When Diversity Meets Hostility: A Study of Domain Squatting Abuse in Online Banking

***Abstract*—In today's digital era, a large number of users rely on banking websites to perform financial transactions. The widespread adoption of online banking and the monetary value associated with each user account make banking websites a potential target for domain squatting. Domain squatting is a common practice in which malicious actors register internet domain names which are similar to popular domains. In this work, we study the prevalence of domain squatting abuse that exploits inconsistent internet domain names used by popular banks across several countries including US, UK, Australia, Germany, China and India. For instance, banks operating within a given country register their domain names in different top-level domains (barclays.co.uk vs. rbsdigital.com) or employ different words in second-level domains (svbconnect.com vs. morganstanleyclientserv.com). An attacker exploits these inconsistencies to generate similar looking domains and use them for malicious purposes such as domain takeover, malware propagation, click fraud, phishing, stealing traffic, distribution of ads and malware.**

In this paper, we present the first context-free grammar (CFG) based algorithm that models inconsistencies in domain names of banking websites and use it to generate candidate domains. We also provide a comprehensive categorization technique to classify candidate domains into four different categories: defensive, malicious, suspicious and unrelated. Our study reveals that more than 3,000 domains that are either malicious or suspicious, targeting popular banks across different countries around the world. Further, we identified three new forms of domain squatting, namely, comboTLDsquatting, fullname squatting and brandname squatting. We found that most of the malicious and suspicious domains are instances of comboTLDsquatting. Our work shows that only few organizations are protecting their brands against domain squatting abuse by performing defensive registration. Further, our study identified different strategies used by malicious actors during domain registration in order to evade detection from security researchers and trick victims into disclosing their credentials. In particular, we discover that malicious actors use similar words, same TLDs, grammar rules and registrar for registering domains which are used in benign domains.

I. Introduction

Online banking has emerged as one of the most important and profitable e-commerce applications in the last decade. It enables customers to perform various activities such as transferring money, balance enquiry and payment of bills at their convenience. Many banks around the world are pursuing a mobile-first strategy,however consumers still prefer and use online banking websites. Recently Deloitte conducted a banking survey with 17100 consumers from 17 different countries [1]. About 73% of the customers use online banking channels at least once a month as compared to 59% who use mobile banking applications. Online banking is used at least once in a month by 94% of mobile banking customers. Also, consumers use mobile banking for relatively simple and quick transactions, such as balance inquiries, but prefer to use online banking to conduct financial transactions. It is predicted that the total number of online and mobile banking users will exceed 3.6 billion by 2024 [2].

Banking sector is one of the most favorite targets for cyber attacks due to the ever increasing adoption of online banking and the monetary nature of the transactions. According to a study conducted by Ponemon institute [3], banking services industry has the highest cost of cybercrime. A report published in Q2 2021 by Anti Phishing Work Group (APWG) identified a total of 616,939 malicious websites of which 29.20% were aimed towards financial institutions [4]. Domain name abuse is one the prominent issues faced by this sector as it leads to a variety of attacks, for example, phishing, drive-by-download, and distribution of ads and malware. In domain name abuse, the attacker would typically register domain names that are confusingly similar to those belonging to popular brands. This practice is commonly known as domain squatting abuse. Domain squatting is also exploited in other abusive activities such as impersonating the original websites to steal traffic, to harvest user credentials and distribution of ads and malware.

Previous work investigated different types of domain squatting including typosquatting (domains that exploit typographical errors) [5], [6], [7], [8], homograph-based squatting (domains that abuse visual similarity of characters from different languages) [9], [10], [11], [12], TLDsquatting (domains that are registered in different top-level domains (TLDs)) and combosquatting (domains that combine a brand name with other words) [13]. Researchers found that combosquatting is 100 times more prevalent than other forms of domain squatting [13]. In this paper, we study and analyze domain squatting abuse in the online banking space. We observed that domain names registered by banking organizations do not follow any common pattern, which could lead to combosquatting, TLDsquatting and other forms of domain squatting abuse.

Table I shows two online banking domain names from each of the five different countries: US, India, China, UK and Canada. As highlighted in the second column of the table, different banks use different words in their second-level domains (SLDs) that too in a different order and are registered in different top-level domains (TLDs). An attacker exploit these inconsistencies to produce new domain names resembling the structure of benign domain names and register them for malicious purposes. This could adversely affect banks, both financially and reputation wise. The generated domains are instances of different domain squatting types including

TABLE I: Inconsistencies in online banking domain names of different countries and candidate domain squatting instances.

Country	Online Banking Domains	Candidate Domain Squatting Instances
US	morganstanley**clientserv**.com, svb**connect**.com	morganstanleyconnect.com, svbclientserv.com, morganstanley.com, svb.com
India	**online**sbi.com, bob**ibanking**.com	onlinebob.com, sbiibanking.com, ibankingsbi.com, bobonline.com, sbionline.com, ibankingbob.com
China	bankofchina.**com**, pingan.**com.cn**	bankofchina.com.cn, pingan.com
UK	yb**online.co.uk**, rbs**digital.com**	ybdigital.co.uk, rbsonline.com, ybonline.com, rbsdigital.co.uk, ybdigital.com, rbsonline.co.uk, yb.co.uk, rbs.com
Canada	royalbank.**com**, nbc.**ca**	royalbank.ca, nbc.com, nationalbankofcanada.com, nationalbankofcanada.ca, rbc.com, rbc.ca

combosquatting and TLDsquatting. The inconsistencies and squatting types are summarized below:

- **Different words in SLDs**: Morgan Stanley bank uses the word *clientserv* in its SLD, whereas Silicon Valley Bank (SVB) uses the word *connect*. The attacker can exchange these two words to produce two **combosquatting** domains, namely morganstanleyconnect.com and svbclientserv.com, and register them (if available) for malicious purposes.
- **Different words in different order**: State Bank of India (SBI) uses the word *online* before the brand name *sbi* in its SLD whereas Bank of Baroda (BOB) uses a different word *ibanking* after the brand name *bob*. In this case, the attacker can exchange the words as well as their positions to obtain six **combosquatting** domains, namely onlinebob.com, sbiibanking.com, ibankingsbi.com, bobonline.com, sbionline.com and ibankingbob.com.
- **Different TLDs**: Bank of China and Pingan bank are registered in two different TLDs, *com* and *com.cn*. The attacker can generate two **TLDsquatting** domains, bankofchina.com.cn and pingan.com, and register them (if available) for malicious purposes.
- **Different words and TLDs**: Yorkshire Bank (YB) and Royal Bank of Scotland (RBS) use different TLDs as well as different words in their SLDs. The attacker can generate six potential candidates by exchanging words and TLDs, ybdigital.co.uk, rbsonline.com, ybonline.com, rbsdigital.co.uk, ybdigital.com and rbsonline.co.uk. The last two instances are combination of both combosquatting and TLDsquatting (as words and TLDs are both different). We refer to this new form of domain squatting as **comboTLDsquatting**.
- **Difference in usage of full name**: Royal Bank of Canada (RBC) uses organization name in its online banking domain (royalbank.ca), whereas National Bank of Canada (NBC) uses acronym (nbc.com). The attacker can register the full name nationalbankofcanada.com if applicable for malicious purposes. We refer to this new form of domain squatting as **fullname squatting**.
- **Difference in usage of brand name**: Most banks use a word alongside their brand name in SLDs. For example, morganstanleyclientserv.com, onlinesbi.com and rbsdigital.com. The attacker can register just the brand names (morganstanley.com, sbi.com and rbs.com) if available and use them for malicious purposes. We refer to this new form of domain squatting as **brandname squatting**.

We use the aforementioned inconsistencies in the benign domain names for generating a context-free-grammar (CFG). The CFG is then used for generating potential candidates for combosquatting, TLDsquatting, comboTLDsquatting, fullname squatting and brandname squatting domains. We collect data pertaining to each of the generated domains, analyze it, and provide new results and insights in the domain squatting landscape of banking institutions. Specifically, our contributions are as follows:

- We design a novel algorithm that exploits inconsistencies present in benign domains and learns a CFG. This is the first study which uses CFG for domain name generation. The resulting grammar then generates new domains which resemble the benign domains. For data generation, we consider 307 online banking domains from 13 major countries. We found that 4,113 candidate domains generated using our CFG based algorithm were already registered.
- We define and identify three new forms of domain squatting, namely comboTLDsquatting, fullname squatting and brandname squatting. We present a more comprehensive categorization technique that employs WHOIS records, DNS records, HTTP status codes and web page content. Of the 4,113 registered domains, we found that only 606 domains (14.73%) are defensive, whereas 3,140 domains (76.34%) are either malicious or suspicious. Thus, only few organizations are protecting their brands against domain squatting abuse by performing defensive registration.
- We uncovered popular words, TLDs, registrars and grammar rules used during benign and malicious domain registration across different countries around the world. We also identified different strategies used by malicious actors during domain registration in order to evade detection from security researchers and trick victims into disclosing their credentials. In particular we discover that, malicious actors use similar words, same TLDs, grammar rules, and registrar as benign domains to register confusingly similar domains.

II. Background

Following section provides a brief overview of the URL structure. The details pertaining to domain name registration

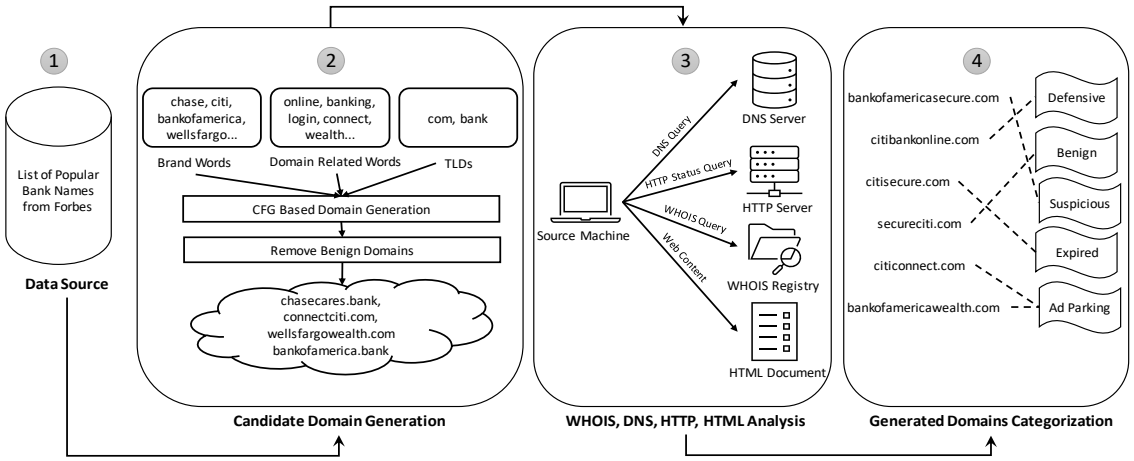


Fig. 1: Our approach to generate and categorize domains related to online banking.

and domain squatting techniques are provided in Appendix A and B respectively.

