成数据集，如图一所示。



图一Typosquatting

Moore等[10]为了更贴近用户输入错误的实际情况，结合键盘上字符排列特征，对编辑距离中的字符替换操作进行改进，提出手指距离替换，即对目标域名中的某个字符只用其对应键盘上的相邻字符进行替换，如facebojk和facebook，o、j、k在标准键盘上相邻。

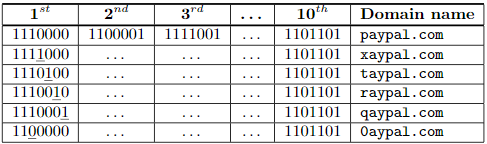
除了主动生成数据集，Khan等[11]针对域名服务器，收集该域名服务器下一个时间段内用户请求的所有域名。文章认为用户访问到错误的域名网站后，会退出并继续访问目标网站，而访问到目标网站后，不会继续访问错误的域名网站，基于这一思想，其引入条件概率模型划分目标域名数据集和Typosquatting数据集。

在拿到Typosqutting的数据集后，需要证明域名是否已经被注册过，Szurdi等[12]将数据集与Zone文件中的域名进行比较，直接筛选出已经注册过的Typosquatting域名。在没有Zone文件时，Nikiforakis等[9]对数据集中的每个域名进行解析，如果能解析到IP地址，则证明已被注册，否则会进一步查询whois记录信息进行确认。

在得到注册域名数据集后，研究人员通常会使用爬虫的方式去获取每个域名随对应的网页特征，然后利用机器学习与人工分析结合的方法对Typosquatting域名特征进行分析。Moore等[10]从页面角度研究错拼抢注域名，将爬虫获取的 URL 与知名广告服务的 URL 进行比对，如果该 URL 存在于广告服务中，直接对域名进行标记。Agten等[13]使用感知哈希算法和局部敏感哈希算法，分别对页面截图和页面结构特征进行聚类，然后结合域名页面内容和重定向信息，研究错拼抢注域名的用途。Halvorson等[14]分析biz顶级域名下的所有域名，查找这些域名在com顶级域名下存在的相同二级域名，根据域名的Zone文件、DNS请求、Whois信息和Web文本特征，比较不同顶级域名下的相同二级域名的差异。

# **3.2 Bitsquatting**

研究人员一般生成Bitsquatting域名的规则如下图二，即将域名转换成二进制编码后，对二进制编码的每一位进行翻转，得到该域名不同的Bitsquatting域名。



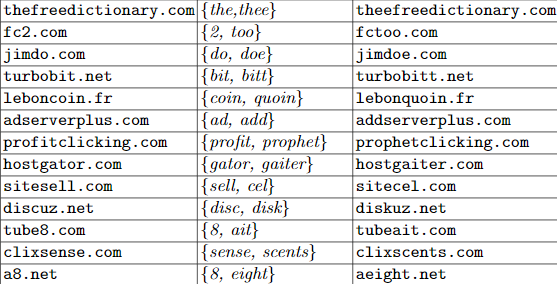
图二Bitsquatting

在Bitsquatting的研究中，与Typosquatting一样一般研究人员普遍会利用最受欢迎的域名作为数据集进行研究，Nikiforakis等[8]针对Alexa排名top500的域名，利用程序自动生成与其相关的符合Bitsquatting规则的域名。Vissers等[15]研究的是域名服务器是否存在Bitsquatting，其瞄准最著名2LD的Zone文件，提取了前10000个域名服务器的数据进行研究。Zeng等[16]选取Alexa排名top300的域名同时选取CN[17]top300的域名以及从某个ISP获取了某段时间内DNS请求的top300作为数据集，然后对每个域名的每一位进行位翻转，形成Bitsquatting数据集。

在得到数据集后，Nikiforakis等[8]利用爬虫对生成的所有域名进行访问，尝试解析其IP。Vissers等[15]对DNS记录进行逐一分析，从被动记录中寻找潜在的发生Bitsquatting的域名。Zeng等[16]逐一对生成的域名进行访问来确定其可访问性。

# **3.3 Soundsquatting**

在Soundsquatting的研究中，研究人员一般根据Alexa排名topN的网址提取出相应的英文单词，然后根据homophone.com网站提供的同音异义词以及维基百科的同音异义词列表，构造出Soundsquatting域名。同音异义词域名一般规则如下图三所示。

