Typo-Squatting: The "Curse" of Popularity

Alessandro Linari[1,2] Faye Mitchell[1] David Duce[1] Stephen Morris[2]

[1]Oxford Brookes University, [2]Nominet UK

{alinari,frmitchell,daduce}@brookes.ac.uk, stephen.morris@nominet.org.uk

Abstract

Typo-squatting is the practice of registering a domain name with the intent to confuse it with the name of a trademark or a famous other domain name. In this paper we study typo-squatting from a statistical point of view. We introduce the concepts of syntactic and visual neighbourhoods of a domain name as the sets of all other domain names which are, respectively, syntactically or visually similar to the original domain. The results of our preliminary experiments on the '`.co.uk`' registry show a strong correlation between the popularity of a domain name and the size of its syntactical and visual neighbourhoods. This suggests the size of the neighbourhood can be used as a reliable indicator for the likelihood of being typo-squatted. We conclude the paper with a brief discussion of the implications of our work in the field of online digital identities.

1 Introduction

Since the early years of the Web, domain names not only represent a mnemonic easy-to-remember way of accessing a website, but are also an important element in the marketing strategy of many companies and brands. Like more traditional media, this has led to the issues with trademark and brand protection.

Typo-squatting is the practice of registering a domain name with the intent to confuse it with the name of a trademark or a famous other domain name. In other words (at least informally) it is a *trademark infringement* on domain names over the Internet. The motivations behind this practice can be very diverse: economical reasons (pay-per-click links), the spreading of viruses and the delivery of unsolicited publicity (e.g. pornographic websites) are but a few of them. In general, typo-squatting poses a threat that may result in a loss of identity, reputation or trust.

The key elements in cases of traditional trademark infringement are the syntactic and graphical *similarity* between the original brand and its imitator and the *confusion* that such similarity induces on the users, in relation to the *context* where the two names are used. Domain names, on the other hand, do not have a graphic element associated with them, nor does the DNS (Domain Name System) provide a context when users manually type names in the browser's address bar or click links in a web page.

Two domain names are said to be *syntactically similar* when they differ by a limited amount of typographical or common misspelling errors. This is the case of '`gopgle`' and '`googel`', two misspellings of '`google`', where a '`p`' has been typed instead of the second '`o`' and the '`l`' and '`e`' have been transposed, respectively. Two domain names are said to be *visually similar* when their appearance can confuse a user who reads them: consider, for example, the strings '`google`' and '`goog1e`', where the digit '`1`' (one) can be easily confused with the character '`l`'.

In this paper, we focus on the similarity between a domain name and its *typo-squatters*, i.e. the domain names that are in a typo-squatting relationship to the given name, and provide a numerical measure to characterise this relationship. We introduce the concept of neighbourhood of a domain name in the space of all existing domains and, by means of a metric, define neighbourhoods of interest, i.e. sets of all other domains which lie within a certain distance according to the chosen metric.

We assess the validity of this approach by applying it to the set of third-level '`.co.uk`' domain names (e.g. in '`example.co.uk`', the label '`uk`' is the top-level domain, '`co`' is the second-level domain and '`example`' is the third-level domain). Although these are preliminary results, they clearly show the soundness of our approach, especially considering that the dataset to which the analysis is applied (the `.co.uk` registry) is one of the largest in the world.

2 Analysis

A common way of estimating the similarity $sim(x_1, x_2)$ between two strings x_1 and x_2 is by means of a distance function $d()$, which is inversely proportional to $sim()$. A general purpose distance function for string comparison is the *edit distance with transpositions*, denoted as $d_{et}()$, which computes the minimum number of operations (insertions, deletions, substitutions of a single character and transpositions of adjacent characters) needed to transform a string x_1 into another string x_2 (a good survey on comparison functions based on the edit distance can be found in [6]).

We estimate the syntactic similarity between domain names using the *keyboard distance* $d_{kb}()$, which has been derived from $d_{et}()$ by giving higher importance to operations which correspond to frequent typing errors. In our implementation, we have assigned a higher similarity (i.e. a lower value of the keyboard distance) to transpositions and substitutions/insertions of adjacent characters in the keyboard. (For simplicity, we have not considered layouts other than that of the "qwerty" keyboard.)

The visual similarity between domain names is estimated by means of the *visual distance* $d_{viz}()$, derived from the similarity function presented by Black in [2] to compare top-level domains[1]. In our implementation, the function has been turned into a distance, i.e. small values correspond to higher similarity, and a normalisation factor has been removed. The resulting $d_{viz}()$ is a function which extends $d_{et}()$ in two ways: it assigns smaller costs to operations which involve visually similar characters (the original weights having been retained) and adds a fifth operation to the original set of four used in $d_{et}()$, to take digraph substitutions into account (e.g. "mn" for "nm" or "cl" for "d").