Components of URL: URL is an acronym for Uniform Resource Locator. Its main purpose is to identify the location of documents and other resources available on the World Wide Web. A URL is comprised of the following three components:

1) **Scheme:** The scheme identifies the protocol to be used to access the resource on the Internet. The scheme name is followed by :// (a colon and two slashes).
2) **Hostname:** The hostname identifies the machine connected to the web that contains the requested resource. The hostname is further sub-divided into two parts: subdomain and domain. Domain is composed of top-level domain (TLD) and second-level domain (SLD).
3) **Path:** The path identifies the location of the requested resource on the host machine. The path is sub-divided into three components: directory, file name and arguments.

Consider the URL **"https://netbanking.bankofamerica.com /online-banking/sign-in?request_locale=en_US"**.Its components are:

1) Scheme: https
2) Hostname: netbanking.bankofamerica.com
   a) Subdomain: netbanking
   b) Domain: bankofamerica.com
   c) SLD: bankofamerica
   d) TLD: com
3) Path: online-banking/sign-in?request_locale=en_US
   a) directory: online-banking
   b) filename: sign-in
   c) arguments: request_locale=en_US

III. Approach

As illustrated in Figure 1, our approach is composed of four steps. The first step consists of collecting a set of benign domains from different countries, the second step is generating candidate domains using a CFG based algorithm, the third step is crawling WHOIS, DNS and web page data pertaining to the generated domains, and the fourth and final step is to use the crawled data to categorize the generated domains into different categories such as defensive, malicious and suspicious.

A. Data Source

To study the prevalence of domain squatting abuse in banking sector, we collected domains registered by **307 popular banks** from **13 major countries**. We refer to these domains as *benign domains*. To collect popular bank names within each country, we used Forbes list of world's best banks published in 2020 [14]. The Forbes list is compiled based on key attributes like digital services, trust, fees, financial advice and general satisfaction. To get online banking domain corresponding to each bank in the list, we searched the bank name along with the phrase *net banking* on Google Search and checked if the search result retrieves online banking domain for the queried bank. If we did not get the domain, we tried the phrases *online banking* and *internet banking*, and repeated the search process. The country-wise count of banks considered in our study is given in Table II.

B. Data Generation

Prior work [13] highlights the need for a generative model to proactively generate combosquatting domains. We designed a context-free grammar (CFG) based domain generation algorithm which exploits inconsistencies in the benign domains and generates combosquatting, TLDsquatting, comboTLD-squatting, fullname squatting and brandname squatting domains. CFGs have been used extensively in the study of natural languages [15], [16], [17], where they are used to generate strings with particular structures. In security domain, CFGs is successfully used in automatic generation of passwords that resemble human-created passwords [18], [19], [20]. We show that CFGs are also useful in generation of domain names that resemble benign domains. A CFG is a tuple of (V, Σ, P, S) where,

- V is a set of non-terminal symbols. Non-terminals are typically represented by capital letters or Greek letters.
- Σ is a set of terminals. Terminals are represented by lowercase letters.

TABLE II: The number of online banking domains for each country along with examples of words used in SLDs and TLDs. The table also shows the number of domains generated using CFG based algorithm for each country.

S.No.	Country	# Benign Domains	Examples of Words Used in SLDs	Examples of TLDs	# Generated Domains
1	US	73	clientserv, online, connect	com	3,312
2	Germany	42	login, online, banking	de, com	2,451
3	India	31	online, net, ibanking	co.in, in, net, net.in, co	5,160
4	Italy	26	onlinebanking, login, securelogin	com, net, it	2,780
5	Australia	21	digital, ibanking, internetbanking	com.au, com	1,326
6	UK	20	secure, personal, login	co.uk, com	916
7	China	17	ibanking, ebank, myebank	cn, com.cn, com	1,101
8	Brazil	15	webbanking, internetbanking, login	com.br, com	367
9	France	14	secure, ebanking, connexion	fr, com	618
10	Canada	13	personal, online, personal-banking	ca, com	594
11	Poland	13	secure, online, login	pl, com.pl	406
12	Russia	13	online, personal, click	ru, com	492
13	Israel	9	portal, login, bank	il, co.il	93
	Total	**307**			**19,616**

- P is a set of production rules. Each rule is of the form $\alpha \rightarrow (\Sigma \cup V)^*$, where $\alpha \in V$.
- $S \in V$ is the start symbol.

Once the CFG is available, we can leverage it to generate new strings by writing down the start symbol S and repeatedly replacing non-terminals V according to the production rules P of the proposed grammar until no non-terminals are left. The language L of the CFG is the set of strings composed only of terminals Σ derivable from the start symbol S, that is, $L = \{w \in \Sigma^* \mid G\ generates\ w\ starting\ from\ S\}$.

We process a list of benign domain names and create a separate CFG for each country. We created separate CFG for each country because there is a difference in the intrinsic structure of URLs of each country. The resulting grammar consists of six non-terminal symbols, S represents start symbol, C represents SLDs, T represents TLDs, F represents organization's full name, B represents brand name used in SLDs and W represents words that are used along with brand names. The production rules for non-terminals S and C are pre-determined and fixed. Specifically, the start symbol S can be substituted in three different ways, $C.T$, $B.T$ and $F.T$, and the non-terminal C can be substituted in two different ways, BW and WB. The production rules of C consider only single word, since most of the combosquatting domains are constructed by adding a single token (word) to the original brand name [13]. The production rules for the remaining non-terminals T, B, W and F are learned from the benign domain name data. The rule $C.T$ generates both combosquatting and comboTLDsquatting domains, whereas the rule $B.T$ generates brandname squatting domains, while the rule $F.T$ generates fullname squatting domains. TLDsquatting domains are generated by all the rules.

For example, consider two Indian banks, State Bank of India (SBI) and Punjab National Bank (PNB), and their respective online banking domain names are onlinesbi.com and pnbibanking.in. The resulting grammar G_1 obtained by processing these two domain names is shown below.

- $S \rightarrow C.T \mid B.T \mid F.T$
- $C \rightarrow BW \mid WB$
- $T \rightarrow com \mid in$
- $B \rightarrow sbi \mid pnb$
- $W \rightarrow online \mid ibanking$
- $F \rightarrow statebankofindia \mid punjabnationalbank$

The aforementioned CFG models inconsistencies in TLDs and SLDs of two banks, SBI and PNB. SBI uses the word $online$ while PNB uses word $ibanking$ in its SLD. SBI is registered in the TLD com, whereas PNB is registered in a different TLD in. Further, SBI's SLD contains the word followed by the brand name ($onlinesbi$) whereas PNB's SLD contains the brand name followed by the word ($pnbibanking$). This is modeled by production rules of non-terminal $C \rightarrow BW \mid WB$. The terminals in CFG are words appearing in the right-hand side of the rules associated with non-terminals T, B, W and F.

The pseudocode for learning a CFG from the set of benign domains is given in Algorithm 1. The $Main$ procedure relies on two subprocedures, $LearnCFG$ and $GenerateDomains$ (lines 1-9). The input to $LearnCFG$ procedure is a set of tuple of benign domain name d (onlinesbi.com), organization name f (State Bank of India) and acronym a (SBI) of the organization name (if it exists). The output of the procedure is a CFG that models inconsistencies in the benign domains. The set of non-terminals V is fixed (line 14). Initially, the set of terminals is empty (line 15). All grammar productions are stored in dictionary P with a non-terminal as key and a set of substitution rules as value. The production rules for the start symbol S and the variable C are predefined (lines 18-19). To learn production rules for the remaining four non-terminals, we process each tuple (d, f, a) (lines 24-33). We split every domain d (onlinesbi.com) into two parts, SLD s (onlinesbi) and TLD t (com). We use strings f (State Bank of India) and a (SBI) to further separate SLD s (onlinesbi) into two parts, brand b (sbi) and word w (online). We add t as a possible substitution for T, b as a possible substitution for B, w as a possible substitution for W and f as a possible substitution for F (lines 28-31). Further, strings t, b, w and f are added to the set of terminals (line 32). After iterating over all tuples, we have our grammar G which is returned to the calling procedure (line 34). We note that the resulting grammar is non-recursive. Further, the resulting grammar could be ambiguous since the brand name b extracted from SLD and fullname f of the corresponding organization can coincide, i.e., $b = f$.

Algorithm 1 CFG Based Domain Generation Algorithm.

