Understanding Domain Registration Abuses

Scott E. Coull[1], Andrew M. White[1], Ting-Fang Yen[2],
Fabian Monrose[1], and Michael K. Reiter[1]

[1] University of North Carolina
{coulls,amw,fabian,reiter}@cs.unc.edu
[2] Carnegie Mellon University
tyen@andrew.cmu.edu

Abstract. The ability to monetize domain names through resale or serving ad content has contributed to the rise of questionable practices in acquiring them, including domain-name speculation, tasting, and front running. In this paper, we perform one of the first comprehensive studies of these domain registration practices. In order to characterize the prevalence of domain-name speculation, we derive rules describing "hot" topics from popular Google search queries and apply these rules to a dataset containing all .com registrations for an eight-month period in 2008. We also study the extent of domain tasting throughout this time period and analyze the efficacy of ICANN policies intended to limit tasting activity. Finally, we automatically generate high-quality domain names related to current events in order to measure domain front running by registrars. The results of our experiments shed light on the methods and motivations behind these domain registration practices and in some cases underscore the difficulty in definitively measuring these questionable behaviors.

1 Introduction

Domain names have become an integral component of modern web browsing by allowing users to navigate to web sites using memorable phrases and keywords. In fact, many users will often assume that domain names based on intuitive keywords will direct them to the desired web site, which is known as type-in navigation. For this reason, domain names have become quite valuable, and has lead to a variety of practices in which domain names are opportunistically registered simply to profit from them. One such dubious domain registration practice is *domain speculation*, where a domain name is registered with the intention of reselling it for a profit at a later date or generating ad revenue from type-in navigation traffic [5]. Though speculation is technically allowed by Internet Corporation for Assigned Names and Numbers (ICANN) rules, it has led to more abusive behaviors by registrars and speculators. For instance, *domain tasting* allows a speculator to register large numbers of domains at no cost for a short grace period during which she can assess the potential value of the domain. Another example is *domain front running*, where domain registrars use queries about domain availability made by their users to preemptively register domains then subsequently resell them to those same users for a profit. The

security problem underlying these behaviors is not unlike that presented by spam during its emergence, in that both activities take advantage of loopholes in existing policy to profit from unintended uses of the systems they abuse. While the security and legal communities have identified certain behaviors as clear abuses of the registration process, the practices and impact of domain speculation, tasting, and front running are still not well-understood.

In this paper, we perform the first in-depth study of questionable domain name registration activity, including a characterization of domain speculation, an analysis of the prevalence of domain tasting, an investigation of the possibility of domain front running by popular domain registrars, and an analysis of the impact of ICANN policies on these abusive behaviors. Specifically, we use popular search terms provided by Google to develop association rules that describe keywords that are likely to occur together in domain names related to current events and "hot" topics. These association rules are then used to generate regular expressions that allow us to search over all .com registrations collected by VeriSign over an eight-month period for evidence of domain speculation. They also allow us to automatically generate quality domain names when measuring domain front-running activities. Our results shed light on the motivations underlying abusive registration practices and the difficulties faced in accurately measuring their prevalence.

2 Related Work

The inherent importance of the Domain Name Service (DNS) in enabling navigation of the web makes it a natural target for attackers seeking to misdirect users to malicious web sites. Due to their prevalence and potential impact on users, these misdirection attacks have been widely studied. For example, a handful of recent studies have focused on measuring the prevalence of typosquatting, [8, 3, 6] and homograph [4] attacks, which take advantage of the user's inability to differentiate the intended domain name from what appears on the screen. Unfortunately, studies of the less malicious, yet still questionable, domain registration activities that we examine in this paper appear to be limited primarily to a series of status reports by ICANN[1]. Their analysis of domain tasting examines the use of the five-day add grace period between June 2008 and August 2009. Overall, they observed a significant decrease in tasting activity after a temporary provision was instituted in July 2008 that limited registrars to a relatively small number of no-cost registration deletions. Similarly, their preliminary statement on domain front running activities based on user complaints found that the majority of claims were due to user error or oversights during the registration process. A follow up ICANN report using a study of 100 randomly generated domains found no evidence of front-running activity by registrars. The obvious limitations of that investigation are its relatively small scale and the use of randomly generated names that were easy to identify and ignore.

[1] *c.f.,* `https://st.icann.org/reg-abuse-wg/`

3 Preliminaries

To successfully achieve our goals we need to overcome two challenges. The first lies in decomposing Google search queries about a given topic into combinations of keywords that are likely to appear in domain names related to that topic. Broadly speaking, we assume that the searches that users make on Google about an event or topic are closely related to the domain names they would navigate to using type-in navigation. These type-in navigation domains are prime targets of domain squatters and front runners, and therefore the focus of our investigation. The second challenge lies in developing a method for determining which domains are pertinent to our study. Before proceeding further, we describe the data sources and methods used to address these challenges.

Data Sources. For our study, we make use of a variety of data sources, including both historical domain name registration data and longitudinal information on Google search query popularity. Our historical analysis of domain name registrations is based on data obtained from VeriSign containing 62,605,314 distinct .com domain registration events from March 7, 2008 to October 31, 2008. The VeriSign data contains domain names, their associated name servers, and the date of each registration event. The data also contains information about deregistration events, which we use in our analysis of domain tasting behaviors. For the remainder of the paper, we refer to the set of all domain name registrations contained in the VeriSign data set as the *background* set.