In order to estimate the likelihood that a domain name x has been typo-squatted, we rely on the *size of the neighbourhood* of x. The *neighbourhood* $\mathcal{N}_d(x)$ of a domain name x is defined as the set of all domain names in the registry whose distance $d()$ from x is lower than a threshold th_d and we denote with $|\mathcal{N}_d(x)|$ its cardinality (or size). The *syntactic neighbourhood* of x, denoted as $\mathcal{N}_{d_{kb}}(x) \equiv \mathcal{N}_{syn}(x)$, is the set of all domain names which are *syntactically* similar to x. Analogously, the *visual neighbourhood* $\mathcal{N}_{d_{viz}}(x) \equiv \mathcal{N}_{viz}(x)$ is the set of domains visually similar to x.

In both cases, we have decided to set $th_{kb} = th_{viz} = 1$, where th_{kb} and th_{viz} are the thresholds associated to the keyboard and visual distances, respectively. As reported in [3], 80% of misspelled words contain one single error, although Pollock and Zamora [7] suggest that the figure is closer to 90-95%. According to $d_{et}()$ this corresponds to a distance of exactly 1. With the introduction of weights, however, operations in both d_{kb} and d_{viz} are associated with costs which can be smaller than 1 and distances can assume fractional values. The result is a net increase of the density of points in the unit disc centred at x (i.e. $|\mathcal{N}_{d_{et}}(x)|$ with threshold 1). As a consequence, defining the neighbourhood of a domain name x as the set of other domain names whose distance from x measured by d_{kb} or d_{viz} is *lower* than 1 is like selecting any domain name which shares with x *additional elements of similarity*, over and above those normally identified by $d_{et}()$.

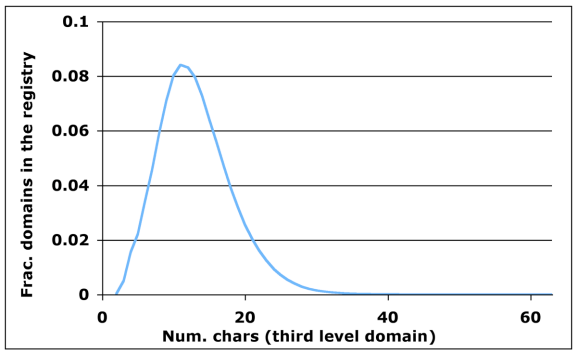
In the next section, we will present our experiments on the size of the syntactic and visual neighbourhoods for both random and popular domain names. Previous studies [1, 5, 9] (as well as practical evidence) suggest that popular domain names are the most affected by typo-squatting. Therefore, we expect, and our experiments confirm, a larger-than-average size of the neighbourhood for this class of domain names, independent with the distance function that is being used.

3 Experiments

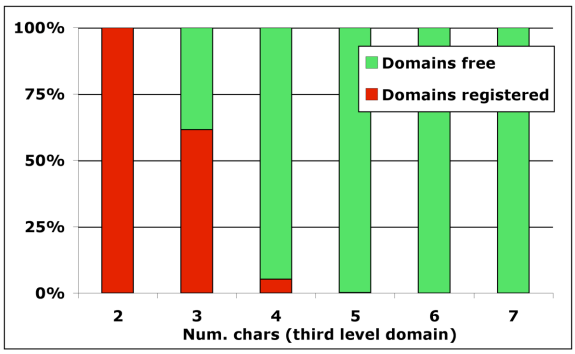
In this section we show preliminary results from the third-level domains `.co.uk` registry. Our experiments were conducted on a snapshot of the database taken in March 2008 which consisted of about six million entries. We have chosen the 1000 most popular domains (according to research conducted by NetCraft[2], March 2008) and partitioned them into three "bands of popularity": the first 100 domains (band A), the domains ranked $101 - 500$ (band B) and the domains ranked $501 - 1000$ (band C). In

[1] The code for Black's similarity function can be downloaded from [2].

[2] `http://www.netcraft.com`. The ranking is determined by the number of accesses the website received from the NetCraft toolbar in the analysed period.



(a)



(b)

Figure 1: (a) Distribution of lengths of `.co.uk` domain names and (b) fraction of registered domains vs. length of the third-level label.

addition, we have defined a fourth band (*Random*) containing 100 domains randomly chosen from the original dataset in the registry.