1: **procedure** $Main$
2: **Input:** A set $I = \{(d_1, f_1, a_1), (d_2, f_2, a_2), \ldots, (d_n, f_n, a_n)\}$ of 3-tuples.
3: **Output:** A set L of potential combosquatting, TLDsquatting, comboTLDsquatting, fullname squatting and brandname squatting domains generated by exploiting inconsistencies in the benign domain names in I.
4: $G = LearnCFG(I)$ //G contains CFG
5: $S = Start(G)$ //S is the start variable in G
6: $L = GenerateDomains(S, G)$ //S is the start variable of grammar G
7: $D = I[1:n, 1]$ //extract domains $\{d_1, d_2, \ldots, d_n\}$ from I
8: $L = L/D$ //remove benign domains from the set L of generated domains
9: **end procedure**
10:
11: **procedure** $LearnCFG$
12: **Input:** A set $I = \{(d_1, f_1, a_1), (d_2, f_2, a_2), \ldots, (d_n, f_n, a_n)\}$ of 3-tuples.
13: **Output:** A grammar $G = (V, \Sigma, P, S)$ that models inconsistencies in the benign domains in I.
14: $V = Set([S, C, B, T, W, F])$
15: $\Sigma = Set()$
16: /*Grammar rules are stored in a dictionary with variable name as a key and a set of productions as value*/
17: $P = HashMap()$
18: $P[S] = Set([C.T, B.T, F.T])$ //$S \rightarrow C.T \mid B.T \mid F.T$
19: $P[C] = Set([BW, WB])$ //$C \rightarrow BW \mid WB$
20: $P[B] = Set()$
21: $P[W] = Set()$
22: $P[T] = Set()$
23: $P[F] = Set()$
24: **for** $(d, f, a) \in I$ **do**
25: /* Split domain d into three parts: brand name b, word w, and TLD t */
26: $s, t = processDomain(d)$
27: $b, w = processSLD(s, f, a)$
28: $P[B].add(b)$ //$B \rightarrow b$
29: $P[W].add(w)$ //$W \rightarrow w$
30: $P[T].add(t)$ //$T \rightarrow t$
31: $P[F].add(f)$ //$F \rightarrow f$
32: $\Sigma.add([b, t, w, f])$ //add terminals
33: **end for**
34: return $G = (V, \Sigma, P, S)$
35: **end procedure**
36:
37: **procedure** $GenerateDomains$
38: **Input:** A non-terminal $\alpha \in V$ and a CFG $G = (V, \Sigma, P, S)$
39: **Output:** The set L of strings generated starting from the variable α
40: **if** $\alpha \notin V$ **then**
41: return $Set()$
42: **end if**
43: $L = Set()$
44: **for** $R \in P[\alpha]$ **do**
45: $A = Set()$
46: **for** $\beta \in R$ **do**
47: **if** $\beta \in V$ **then**
48: $L_v = GenerateDomains(\beta, P)$
49: $A = concatenate(A, L_v)$
50: **else if** $\beta \in \Sigma$ **then**
51: $A = concatenate(A, Set([\beta]))$
52: **else** $exit()$
53: **end if**
54: **end for**
55: $L = L \cup A$
56: **end for**
57: return L
58: **end procedure**

Hence, the same candidate domain can be generated using two different rules, $S \rightarrow B.T$ and $S \rightarrow F.T$.

Once we obtain the grammar G (line 4), we use $GenerateDomains$ procedure to generate candidate combosquatting, TLDsquatting, comboTLDsquatting, brandname squatting and fullname squatting domains. The procedure takes as its input a non-terminal α and grammar G. It returns all strings that can be derived starting from the non-terminal α using production rules of G. The procedure is called with G learned from a set of benign domains and setting $\alpha = S$ (lines 5-6). Hence, we get the set of all strings generated using G. The procedure begins by checking whether the symbol α is indeed a non-terminal in G. If it is not, then it returns an empty set (lines 40-42). If the symbol α is a non-terminal, then we iterate through all production rules associated with α. For each production rule $\alpha \rightarrow R$, the inner for loop produces the set of strings derivable from R and stores it in set A (lines 46-54). Initially the set A is empty (line 45). The inner for loop scans the string R from left-to-right and checks for a non-terminal symbol. If it encounters a non-terminal symbol β then the procedure $GenerateDomains$ is called again which recursively finds all possible strings generated by the non-terminal β (line 48). Subsequently, all strings derived from the non-terminal β are concatenated with strings stored in set A (line 49). The procedure $Concatenate$ is a utility function which concatenates every string with every other string from two sets S_1 and S_2, i.e., S = {xy $|x \in$S$_1$ and y $\in$S$_2$}. At the end of the inner for loop, all possible strings generated from the rule R are available in set A which is subsequently combined with set L (line 55). In this way, all possible strings derivable from the start symbol S (and hence the grammar G) are produced and stored in L. Finally, the set L is returned (line 57). The grammar also generates benign domains which are removed by the $Main$ procedure (lines 7-8).

The outer for loop of $GenerateDomains$ procedure iterates over all possible production rules associated with non-terminal α, which are finite. The inner for loop parses a given production rule R which is of finite length. Further, as the grammar is non-recursive, the recursion depth of the $GenerateDomains$ procedure is finite as well. Specifically, strings generated using the rule $S \rightarrow C.T$ have depth three, strings generated using the rules $S \rightarrow B.T$ and $S \rightarrow F.T$ have depth two. Hence, the algorithm $GenerateDomains$ eventually halts and returns the set L of all possible strings generated by the grammar G. Let $|\alpha|$ represent the number of terminal strings derived from the non-terminal $\alpha \in V$. As the resulting grammar could be ambiguous, the total number of domains n that can be generated beginning from the start symbol S is at most $|S|$.

$$n \leq |S| \leq 2 \cdot (|W| + 1) \cdot |I| \cdot |T| \quad (1)$$

The derivation for equation (1) is given in Appendix C, where we use the fact that both the number of brand names and organization names are equal to the number of tuples in the input set I, that is, $|B| = |F| = |I|$, however, the set of generated domains L also contains benign domains from I. We remove the benign domains from L. For the CFG G_1 derived from the input set $I = \{(onlinesbi.com, StateBankofIndia, SBI), (pnbibanking.in, PunjabNationalBank, PNB)\}$, we have $|W| = |B| = |F| = |T| = |I| = 2$. The number of generated domains is 24 out of which 2 are benign. Newly generated 22 domains along with their squatting types are given below:

- **Combosquatting** : sbionline.com, sbiibanking.com, ibankingsbi.com, pnbonline.in, onlinepnb.in, ibankingpnb.in
- **TLDsquatting** : onlinesbi.in, pnbibanking.com
- **ComboTLDsquatting** : sbionline.in, sbiibanking.in, ibankingsbi.in, pnbonline.com, onlinepnb.com, ibanking-

pnb.com

- **Fullname squatting** : statebankofindia.com, statebankofindia.in, punjabnationalbank.com, punjabnationalbank.in
- **Brandname squatting** : sbi.com, sbi.in, pnb.com, pnb.in

Figure 10 (Appendix D) shows parse tree derivation of comboTLDsquatting domain sbionline.in. The advantage of the CFG based approach is that it allows incremental training. If we add some new benign domains to the original set, then we can still reuse the existing CFG and simply extend it for additional domain generation.

C. Data Crawling

After generating candidate domains for each country, we gathered WHOIS record, DNS record and web page (if it exists) for each of the generated domain. A number of studies [7], [8], [11], [13] have leveraged WHOIS record, DNS record and web page information for categorization. To obtain this information, we set up three automated crawlers which are explained below.

- **WHOIS Lookup:** The first crawler was configured to perform WHOIS lookup for benign and candidate domains. We used **python-whois** library which supports extraction of WHOIS data for the TLDs considered in our experiment. Out of 19,616 candidate domains, we obtained WHOIS data for 4,113 domains. Out of 4,113 domains, WHOIS records of 224 domains were redacted for privacy. The redacted WHOIS record still provides information about domain creation date, updation date and expiration date, and details like city, state and country of the registrant, which can be useful. The crawler saves the entire WHOIS record to the disk.
- **DNS Lookup:** For each candidate domain with WHOIS record, the second crawler determined whether the domain resolves to an IP address. We used **dnspython** library to perform DNS lookup. The crawler saves the IP address (if it exists) to the disk. A total of 2,492 domains resolved to an IP address during our study.
- **Web Page Crawling:** For each candidate domain with DNS record, the third crawler visited the web page hosted on the domain using Selenium, a headless JavaScript-enabled web browser. After loading the web page, the crawler waits for 10 seconds, allowing the page to load dynamic content or perform redirection. Finally, the crawler saves the final URL, HTML body, HTTP status code and a screenshot of the page to the disk.

D. Categorization of Candidate Domains

Based on the analysis of WHOIS records, DNS records and web page content, we classify the candidate domains into four different categories, namely *defensive*, *suspicious*, *malicious* and *unrelated*. These categories are further subdivided as shown in Figure 11 (Appendix E). Such categorization technique can be found in prior work [8], [11], [13], [21], [22]. Building upon it, we present a more comprehensive categorization technique which is given below:

1) **Defensive:** First, we compare the WHOIS record of the candidate domain and the benign domain. If it matches then we call it as defensive. For the redacted WHOIS record, if the domain redirects to benign website we consider it as defensive. Such domains are proactively registered by an authoritative domain owner to curb abuse from domain squatters. Defensive category is subcategorized into six types given below:
   a) **Expired:** If the expiry date is less than the current date then the domain is classified as *expired*. As expired domains are potentially dangerous [23], [24], [25], [26], banks should be vigilant of these domains and proactively register them before attackers do.
   b) **Not Live:** We check if the domain resolves to some IP address (via DNS lookup). If it does not, then the domain is defensively registered but not being used.
   c) **Redirection to benign:** The domain resolves to an IP address and redirects to the benign website [8].
   d) **Server throws error:** The website page shows an error (e.g., HTTP status code 404).
   e) **Coinciding:** The domain hosts the same web page content as that of the benign domain [8].
   f) **Content does not match:** The web page has different content from that of the benign web page or it is blank.

   Decision making flowchart for Defensive subcategories is given in Figure 2.