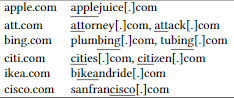


图三Soundsquatting

在生成了Soundsquatting域名后，与Typosquatting中研究人员采取的方法类似，利用爬虫[9]等方法对每一个域名进行解析，看其是否被注册。

# **3.4 Combosquatting**

在Combosquatting的研究中，研究人员会根据Alexa排名，对知名网站的域名进行一定的扩充，如下图四所示。或者从被动DNS记录中筛选域名，如果一个域名完整包含了某个知名域名，那么它就可以算是 Combosquatting域名。



图四Combosquatting

Typosquatting的构造方式是有限的，而Combosquatting是无限的。事实上，检测出的Combosquatting域名也比Typosquatting域名数量多得多，高出两个数量级。虽然Combosquatting域名具有无限的构造空间，但是实际上这些注册者会精心构造域名、精心选词。通过分词研究人员发现，90%的Combosquatting域名只有2-3个英文单词，因为不会有用户愿意敲非常长的域名，非常长的域名通常是可疑的。

Kintis等[24]在收集到Combosquatting域名后，利用爬虫对这些域名进行IP解析以及网页特征获取，在获取网页特征后，对Combosquatting域名的使用进行分类，如网络钓鱼、恶意软件、APT攻击等。

# 4数据集

|  |  |  |
| --- | --- | --- |
| Typosquatting： | 原始数据集 | 新数据集 |
| Moore[10] | AlexaTop6000中选取了3 264个域名 | 利用3.1中的规则针对原始数据集中的域名进行生成，共生成938000个域名用于后续分析 |
| Khan[11] | 美国大型公立大学网络的PDNS记录  某企业一个月的PDNS记录 | 利用类似3.1中的规则对原始数据集中的域名进行生成，共生成大约13500个域名 |
| Szurdi[12] | .com域名的Zone文件约1.06亿个域名  AlexaTop100万  12个不同的域名黑名单检查.com的错字域名（为了揭示错别字域的恶意使用，） | 利用类似3.1中的规则对原始数据集中的域名进行生成，共生成大约470万域名 |
| Agten[13] | AlexaTop500域名 | 利用类似3.1中的规则对原始数据集中的域名进行生成，共生成大约28179域名 |
| Zeng[16] | 将AlexaTop300域名、  CNTop300域名、  某ISP的DNS记录中某段时间内的Top300域名  去重共得到786个 | 利用3.1中的规则针对原始数据集中的域名进行生成，共生成24684个域名用于后续分析 |
| Wang[18] |  |  |

|  |  |  |
| --- | --- | --- |
| Bitsquatting： | 原始数据集 | 新数据集 |
| Nikiforakis[8] | AlexaTop500域名 | 利用3.2中的位翻转规则针对原始数据集中的域名进行生成，利用爬虫在270天内确定域名是否被注册，共确定了5366个域名 |
| Zeng[16] | 将AlexaTop300域名、  CNTop300域名、  某ISP的DNS记录中某段时间内的Top300域名  去重共得到786个 | 利用3.2中的位翻转规则针对原始数据集中的域名进行生成，共生成6307个域名用于后续分析 |

|  |  |  |
| --- | --- | --- |
| Soundsquatting： | 数据集 | 新数据集 |
| Nikiforakis[9] | AlexaTop10000域名 | 利用同音异义词的规则，针对原始数据集中的域名进行生成，共生成8476个域名用于后续分析 |
| Zeng[16] | 将AlexaTop300域名、  CNTop300域名、  某ISP的DNS记录中某段时间内的Top300域名  去重共得到786个 | 利用同音异义词的规则以及中文拼音相近的规则，针对原始数据集中的域名进行生成，共生成2157个域名用于后续分析 |

|  |  |  |
| --- | --- | --- |
| Combosquatting： | 数据集 | 新数据集 |
| Kintis[24] | 5年中某运营商提供的被动DNS记录、  1年的主动DNS记录、  4个公开域名黑名单 | 在数据集中利用3.4中的规则进行分析，共得到110万个域名用于后续分析。 |
| Zeng[16] | 被动DNS记录 | 利用50000个常用英语单词与其获取的786个流行域名进行组合，在数据集中进行检索，共得到9600个域名用于后续分析 |