Figure 1(a) shows that the distribution of third-level domain names is not uniform with respect to their length. It should be noted that the length of the domain names also has a strong influence on the distance functions presented in the previous section: give any two domain names x_1 and x_2 of lengths $len(x_1)$ and $len(x_2)$, respectively, their distance $d(x_1, x_2)$ is bounded by $[0, d_{max}(x_1, x_2)]$, where $d_{max}(x_1, x_2) = max(len(x_1), len(x_2))$. As a consequence, for any domain x in the registry, the probability of finding any other domain at a short distance is inversely proportional to its length $len(x)$. In general, we intuitively expect:

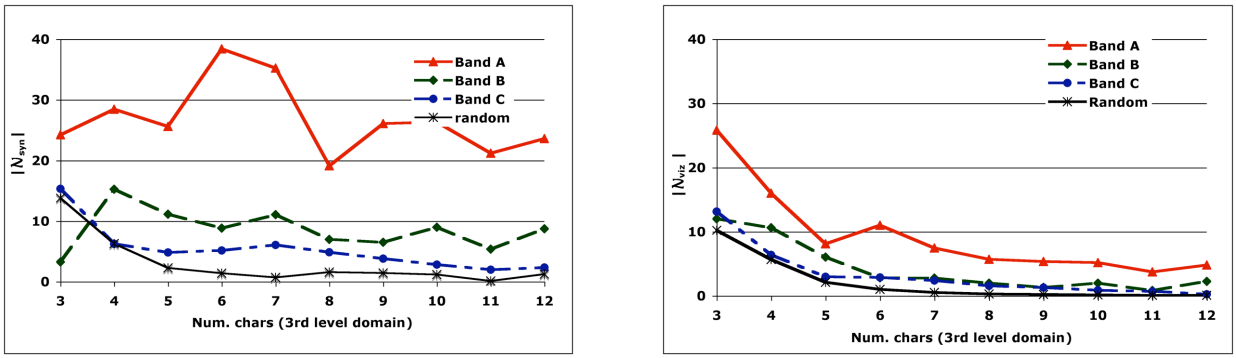
- $|\mathcal{N}_{syn}(x)|$ an d $|\mathcal{N}_{viz}(x)|$ to be inversely proportional to $len(x)$;
- The probability of "random collisions" (i.e. two domain names that are in each other's neighbourhood, but are not in a typo-squatting relationship) to be inversely proportional to $len(x)$.

A direct consequence of the latter is that our analysis is more reliable for longer domain names because the probability of false positives decreases. If we now consider the fraction of name space that has been registered shown in Figure 1(b), we notice that all 2-character and the vast majority of 3-character domains have already been registered, which increases the probability of random collisions in this region.

In general, given that a significant part of the neighbourhood of the 4-character domains resides in the 3-character space and given the high level of occupancy of this space, we suggest that reliable results are only obtained for domain names longer than four characters. This is not a too critical limitation, because the 2-4 character space accounts for less than 1.5% of the whole registry. In the reminder of the paper, we have excluded any further reference to 2-character domain names (which cannot be freely registered) and have kept the statistics for 3- and 4- character domains for reference. Note that previous studies (such as [1, 9]) did not investigate this aspect, nor did they consider the effect of the length of a domain name on its associated statistics.

Figures 2(a) and 2(b) show $|\mathcal{N}_{syn}(x)|$ and $|\mathcal{N}_{viz}(x)|$ respectively, for domain names whose lengths range between 3 and 12 characters. (Our analysis does not consider longer domain names because there were very few popular domains in that region of the space.) The curves are not monotonic because they depend on the specific domains in the bands, an effect that is larger for the bands with the more popular domain names. Band A, for example, contains 100 domains *in total* and between 5 and 12 domains per each length. We notice the spikes in the graphs for six-character domain names: these are caused by `google.co.uk` and `amazon.co.uk`, both well-known typo-squatted domains.

The figures show that for both distance functions, and for all lengths considered, there is a strong correlation between the popularity of a domain name and the size of its neighbourhood. In the case of very popular domain names (i.e. those in band A), this correlation holds even for small lengths. This



(a) (b)

Figure 2: Size of the (a) syntactic and (b) visual neighbourhoods for domain names of different popularity.

might suggest an unexpected robustness of the methodology, although additional testing is still needed.

In Figure 2(a), the size of the syntactic neighbourhood for band A domains does not depend on the length. The dependence is partially visible for bands B and C and is stronger for random domains. This apparently contradictory phenomenon (we were expecting $|\mathcal{N}_{syn}(x)| \propto len(x)^{-1}$) suggests that the keyboard distance is particularly accurate in identifying those areas of the space where it is more likely to find typo-squatting domain names. This explains why, as the popularity of a domain name decreases, the curves tend to the behaviour of a randomly chosen domain name.