   Suspicious, Malicious and Unrelated categories are considered when the organization name and address fields in the WHOIS record of the candidate domain do not match with that of the benign domain.
2) **Suspicious:** The domain doesn't resolve to an IP address OR if it resolves to an IP address but the page hosted by the website is blank or displays an error message [13].
3) **Malicious:** If a valid page exists, then we analyze it using an image hashing based technique [8] described in Appendix F and determine whether it displays any fraudulent content. Malicious category is further subdivided into six subcategories as follows:
   a) **Expired:** If the expiry date is less than the current date then the domain is classified as *expired* and it is available for registration [23], [25].
   b) **Phishing:** The page poses as a reputed brand that deceives users to enter their personal sensitive information like username, password or PIN [13], [27].
   c) **Social Engineering:** The page displays surveys, scams and malicious downloads that trick users to give away their information [13], [21], [8].
   d) **Ad Parking:** The page displays advertisement related links provided by commercial parking vendors such as parkingcrew, sedoparking and parklogic. Recent work [8], [28], [23], [29], [30] consider ad parking pages as malicious because they are involved in malware propagation, click fraud, malvertising practices, traffic spam, fake antivirus warnings, traffic stealing, hosting malicious content, and are vulnerable to AWS hijack-

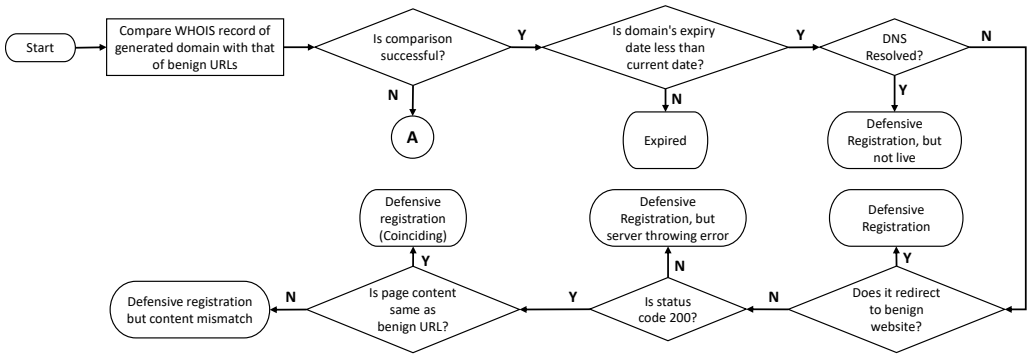


Fig. 2: Flowchart for Defensive subcategories.

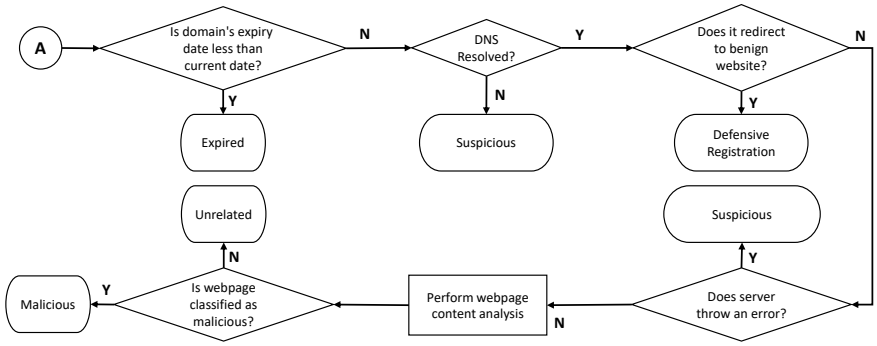


Fig. 3: Flowchart for Suspicious, Defensive, Malicious and Unrelated categories.

ing, domain takeover, etc.

e) **Domain for sale:** The generated domain is put up for sale on an auction website. Such domains, when visited, redirect users to unwanted pages, and are hence considered as malicious [8], [22], [31].

f) **Adult Content:** The page displays adult content [8].

4) **Unrelated:** If the page does not contain any malicious content then we call it unrelated (seemingly benign and unrelated websites) [13].

Decision making flowchart for Suspicious, Defensive, Malicious and Unrelated categories is given in Figure 3. Note that letter "A" in Figure 3 indicates that WHOIS record of generated domain doesn't match with that of benign URLs.

IV. Results

A total of 19,616 candidate domains were generated using 307 online banking domains from 13 major countries, out of which we found that 4,113 domains were registered. For analysis, we focus only on the registered domains. Table III shows country-wise distribution of defensive, unrelated, malicious and suspicious domains. As the number of banks considered in each country is different (e.g., 73 in US while 13 in Poland) we normalized the count of domains in all categories by dividing it by the number of banks considered in the respective country. This allows fair comparison across countries. For example, the actual number of defensive registrations in US is 187 (Actual Count: AC), after normalization (dividing by the number of US banks 73), the number of defensive registrations per bank is 2.56 (Normalized Count : NC). As the number of registered domains in Poland and Israel are 21 and 10 respectively, (too small to draw insightful observations), we skip them from country-wise analysis.

A. Overall Category Distribution

Out of 4,113 registered domains, using our flowchart, we determined that 606 (14.73%) domains are defensively registered, 1,255 (30.51%) are malicious, 1,885 (45.83%) are suspicious and 367 (8.92%) domains have unrelated pages. Figure 4 shows the distribution of subcategories within each main category.

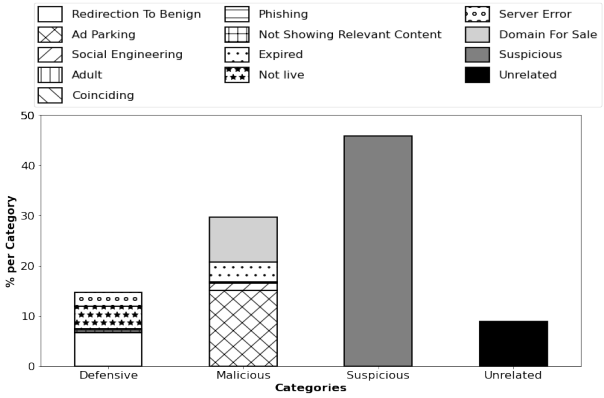


Fig. 4: Overall distribution of all categories.

TABLE III: Category-wise distribution of registered domains for each country. For each category, we report actual count (AC) as well as normalized count (NC), where NC = AC/(no. of banks considered for generating domains).

Country	No. of Banks	Defensive (AC/NC)	Malicious (AC/NC)	Suspicious (AC/NC)	Unrelated (AC/NC)	Total (AC/NC)
US	73	187/2.56	301/4.12	190/2.60	86/1.18	**764/10.46**
Germany	42	45/1.07	59/1.40	65/1.55	30/0.71	**199/4.74**
India	31	80/2.58	**407/13.12**	**246/7.94**	97/3.13	**830/26.78**
Italy	26	**89/3.42**	**120/4.62**	**899/34.58**	26/1	**1134/43.61**
Australia	21	26/1.24	47/2.24	68/3.24	18/0.86	**159/7.6**
UK	20	34/1.7	**98/4.9**	111/5.55	30/1.5	**273/13.65**
China	17	1/0.06	56/3.29	**100/5.88**	21/1.23	**178/10.5**
Brazil	15	**40/2.67**	37/2.47	45/3	12/0.8	**134/8.93**
France	14	21/1.5	17/1.21	53/3.79	17/1.21	**108/7.71**
Canada	13	**60/4.62**	46/3.54	34/2.62	10/0.77	**150/11.54**
Poland	13	8/0.61	6/0.46	7/0.54	0/0	**21/1.61**
Russia	13	13/1	**60/4.62**	60/4.62	20/1.54	**153/11.76**
Israel	9	2/0.22	1/0.11	7/0.78	0/0	**10/1.11**
Total	**307**	**606/1.97**	**1255/4.09**	**1885/6.14**	**367/1.19**	**4113/13.40**

Out of 606 defensively registered domains, 183 domains failed to resolve to an IP address (not live), 277 domains redirected to their primary benign websites (defensive), 116 domains showed error pages, 13 domains displayed web pages having different content from that of the respective benign web page, 5 domains displayed pages that matched with the content on the respective primary domains (coinciding) and 12 domain registrations are expired. Thus, only 290 domains (277 defensive and 13 coinciding) showed correct behavior while the rest of the defensive registrations did not redirect to the primary benign website. Further, the expired defensive domains can now be registered for malicious purposes.

Out of 1,255 malicious domains, 655 domains are parked, 369 domains are put up for sale, 164 domains are expired, 58 domains are involved in social engineering, 7 domains displayed adult content and 2 domains are involved in phishing attack. Overall, we found that India alone contributes to around 32.43% of the malicious registrations whereas Italy contributes to 47.70% of the suspicious registrations.

Ad parking domains comprise 52.19% of malicious domains. Few examples of parked domains are mentioned below:

- US: bankofamericawealth.com, chaselogin.com, citibanking.com
- China: icbcbank.com, cmbcbank.com.cn, bankboc.cn
- India: hdfcnetbank.in, sbibank.com, icicinetbanking.com
- UK: hsbconlinebanking.co.uk, lloydsonline.co.uk, barclayslogin.com

Domains for sale comprise 29.40% of malicious domains. Few examples of domains that are put up for sale are given below:

- US: citiwealth.com, chaseaccounts.com, onlineciti.com
- China: icbcchina.com, bocchina.com, abcchina.cn
- India: hdfcindia.in, onlineicici.com, punjabnationalbank.net
- UK: lloydsdigital.com, sconlinebanking.com

Few examples of suspicious domains are given below:

- US: bankofamericasecure.com, chaseonline.com, citiclient.com
- China: abbank.com.cn, chinaboc.cn, cnbcchina.com
- India: hdfcindia.com, sbinetbanking.co.in, pnbnet.co.in
- UK: hsbcpersonal.com, barcalysonline.co.uk, rbsonlinebanking.co.uk

As shown in Table III, the top three countries having most defensive registrations are Canada, Italy and Brazil with 4.62, 3.42 and 2.67 NC respectively. For malicious category, India, UK, Italy and Russia are at the top positions with 13.12, 4.9, 4.62 and 4.62 NC respectively. For suspicious category, Italy, India and China are the top three countries with 34.58, 7.94 and 5.88 NC respectively.

B. Defensive Subcategories

We found that among different defensive subcategories, redirection to benign (45.70%), not live (30.20%) and server throwing error (19.14%), together contribute more than 95% of the defensive registrations. Figure 5 shows country-wise plots for these three subcategories.

Redirection to benign: As depicted in Figure 5a, Canada, Brazil and Italy are the top 3 countries with 2.15, 1.60 and 1.35 NC respectively that redirect to the primary domains. On the contrary, China and Russia have the least number of registrations (0.05 and 0.38 NC respectively) that redirect to the benign domains. We found that Canada's Desjardins Group with 6 domains, Brazil's Banco Bradesco with 14 domains and Italy's Banco Posta with 5 domains are the top banks that redirect to primary website (benign domain).