**上述研究中的新数据集均未开源。Github有很多typosquatting域名的检测工具[20,22]以及生成工具[21,23]。**

# 5参考文献

1. Tyler Moore and Richard Clayton. 2014. The Ghosts of Banking Past: Empirical Analysis of Closed Bank Websites. In Financial Cryptography and Data Security.
2. Chaz Lever, Robert Walls, Yacin Nadji, David Dagon, Patrick McDaniel, and Manos Antonakakis. 2016. Domain-Z: 28 Registrations Later. In Proceedings of the 37th IEEE Symposium on Security and Privacy.
3. Johann Schlamp, Georg Carle, and Ernst W Biersack. 2013. A forensic case study on as hijacking: The attacker’s perspective. ACM SIGCOMM Computer Communication Review 43, 2 (2013), 5–12.
4. Johann Schlamp, Josef Gustafsson, Matthias Wählisch, Thomas C Schmidt, and Georg Carle. 2015. The abandoned side of the Internet: Hijacking Internet resources when domain names expire. In International Workshop on Tra?c Moni-toring and Analysis. Springer, 188–201.
5. T. Moore and B. Edelman. Measuring the perpetrators and funders of typosquat-ting. In Financial Cryptography and Data Security, pages 175{191, 2010.
6. R. Ferguson. Tvviter Typosquatting Phishing Site. http://countermeasures. trendmicro.eu/tvviter-typosquatting-phishing-site/.
7. Nikiforakis, N., Invernizzi, L., Kapravelos, A., Van Acker, S., Joosen, W., Kruegel, C., Piessens, F., and Vigna, G. You Are What You Include: Large-scale Evaluation of Remote JavaScript Inclusions. In Proceedings of the ACM Conference on Computer and Communications Security (CCS) (2012).
8. Nikiforakis N, Van Acker S, Meert W, et al. Bitsquatting: Exploiting bit-flips for fun, or profit?[C]//Proceedings of the 22nd international conference on World Wide Web. 2013: 989-998.
9. Nikiforakis N, Balduzzi M, Desmet L, et al. Soundsquatting: Uncovering the use of homophones in domain squatting[C]//International Conference on Information Security. Springer, Cham, 2014: 291-308.
10. Moore T, Edelman B. Measuring the perpetrators and funders of typosquatting[C]//International Conference on Financial Cryptography and Data Security. Springer, Berlin, Heidelberg, 2010: 175-191.
11. Khan M T, Huo X, Li Z, et al. Every second counts: Quantifying the negative externalities of cybercrime via typosquatting[C]//2015 IEEE Symposium on Security and Privacy. IEEE, 2015: 135-150.
12. Szurdi J, Kocso B, Cseh G, et al. The long “taile” of typosquatting domain names[C]//23rd {USENIX} Security Symposium ({USENIX} Security 14). 2014: 191-206.
13. Agten P, Joosen W, Piessens F, et al. Seven months' worth of mistakes: A longitudinal study of typosquatting abuse[C]//Proceedings of the 22nd Network and Distributed System Security Symposium (NDSS 2015). Internet Society, 2015.
14. Halvorson T, Szurdi J, Maier G, et al. The BIZ top-level domain: ten years later[C]//International Conference on Passive and Active Network Measurement. Springer, Berlin, Heidelberg, 2012: 221-230.
15. Vissers T, Barron T, Van Goethem T, et al. The wolf of name street: Hijacking domains through their nameservers[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017: 957-970.
16. Zeng Y, Zang T, Zhang Y, et al. A comprehensive measurement study of domain-squatting abuse[C]//ICC 2019-2019 IEEE International Conference on Communications (ICC). IEEE, 2019: 1-6.
17. "CNTopSites."http://www.alexa.cn/siterank/.
18. Wang Y M, Beck D, Wang J, et al. Strider Typo-Patrol: Discovery and Analysis of Systematic Typo-Squatting[J]. SRUTI, 2006, 6(31-36): 2.2-2.3.
19. List of dialect-independent homophones. http://en.wiktionary.org/wiki/ Appendix:List\_of\_dialect-independent\_homophones.
20. https://github.com/monkeym4ster/DomainFuzz
21. https://github.com/zntrio/typogenerator
22. https://github.com/elceef/dnstwist
23. https://github.com/urbanadventurer/urlcrazy
24. Kintis P, Miramirkhani N, Lever C, et al. Hiding in plain sight: A longitudinal study of combosquatting abuse[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017: 569-586.
25. Dam T, Klausner L D, Schrittwieser S. Typosquatting for Fun and Profit: Cross-Country Analysis of Pop-Up Scam[J]. arXiv preprint arXiv:2004.01749, 2020.
26. Dam T, Klausner L D, Buhov D, et al. Large-Scale Analysis of Pop-Up Scam on Typosquatting URLs[C]//Proceedings of the 14th International Conference on Availability, Reliability and Security. 2019: 1-9.
27. Duane Wessels. 2012. Evidence of Bitsquatting in COM/NET Queries. https://

www.nanog.org/meetings/nanog54/presentations/Tuesday/Wessels.pdf. (2012).