Figure 2(b) shows the correlation between popularity and the size of the visual neighbourhood. The graph still shows the expected behaviour, in that $|\mathcal{N}_{viz}(x)| \propto popularity(x)$. However, it does not appear to show a good discrimination as the syntactic measure.

4 Related work

A number of industrial reports have dealt with the problem of typo-squatting. One that caught the attention of the media was published by McAfee at the end of 2007 [5] where the spread of typo-squatting among a number of top-level domains was studied. On the academic side, research on typo-squatting has been very limited. Wang *et al.* [9] adopted a definition of neighbourhood of a domain name which is analogous to our syntactic neighbourhood. They also consider typo-squatting domain names obtained by removing the '.' (dot) in the first part of the identifying URL, i.e. domains such as `www<name>.co.uk`, where `<name>` is the typo-squatted domain name. In [4], Holgers *et al.* investigate the spread of the *homograph attack*, which consists in registering a domain name where single characters are substituted with visually similar ones. They considered Unicode characters, but restricted their study to single-character substitutions. Interestingly enough, their results are consistent with our findings on the correlation between popularity and likelihood of being typo-squatted. Banerjee *et al.* [1] choose popular domain names and analysed three classes of typo-squats, those that can be obtained by deletion, insertion and substitution of a single character.

A characteristic common to all previous studies of typo-squatting is the lack of access to the registry, which forces the estimation of $|\mathcal{N}_d(x)|$ by generating all possible typo-squatting domain names and testing if they were registered. A second limitation consists in not having considered the effect of the length of the domain name. Consider, for example, the '.com' name space, which has been the subject of various studies (e.g. [1, 4, 9]). At the time those analyses took place, this registry contained more than 60 millions domains [8], which is one order of magnitude higher than our dataset. Although we have not checked, we would not be surprised if the fraction of the 4-character namespace registered is equal to or higher to that in `.co.uk`.

5 Discussion

The use of a distance function and of the concept of neighbourhood allows us to give a geometrical interpretation to the results presented so far. If we consider the space of all possible domain names, the neighbourhood is a cluster of domains close to each other (according to the distance function). According to our analysis, a cluster is evidence that a typo-squatting activity has taken place in that region of the space. Therefore, it identifies a region where a similar activity might take place in the future, but also suggests a method of observing and measuring the evolution of typo-squatted clusters.

The theoretical approach we are undertaking can be applied to other contexts, to characterise analogous malicious activities that can be associated to the very general case of the *theft of identity*. In social networks and Web 2.0 applications, for example, the concept of *online identity* is still very loose, but plays an increasingly important role in the future development of these technologies. The risks associated with a phenomenon similar to typo-squatting in such a scenario are still not fully understood, but might have a disruptive effect in terms of trust and reputation to the person or company/association to whom the identity belongs.

6 Acknowledgements

This work was supported by the UK government as part of the Knowledge Transfer Partnerships programme number 6502.

We thank Nominet for providing full access to the '.co.uk' database and all staff for their support and useful feedback.

References

[1] A. Banerjee, D. Barman, M. Faloutsos, and L. N. Bhuyan. Cyber-fraud is one typo away. In *INFOCOM*, 2008.

[2] P. E. Black. Visual similarity of top-level domains. NIST website. `http://hissa.nist.gov/ black/GTLD/` (valid Feb. 2009).

[3] F. Damerau. A technique for computer detection and correction of spelling errors. *Commun. ACM*, 7(3):171–176, 1964.

[4] T. Holgers, D. E. Watson, and S. D. Gribble. Cutting through the confusion: A measurement study of homograph attacks. In *USENIX Annual Technical Conference, General Track*, pages 261–266, 2006.

[5] McAfee. What's in a name: The state of typo-squatting 2007. `http://www.mcafee.com/typosquatters` (valid Jan. 2009).

[6] G. Navarro. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88, 2001.

[7] J. J. Pollock and A. Zamora. Automatic spelling correction in scientific and scholarly text. *Commun. ACM*, 27(4):358–368, 1984.

[8] VeriSign. The domain name industry brief. 6(1), Feb. 2009. Avail. at `http://www.verisign.com/static/044518.pdf` (valid Feb. 2009).

[9] Y.-M. Wang, D. Beck, J. Wang, C. Verbowski, and B. Daniels. Strider typo-patrol: Discovery and analysis of systematic typo-squatting. In *SRUTI (Usenix WS)*, 2006.