Not Live: As depicted in Figure 5b, Canada, Italy and US have at least one defensively registered domain per bank that doesn't resolve to an IP address. Thus, banks have registered the domains defensively but they are not using it. We observed that US's Bank of America has 12 such domains, Italy's Banca Generali Private has 10 such domains and Canada's Canadian Imperial Bank of Commerce has 8 such domains.

Server throwing error: As depicted in Figure 5c, India with 1.22, Italy with 0.73 and US with 0.50 NC are the top 3 countries that display error pages on the defensively registered domains. We found that 5 domains of US's Capital One bank, 7 domains of India's Punjab National Bank and 5 domains of Italy's Credito Emiliano bank show an error page.

Expired: We found a total of 12 domains, 5 from Italy, 4 from US, 2 from India and 1 from Brazil, that were once defensively registered but are now expired (available for registration). The five expired domains from Italy are unicreditonline.com, unicreditonline.it, bperbanca.it, credembanca.it and creditoemil-

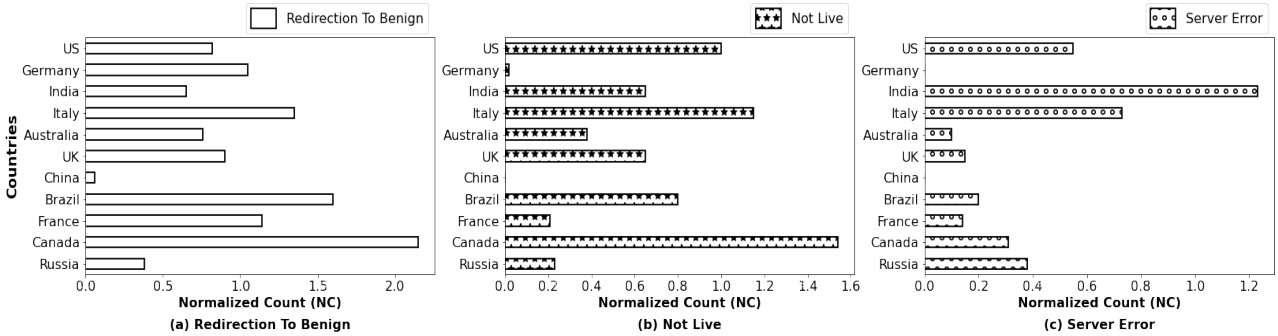


Fig. 5: Country-wise distribution of defensive subcategories.

iano.it, and they belong to 3 different banks, namely Unicredit Bank, BPER Banca and Credito Emiliano. The four expired domains from US are citibankingonline.com, citisecure.com, morganstanleyonline.com and svbwealth.com, and they belong to 3 different banks, namely Citi Bank, Morgan Stanley and Silicon Valley Bank. The two expired domains from India are pnbibanking.in and pnb.co.in, both belonged to Punjab National Bank. The expired domain santanderbrasil.com.br from Brazil belonged to Banco Santander Brasil Bank.

Content does not match: A total of 13 defensively registered domains displayed web pages having different content from that of the respective benign web page. Of these 13 domains, 7 are from US and 6 are from Canada. The domains from US are citionline.com, citilogin.com, morganstanleylogin.com, onlinemorganstanley.com, pncconnect.com, pncdirect.com and nbtbancorp.com, and they belong to four different banks, namely Citi Bank, Morgan Stanley, PNC Financial Services and NBT Bank. The domains from Canada are banksimplii.com, banksimplii.ca, simpliifinancial.ca, laurentianbankofcanada.ca, tdwealth.com and tdbanque.com, and they belong to three different banks, namely Simplii Financial, Laurentian Bank of Canada and Toronto-Dominion Bank.

Coinciding: Only 5 domains were found in coinciding subcategory, 2 from US and 3 from Canada. These are midfirstbank.com, metafinancialgroup.com, scotiabank.ca, banquescotia.com and banquescotia.ca and they belong to 3 different banks, namely Midfirst Bank, MetaBank and Bank of Scotia.

C. Malicious Subcategories

We found that among different malicious subcategories, parked domains (52.20%), domain for sale (29.40%) and expired domains (13.06%), together contribute to more than 94% of the malicious registrations. Figure 6 shows country-wise plots for these three subcategories.

Ad Parking: As given in Figure 6a, the top 3 countries in ad parking subcategory are India, US and UK with NC of 7.58, 2.65 and 2.60 parked domains respectively. For instance, we found that 22 parked domains were derived from India's Allahabad bank, 8 were derived from US's JPMorgan Chase bank, and 9 parked domains were derived from UK's Standard Chartered bank.

Domain for Sale: As shown in Figure 6b, the most popular countries in domain for sale subcategory are India, China and Russia with NC of 3.38, 2 and 1.85 respectively. For instance, we found that 13 domains were derived from India's Federal Bank, 6 domains from China's Bank Of Jiangsu Co and 5 domains from Russia's Ros bank were up for sale.

Expired: As depicted in Figure 6c, the highest number of expired domains were found for Italian banks with NC of 2.38. The next two countries are India and Russia with 1.51 and 0.615 NC respectively. For instance, we found that 14 domains derived from Italy's Credito Emiliano, 5 from India's Allahabad bank and 2 from US's City National Bank are expired and available for registration.

Social Engineering: Around 4.6% malicious domains performed social engineering attacks. Most of these domains were derived from India, France and Russia. We noticed that 6 domains derived from India's Allahabad bank and Yes bank, 5 from France's Milleis Banque SA and 3 from Russia's Promsvyaz bank were involved in social engineering attacks.

Adult Content: We found 7 domains with adult content, out of which 3 domains were derived from China, 2 domains from India and 2 domains from US. These are denain.com, bankfederal.com, unitedbanking.com, client53.com, bocebs.com, cebchina.com and pbankpsbc.com.

Phishing: We found two domains hosting phishing content (paytmindia.net and iniob.com) derived from two Indian banks (Paytm Payments Bank and Indian Overseas Bank).

D. Grammar Rules

Our CFG based algorithm employed four productions rules for generating candidate domains, namely $S \rightarrow BW.T \mid WB.T \mid B.T \mid F.T$. Out of 4,113 registered domains, 52.68% were generated from $BW.T$, 30.93% from $WB.T$, 10.77% from $B.T$ and 5.62% from $F.T$. Figure 7 shows the rule-wise distribution of each of the four categories: defensive, malicious, suspicious and unrelated. Only 36.36% of the domains generated using the rule $F.T$ and 18.48% of the domains generated using the rule $B.T$ are defensively registered. Thus, majority of organizations do not bother to register either their full organization names or brand name which are taken by malicious entities. This is the root cause of fullname squatting and brandname squatting. We note that the rule $B.T$ has the highest percentage of unrelated domains

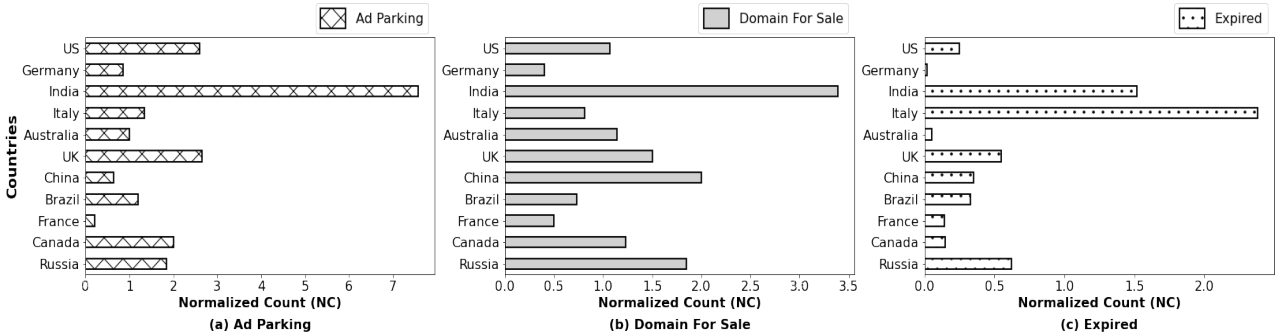


Fig. 6: Country-wise distribution of malicious subcategories.

(23.26%). This is explained by an empirical observation that multiple companies, including those which are not related to banking have same acronyms as banking organizations and use them as brand names. Examples of malicious fullname squatting domains are given below:

- US: firstcommonwealthbank.com, carterbanktrust.com
- UK: yorkshirebank.com, clydesdalebank.com
- China: bankofbeijing.cn, chinaminshengbank.com
- India: statebankofindia.co.in, punjabnationalbank.com

56.21% of the domains produced by the rule $WB.T$ are suspicious, 28.77% are malicious while only 7.86% of the domains are registered defensively. Similarly, 42.96% of the domains produced by the rule $BW.T$ are suspicious, 34.10% are malicious and only 15.78% are defensive. The domains produced by the rules $WB.T$ and $BW.T$ are combosquatting, TLDsquatting and comboTLDsquatting domains. Out of candidate domains produced by the rule $BW.T$, 58.51% are combosquatting, 32.25% are comboTLDsquatting, and 9.22% are TLDsquatting. The rule $WB.T$ also produced three types of squatting domains, namely, combosquatting (63.36%), comboTLDsquatting (33.72%) and TLDsquatting (2.9%).

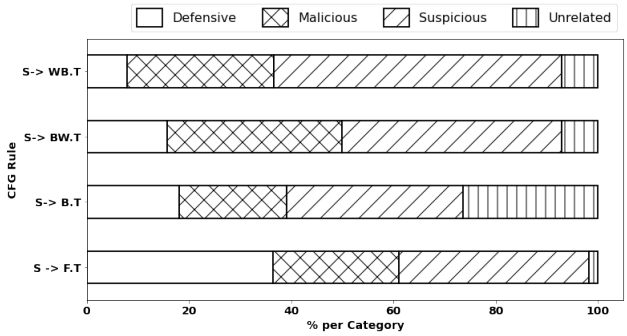


Fig. 7: Overall category distribution in CFG rules.

E. Popular Words, TLDs and Registrars

Table IV shows top 3 words along with top TLDs used in malicious and suspicious domain registrations. The words *online* and *bank* are very popular. The word *bank* appears in 10 out of 11 countries (except US), whereas the word *online* is the most popular word in 6 out of 11 countries. We also found that country specific words such as *india* and *china* are being used to register malicious and suspicious domains within those specific countries. Thus, the choice of words by adversaries show that most registered domains are intentionally designed to confuse users. The generic TLD *com* is the most popular TLD for registering malicious and suspicious domains in all countries except Italy. Moreover, the second popular choice is the country specific TLDs, i.e., ccTLDs. If benign banking domains are registered in gTLDs (*com* and *net*), then the attacker tries to register them in ccTLDs and vice versa.

Top 2 registrars in both defensive and malicious registrations for each country are provided Table V. We found that overall CSC Corporate Domains is the most popular registrar in defensive category whereas GoDaddy is the most popular registrar in malicious category. CSC Corporate Domains registrar is used in 6 out of 11 countries for defensive registrations, whereas GoDaddy is used for malicious registrations in 9 out of 11 countries. Further, GoDaddy is common in both defensive and malicious category in Brazil. Similarly, for India, Net 4 India limited, and for China, Alibaba Cloud Computing (Wanwang) are involved in both defensive and malicious domain registrations.

TABLE IV: Top 3 words and top TLDs in defensive and malicious domain registrations for each country.

Country	Top 3 Words (%)	Top TLDs (%)
US	**online**(13.8), direct(8.4), banking(8.4)	com(100.0)
Germany	**online**(16.9), my(16.1), **bank**(12.9)	com(62.1), de(37.9)
India	**online**(14.1), **bank**(13.8), india(10.1)	com(43.5), net(19.9), in(17.9), co.in(12.4), net.in(6.3)
Italy	**online**(8.0), **bank**(7.2), login(6.2)	it(84.0), com(9.7), net(6.3)
Australia	**online**(20.0), digital(13.0), **bank**(10.4)	com(87.0), com.au(13.0)
UK	**online**(19.1), digital(12.4), **bank**(11.5)	com(60.8), co.uk(39.2)
China	china(31.4), **bank**(23.1), ebank(6.4)	com(40.4), cn(34.0), com.cn(25.6)
Brazil	**bank**(26.8), net(23.2), banco(22.0)	com(73.2), com.br(26.8)
France	banque(35.7), **bank**(25.7), secure(11.4)	com(62.9), fr(37.1)
Canada	**bank**(18.8), **online**(16.2), banking(12.5)	com(80.0), ca(20.0)
Russia	i(20.0), **online**(15.8), **bank**(15.8)	com(61.7), ru(38.3)

F. Overall Distribution of Categories in Squatting Types

Domain squatting wise distribution is as follows: 2,074 (50.42%) are combosquatting domains, 1,128 (27.42%) are comboTLDsquatting domains, 443 (10.77%) are brandname squatting domains, 237 (5.76%) are TLDsquatting domains and 231 (5.61%) are fullname squatting domains. The overall distribution of different categories across different forms of domain squatting is depicted in Figure 8. The highest percentage of defensive registrations were observed for TLDsquatting domains (37.97%) followed by fullname squatting domains (36.36%). However, the number of malicious and suspicious

TABLE V: Registrars' popularity in defensive and malicious domain registrations for each country.

Country	Top 2 Registrars in Defensive Registrations	% of Defensive Registrations	Top 2 Registrars in Malicious Registrations	% of Malicious Registrations
US	**CSC CORPORATE DOMAINS**, MarkMonitor	67.91	**GoDaddy.com**, ENOM	34.55
Germany	PSI-USA dba Domain Robot, Ascio Technologies	28.89	**GoDaddy.com**, UNIREGISTRAR CORP	23.73
India	Net 4 India Limited, Endurance Domains Technology	70	**GoDaddy.com**, Net 4 India Limited	35.63
Italy	COREhub, Telecom Italia s.p.a.	24.72	**GoDaddy.com**, TurnCommerce DBA NameBright.com	12.5
Australia	**CSC CORPORATE DOMAINS**, Corporation Service Company (Aust) Pty Ltd	96.15	**GoDaddy.com**, DYNADOT	31.91
UK	MarkMonitor, **CSC CORPORATE DOMAINS**	50	**GoDaddy.com**, Media Elite Ltd	21.43
China	Alibaba Cloud Computing (Wanwang)	100	Alibaba Cloud Computing (Beijing), Alibaba Cloud Computing (Wanwang)	23.21
Brazil	**CSC CORPORATE DOMAINS**, **GoDaddy.com**	25	**GoDaddy.com**, Megazone Corp dba HOSTING.KR	24.32
France	**CSC CORPORATE DOMAINS**, NAMESHIELD	47.62	OVH, SAS	35.29
Canada	CSC Corporate Domains (Canada) Company, **CSC CORPORATE DOMAINS**	46.67	**GoDaddy.com**, PDR Ltd. d/b/a PublicDomainRegistry.com	21.74
Russia	RU-CENTER-RU, JSC dba RU-CENTER	92.31	**GoDaddy.com**, SALENAMES-RU	25

registrations together still exceeds defensive registrations in these two squatting types with 56.54% and 61.90% respectively. The percentage of defensive registrations for comboTLDsquatting domains were the lowest (6.82%). We note that it is much easier for an organization to do defensive registration for TLDsquatting domains as the number of TLDs is few. However, it is more difficult in case of comboTLDsquatting as the number of possibilities is large.

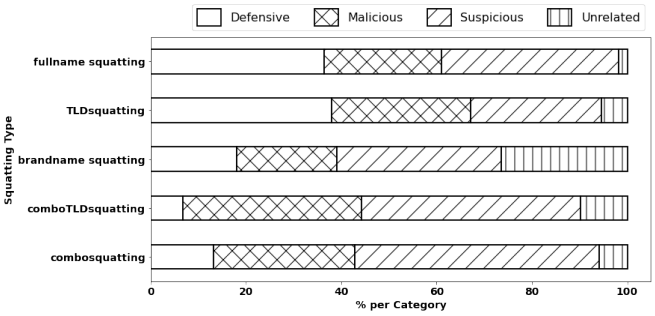
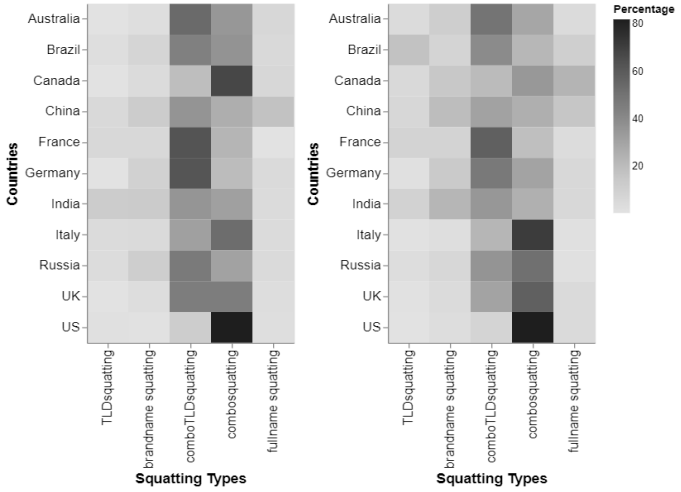


Fig. 8: Overall category distribution in domain squattings.

Both malicious and suspicious categories were dominant in all squatting types. Particularly, combosquatting has the largest fraction of suspicious domains (51.25%) followed by comboTLDsquatting (45.92%). Further, comboTLDsquatting has the highest fraction of malicious domains (37.5%) followed by combosquatting (29.55%).

Figures 9 shows country-wise distribution of squatting types for malicious and suspicious domains. Percentages are shown in grayscale and their sum for each country is 100%. In both Figure 9a and Figure 9b, combosquatting and comboTLDsquatting domains are prevalent. Most of the malicious domains in US (84.39%), Canada (69.57%), Italy (54.16%) employ combosquatting. The top 3 countries with most comboTLDsquatting domains are France (64.71%), Germany (64.41%), Australia (55.32%). Most of the suspicious domains in US (82.11%), Italy (71.41%) and UK (58.56%) are combosquatting. The top 3 countries with most comboTLDsquatting domains are France (58.49%), Australia (50%) and Germany (47.69%).



(a) Malicious Registrations (b) Suspicious Registrations

Fig. 9: Country-wise distribution of squatting types in a) Malicious and b) Suspicious domain registrations.

V. Related Work

Phishing is a well-known and well-studied problem, in this section we review prior work on domain squatting attacks and techniques used to classify domains into malicious or benign. Agten et al. [8] performed a seven-month-long longitudinal experiment in which they visited the typosquatting domains targeting the 500 most popular websites. They found that 95% of the most popular domains are targeted by phishers, and most of them do not use any defensive registrations. A similar study was performed by Kintis et al. [13] to assess the impact of combosquatting abuse. They analyzed more than 468 billion DNS records and identified 2.7 million combosquatting domains that targeted more than 268 most popular trademarks in US. Their work also established that combosquatting domains are 100 times more prevalent than typosquatting domains. Quinkert et al. [11] studied homograph domains by monitoring a daily feed of newly registered domains over a

period of eight months. They detected around 3000 candidate domains targeting 819 distinct reference domains.They also observed that defensive registrations of homograph domains were done in very limited scope. There are some well-known tools that generate synthetic URLs from a seed URL, for example *dnstwist*[1], however they are primarily designed for typosquatting domains. Szurdi et al. [7] studied typosquatting domain registrations within 'com' TLDs. For this they designed a tool which considers passive and active domain features such as WHOIS record information, DNS record information and web page content information for identifying and categorizing domains into different categories. In this paper we are extending this approach for other TLDs, especially we consider different gTLDs (for example, com and net), as well as ccTLDs (for example, de, in, uk, cn, it). It is observed that financial institution are at a higher risk of phishing attacks. Vargas et al. [32] presented their findings on targeted attacks on a major financial institution in the US. They used HTML structure, content analysis, DNS RRSets and domain registration records for finding patterns and correlations between phishing attacks. They identified different strategies used by criminal organizations, valuable insight into who is targeting the institution and attacker's modus operandi. The vulnerable situation of financial institution are further studied by Bijmans et al. [33]. In their work, they studied the use of TLS certificates by malicious actors to uncover possible phishing domains targeting the Dutch financial sector. They collected 70 different Dutch phishing kits in the underground economy and identified ten distinct kit families. The ongoing Covid-19 pandemic has further contributed to the number of phishing attacks. Bitaab et al. [34] performed a comprehensive study of phishing attacks at the start of the pandemic by collecting and analyzing DNS records, TLS certificates, phishing URLs, and source code of phishing websites. Based on their collected datasets, they tracked trends and consequences of phishing activities in the early months of pandemic.

In this work we are primarily focusing on prevalence of domain squatting attacks on the financial sector as it is most affected area, and also it directly affects the end users. We propose a novel method to generate domain squatting instances which includes combosquatting, tldsquatting and combination of both by exploiting inconsistencies present in benign domains using a Context Free Grammar (CFG). Proactive domain name generation from benign domains and their analysis is shown to be effective for typosquatting by Wang et al. [5], however no such work exists for combosquatting. Similarly, Bahnsen et al. [35] proposed DeepPhish algorithm that learns the intrinsic patterns from malicious URLs for generating synthetic domains, however their work is aimed at bypassing the AI phishing detection algorithms, whereas our focus is on defensive registration.

[1]https://github.com/elceef/dnstwist (Accessed: 20 September, 2021)

VI. Discussion and Conclusion

In this paper, we studied the online banking domain squatting landscape of 13 major countries. We showed that online domain names registered by banking organizations do not follow any common pattern which can be exploited by domain squatters. We designed the first CFG based domain generation algorithm that models inconsistencies in the benign domain names of each country and generates candidates from five different domain squatting techniques, namely combosquatting, TLDsquatting, comboTLDsquatting, fullname squatting and brandname squatting. We used WHOIS records, DNS records and web page content of candidate domains to categorize them into 4 different categories, namely defensive, malicious, suspicious and unrelated. We found that 4,113 candidate domains were already registered, of which only 606 domains (14.73%) are defensive, whereas 3,140 domains (76.34%) are malicious or suspicious. Out of 606 defensively registered domains, only 45.70% redirect to primary website (benign), whereas almost 50% either fail to resolve to an IP address or show an error page. A large number of malicious and suspicious registrations are instances of combosquatting and comboTLDsquatting. We observed that only few organizations were protecting their brands against domain squatting abuse by performing defensive registration. Further, we also identified different strategies used by malicious actors during domain registration in order to evade detection from security researchers and trick victims into disclosing their credentials. We recommend banking organizations to use the proposed CFG model to generate probable squatting domains, monitor their activity, try to resolve them through ICANN's Uniform Domain-Name Dispute-Resolution Policy (UDRP) and take action accordingly.

In this study, we highlighted inconsistencies in online banking domains in major countries. However, we also observed diversity in the domain names registered by banking organizations for other services such as credit card, loans and insurance. We aim to study their domain squatting landscape using our CFG algorithm in the future. Also, we plan to extend this study by including other types of attacks such as homograph attack, and compare their effectiveness.

References

[1] "The value of online banking channels in a mobile-centric world," https://www2.deloitte.com/us/en/insights/industry/financial-services/online-banking-usage-in-mobile-centric-world.html, 2020.

[2] "Digital Banking: Banking-as-a-Service, Open Banking & Digital Transformation 2020-2024," https://www.juniperresearch.com/researchstore/fintech-payments/digital-banking-trends-report, 2020.

[3] "The Cost of Cybercrime," https://www.accenture.com/_acnmedia/pdf-96/accenture-2019-cost-of-cybercrime-study-final.pdf, 2020.

[4] Anti-Phishing Work Group, "Malicious Activity Trends Report: 2nd Quarter 2021," https://docs.apwg.org/reports/apwg_trends_report_q2_2021.pdf, 2021.

[5] Y.-M. Wang, D. Beck, J. Wang, C. Verbowski, and B. Daniels, "Strider Typo-patrol: Discovery and Analysis of Systematic Typo-squatting," in *Proceedings of the 2Nd Conference on Steps to Reducing Unwanted Traffic on the Internet - Volume 2*, ser. SRUTI'06. Berkeley, CA, USA: USENIX Association, 2006, pp. 5–5. [Online]. Available: http://dl.acm.org/citation.cfm?id=1251296.1251301

[6] A. Banerjee, D. Barman, M. Faloutsos, and L. N. Bhuyan, "Cyber-fraud is one typo away," in *IEEE INFOCOM 2008 - The 27th Conference on Computer Communications*, April 2008, pp. 1939–1947.

[7] J. Szurdi, B. Kocso, G. Cseh, J. Spring, M. Felegyhazi, and C. Kanich, "The long "taile" of typosquatting domain names," in *23rd USENIX Security Symposium (USENIX Security 14)*. San Diego, CA: USENIX Association, Aug. 2014, pp. 191–206.

[8] P. Agten, W. Joosen, F. Piessens, and N. Nikiforakis, "Seven Months' Worth of Mistakes: A Longitudinal Study of Typosquatting Abuse," in *Proceedings of the 22nd Network and Distributed System Security Symposium (NDSS 2015)*. Internet Society, 2015.

[9] E. Gabrilovich and A. Gontmakher, "The Homograph Attack," *Commun. ACM*, vol. 45, no. 2, pp. 128–, Feb. 2002. [Online]. Available: http://doi.acm.org/10.1145/503124.503156

[10] T. Holgers, D. E. Watson, and S. D. Gribble, "Cutting through the confusion: A measurement study of homograph attacks," in *Proceedings of the Annual Conference on USENIX '06 Annual Technical Conference*, ser. ATEC '06. Berkeley, CA, USA: USENIX Association, 2006, pp. 24–24. [Online]. Available: http://dl.acm.org/citation.cfm?id=1267359.1267383

[11] F. Quinkert, T. Lauinger, W. Robertson, E. Kirda, and T. Holz, "It's not what it looks like: Measuring attacks and defensive registrations of homograph domains," in *2019 IEEE Conference on Communications and Network Security (CNS)*, June 2019, pp. 259–267.

[12] V. Le Pochat, T. Van Goethem, and W. Joosen, "Funny accents: Exploring genuine interest in internationalized domain names," in *International Conference on Passive and Active Network Measurement*. Springer, 2019, pp. 178–194.

[13] P. Kintis, N. Miramirkhani, C. Lever, Y. Chen, R. Romero-Gómez, N. Pitropakis, N. Nikiforakis, and M. Antonakakis, "Hiding in Plain Sight: A Longitudinal Study of Combosquatting Abuse," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '17. New York, NY, USA: ACM, 2017, pp. 569–586. [Online]. Available: http://doi.acm.org/10.1145/3133956.3134002

[14] "Forbes list of world's best banks," https://www.forbes.com/worlds-best-banks/#57edf46f1295, 2020.

[15] N. Chomsky, "Three models for the description of language," *IRE Transactions on Information Theory*, vol. 2, no. 3, pp. 113–124, 1956.

[16] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[17] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 1st ed. USA: Prentice Hall PTR, 2000.

[18] M. Weir, S. Aggarwal, B. d. Medeiros, and B. Glodek, "Password cracking using probabilistic context-free grammars," in *2009 30th IEEE Symposium on Security and Privacy*, 2009, pp. 391–405.

[19] S. Houshmand, S. Aggarwal, and R. Flood, "Next gen pcfg password cracking," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 8, pp. 1776–1791, 2015.

[20] D. Wang, Z. Zhang, P. Wang, J. Yan, and X. Huang, "Targeted online password guessing: An underestimated threat," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1242–1254. [Online]. Available: https://doi.org/10.1145/2976749.2978339

[21] N. Miramirkhani, O. Starov, and N. Nikiforakis, "Dial one for scam: A large-scale analysis of technical support scams," *Proceedings 2017 Network and Distributed System Security Symposium*, 2017. [Online]. Available: http://dx.doi.org/10.14722/ndss.2017.23163

[22] "Recursive malvertising: over one thousand URLs for sale redirect users to pages from the denylist –including malicious ones," https://usa.kaspersky.com/about/press-releases/2020_recursive-malvertising-over-one-thousand-urls-for-sale-redirect-users-to-pages-from-the-denylist-including-malicious-ones, 2020.

[23] N. Miramirkhani, "Methodologies and tools to study malicious ecosystems," 2020.

[24] T. Lauinger, A. Chaabane, A. S. Buyukkayhan, K. Onarlioglu, and W. Robertson, "Game of registrars: An empirical analysis of post-expiration domain name takeovers," in *26th {USENIX} Security Symposium ({USENIX} Security 17)*, 2017, pp. 865–880.

[25] "Expired domain names and malvertising," https://blog.malwarebytes.com/threat-analysis/2017/09/expired-domain-names-and-malvertising/, 2020.

[26] "A Plugin's Expired Domain Poses a Security Threat to Websites," https://blog.sucuri.net/2016/08/plugin-expired-domain-security-threat.html, 2016.

[27] M. Wu, R. C. Miller, and S. L. Garfinkel, "Do security toolbars actually prevent phishing attacks?" in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 601–610. [Online]. Available: https://doi.org/10.1145/1124772.1124863

[28] "Over 60,000 parked domains were vulnerable to AWS hijacking," https://www.bleepingcomputer.com/news/security/over-60-000-parked-domains-were-vulnerable-to-aws-hijacking/, 2021.

[29] S. Alrwais, K. Yuan, E. Alowaisheq, Z. Li, and X. Wang, "Understanding the dark side of domain parking," in *23rd USENIX Security Symposium (USENIX Security 14)*. San Diego, CA: USENIX Association, Aug. 2014, pp. 207–222.

[30] T. Vissers, W. Joosen, and N. Nikiforakis, "Parking sensors: Analyzing and detecting parked domains," in *Proceedings of the 22nd Network and Distributed System Security Symposium (NDSS 2015)*. Internet Society, 2015, pp. 53–53.

[31] N. Nikiforakis, S. Van Acker, W. Meert, L. Desmet, F. Piessens, and W. Joosen, "Bitsquatting: Exploiting bit-flips for fun, or profit?" in *Proceedings of the 22nd International Conference on World Wide Web*, ser. WWW '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 989–998. [Online]. Available: https://doi.org/10.1145/2488388.2488474

[32] J. Vargas, A. C. Bahnsen, S. Villegas, and D. Ingevaldson, "Knowing your enemies: Leveraging data analysis to expose phishing patterns against a major us financial institution," in *2016 APWG Symposium on Electronic Crime Research (eCrime)*. IEEE, 2016, pp. 1–10.

[33] H. Bijmans, T. Booij, A. Schwedersky, A. Nedgabat, and R. van Wegberg, "Catching phishers by their bait: Investigating the dutch phishing landscape through phishing kit detection," in *Proceedings of the 30th USENIX Security Symposium*. USENIX Association, 2021, pp. 3757–3774.

[34] M. Bitaab, H. Cho, A. Oest, P. Zhang, Z. Sun, R. Pourmohamad, D. Kim, T. Bao, R. Wang, Y. Shoshitaishvili *et al.*, "Scam pandemic: How attackers exploit public fear through phishing," in *2020 APWG Symposium on Electronic Crime Research (eCrime)*. IEEE, 2020, pp. 1–10.

[35] A. C. Bahnsen, I. Torroledo, L. D. Camacho, and S. Villegas, "Deephish: simulating malicious ai," in *2018 APWG symposium on electronic crime research (eCrime)*, 2018, pp. 1–8.

APPENDIX A
DOMAIN NAME REGISTRATION

Domains are important element in the current internet space. The Domain Name System (DNS) is used to translate domain names to IP addresses and vice versa. In order to have online presence, organizations need to register domains with domain registrars that are accredited by ICANN. For registering a domain name, the organization needs to submit the following information to a registrar:

- Domain name
- Domain registrant's information including name and contact information such as email id, contact phone number and physical address, administrative and billing contact details
- Domain registration period

After receiving these details, registrar sends the domain name request to the relevant domain name registry which maintains the database of all domain names in the requested TLD. Domains can be registered in different TLDs such as generic TLDs (gTLDs) and country-code TLDs (ccTLDs).

Generic TLDs are not restricted to any country, and organizations across the world can apply to register their domain names in gTLDs. Few examples of gTLDs are *com*, *net*, *info* and *org*. Country-code TLDs are specific to countries. Few examples are *uk* for United Kingdom, *it* for Italy, *in* for India and *cn* for China.

Once the domain is registered, its WHOIS record is created which consists of domain registrant's information, registrar's information, and domain's creation, updation and expiration date. WHOIS records play an important role in deciding whether the registered domain is benign or malicious [8], [11], [13].

Appendix B
Domain Squatting Techniques

Domain squatting is the practice of strategically registering domain names that are confusingly similar to those belonging to popular brands. Such domains are used for abusive activities such as phishing, distribution of ads and malware, and social engineering attacks. There are different forms of domain squatting including typosquatting, homophone-based squatting, bitsquatting, homograph-based squatting and combosquatting. Combosquatting is more prevalent than other squatting types [13].

In this paper, in addition to *combosquatting* and *TLDsquatting*, we study three new forms of domain squatting, namely *comboTLDsquatting*, *fullname squatting* and *brandname squatting*. We explain all squatting types using the benign domain rbsdigital.com registered by Royal Bank of Scotland (RBS).

Combosquatting domains combine a popular brand name with a phrase. The domains rbsonline.com, onlinerbs.com and digitalrbs.com are examples of combosquatting.

TLDsquatting is the act of replacing original TLD within benign domains with other TLDs. An example of TLDsquatting domain is rbsdigital.co.uk.

ComboTLDsquatting domains are the combination of combosquatting and TLDsquatting. These domains not only combine a popular brand with a phrase but also use a different TLD. The domain rbsonline.co.uk is a comboTLDsquatting domain since it uses a different word (online) and TLD (co.uk) compared to the benign domain rbsdigital.com.

In ***fullname squatting***, the attacker registers the full name of an organization. For instance, the fullname squatting domain for the Royal Bank of Scotland would be royalbankofscotland.com.

In ***brandname squatting***, the brand name used in SLD of the benign domain is registered. For instance, the brand name in rbsdigital.com is rbs. Hence, the domain rbs.com is an example of brandname squatting.

Appendix C
Derivation of Maximum Cardinality of Generated Domains using CFG

We calculate the upper bound for the number of domains that can be generated from the starting symbol S of our CFG. As the resulting grammar could be ambiguous, the total number of domains n that can be generated beginning from the start symbol S is at most $|S|$. In the derivation below, we use the fact that both the number of brand names $|B|$ and organization names $|F|$ are equal to the number of tuples in the input set I, i.e., $|B| = |F| = |I|$. However, the set of generated domains L also contains the benign domains from I. Hence, we remove such domains from L.

$$
\begin{aligned}
n &\leq |S| \\
&= |C.T| + |B.T| + |F.T| \\
&= |C| \cdot |T| + |B| \cdot |T| + |F| \cdot |T| \\
&= (|B| \cdot |W| + |W| \cdot |B|) \cdot |T| + |B| \cdot |T| + |F| \cdot |T| \\
&= (2 \cdot |W| \cdot |B| + |B| + |F|) \cdot |T| \\
&= (2 \cdot |W| + 2) \cdot |I| \cdot |T| \\
&= 2 \cdot (|W| + 1) \cdot |I| \cdot |T| \\
n &\leq 2 \cdot (|W| + 1) \cdot |I| \cdot |T|
\end{aligned}
$$

Appendix D
Parse Tree

Figure 10 shows the derivation of comboTLDsquatting domain sbionline.in using a parse tree. As usual, the derivation begins at start symbol S which is replaced with the rule $C.T$. The non-terminal C is replaced with BW. B is substituted with sbi, W with $online$ and T with in. The derived string is available in leaf nodes and is read from left-to-right. The height of the parse tree is the longest path (number of edges) from the root node to a leaf node. In this case, the height of the tree is 3. In fact, the maximum height of the parse tree for any string generated using our CFG never exceeds 3.

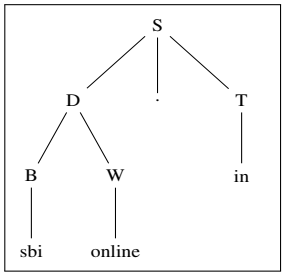


Fig. 10: Parse tree of sbionline.in, a potential comboTLDsquatting domain generated using CFG G_1.

Appendix E
Different Categories

Based on the analysis of WHOIS records, DNS records and web page collected for each candidate domain, we classify the domains into four different categories, namely *defensive*, *unrelated*, *suspicious* and *malicious*. These categories are further subdivided as shown in Figure 11.

Appendix F
Web Page Analysis

The flowchart categorized around 61% of the registered candidate domains based on WHOIS information, HTTP status

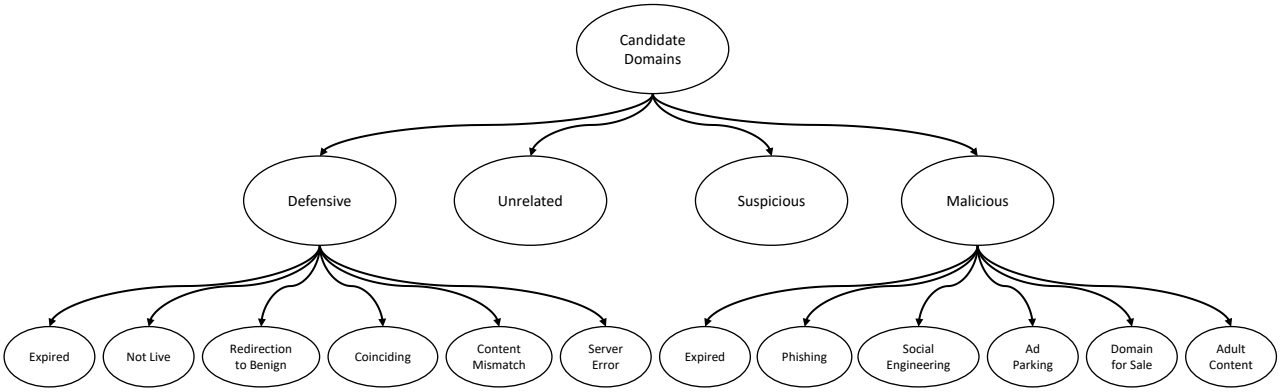


Fig. 11: Different categories assigned to the generated candidate domains.

codes and DNS records. The remaining 39% of the candidate domains required web page analysis. We used visual characteristics of a web page to decide whether it belonged to phishing, social engineering, ad parking, domain for sale and adult content category. During the data crawling phase, we had gathered screenshots for each domain that resolved to an IP address. We found that ad parking, domain for sale and suspicious web pages are visually similar across all countries. We used this observation to automate the task of classifying web pages using an image based hashing technique. We calculated average hash values of the screenshots as they are robust to small changes in the input as opposed to crypotographical hashing techniques. The average hash converts the visual characteristics present in the screenshot of a web page in a numeric form [8].

Our methodology to categorize a web page based on its screenshot is as follows. We labeled a small subset of web pages pertaining to ad parking, domain for sale and suspicious manually. We computed the average hash value of each web page and stored it in a template dictionary. To categorize the remaining set of web pages, we calculated the average hash value of each web page and compared it with the hash value of each image stored in the dictionary. If the difference is less than a threshold value (determined heuristically), we assigned the web page the category of the template that produced the least difference. If the difference is above threshold, we inspected the web page manually and assigned it a relevant category. Of the web pages requiring content analysis, 45% were classified using the image based hashing technique and the rest were classified manually.