Furthermore, in order to gain a sense of the popularity of various topics or events, we make use of data provided by Google via its Insights for Search and Trends services. These services rank the top searches made by users over a given time frame, and provide up to ten related searches for each. Our methods assume that these queries adequately represent the hot topics that caused their increase in popularity in the Google search engine, and that this increase in search popularity is an indicator of the desirability of domains related to the hot topic. In our study, we use the Insights for Search service to derive rules used to search for domains in our VeriSign data, while the Trends service provides real-time search rankings that are used to generate high-quality domains names to be used in our domain front-running experiment. A *topic* in our study is defined to be a top ranked search, along with its ten related searches.

Association Rule Mining. To decompose each search query into combinations of keywords that best represent the topics associated with them, we apply association rule mining techniques [1]. These techniques consider an itemset $I = \{i_1, \ldots, i_m\}$ containing all items that can be found in a transaction, and a transaction set $T = \{t_1, \ldots, t_n\}$, where t_α is a set containing the items associated with the α^{th} transaction. The *support* of a set of items X is defined as $supp(X) = n_X/n$, where n_X is the number of transactions containing all items in X. An *implication* between sets of items X and Y, denoted as $X \Rightarrow Y$, indicates that the presence of the items in X implies items in Y will also be present. The *confidence* of an implication $X \Rightarrow Y$ is defined as

$conf(X \Rightarrow Y) = supp(X \cup Y)/supp(X)$. An implication is considered to be a *rule* if the sets have a sufficient level of support and confidence.

For our purposes, we use the notions of support and confidence defined above to decompose each Google query into groups of keywords that are specific to the topic at hand. To do so, we consider each of the n searches for a given topic to be transactions with each keyword in the search acting as an item in the transaction's set. We then decompose those searches into sets of co-occurring keywords based on the confidence of the keywords' pairwise implications. Specifically, we first examine each ordered pair of keywords in the search query to discover all of the bidirectional rules (*i.e.,* implications where the confidence in both directions is above our threshold), and merge them together by assuming transitivity among the implications. These bidirectional rules describe the groups of keywords that must appear together in order to be meaningful to the given topic. Next, we augment the rule set by examining unidirectional implications, which indicate that the antecedent of the implication should only be present where the consequent also exists. As before, we assume transitivity among the rules to merge them appropriately. If a keyword is not the antecedent in any rules, we add it as a singleton set. The algorithm returns the union of the rule sets for each of the search queries, which contain all of the groups of keywords that represent the topic associated with those searches.

Due to the inherently noisy nature of the data used in our study, it is important to carefully set thresholds used in our rule mining and other selection procedures. The threshold selection methodology is complicated by the fact that our data provides no notion of what values might be related to a given topic and what is not (*i.e.,* the data is unlabeled). Therefore, we make use of cluster analysis techniques to automatically set the thresholds used in our study, rather than appealing to manually derived thresholds. Specifically, we make the observation that we need only separate two classes of unlabeled values: those that are interesting with respect to our analysis and those that are not. Thus, to determine a threshold we first use the k-means++ algorithm [2] with $k = 2$ to partition the unlabeled values into the sets S_1 and S_2 (*i.e.,* interesting and uninteresting), and set the threshold as the midpoint between these two clusters.

4 Domain Name Speculation

Our first objective is to examine the relationship between new registrations and so-called hot topics in an effort to gain a better understanding of domain speculation. To do so, we follow an iterative process that consists of: (i) generating rules that are specific to the topic at hand, (ii) converting those rules into regular expression to select domains, and (iii) pruning and verifying the set of selected domains to ensure they are, in fact, related to the topic.

First, we gather Google Insights data for each month in 2008, and treat the set of searches related to the topics as transactions, which are used in our association rule mining algorithm to generate rules. Recall that a threshold confidence value dictates which implications in our set of transactions should be considered rules.

Google Query	Association Rule	Regular Expression
clinton vs obama	vs ⇒ clinton ⇒ obama	(.*obama.*) or (.*clinton.*) & (.*obama.*) or (.*clinton.*) & (.*obama.*) & (.*vs.*)

Table 1. Decomposition of a Google query into rules and regular expressions.

To determine this threshold, we calculate the confidence between all pairs of keywords within a topic and use the threshold selection method discussed in Section 3 on these values to set the appropriate threshold. The resulting set of rules are further pruned to ensure that non-specific rules are discarded. To do so, we score each rule r_i for a topic as $S(r_i) = \sum_{k \in r_i} supp(k) \times |k|$, where r_i is a rule for the current topic (represented as a set of keywords), $supp(k)$ is the support of keyword k, and $|k|$ denotes the string length of the keyword k. Intuitively, this procedure produces rules for a topic that contain predominately long, important keywords, and removes those rules that may introduce irrelevant domains due to more general or shorter keywords. Again, we use the threshold selection method to set a threshold score for the rules associated with a topic, where all rules above that threshold are retained.

Given the high-quality rules generated for each topic, we convert them to regular expressions by requiring all keywords in a rule to be found as a substring of the domain, and that keywords in bidirectional implications appear in their original ordering. To add a level of flexibility to our regular expressions, we also allow any number of characters to be inserted between keywords. The domains selected by the regular expressions for a given topic undergo one more round of pruning wherein the domain is assigned a score equal to the sum of the scores for each of the rules that matched it. These domain scores are given as input to the threshold selection algorithm, and any domains with scores above the threshold are manually verified to be related to the associated topic. These related domains are herein referred to as the *relevant* set. Table 4 shows an example of the conversion process from search query, to rule sets, and finally to regular expressions used to search our domain registration data.

Results. Our search methodology selected 21,103 distinct domain names related to 116 of the 120 hot topics from the VeriSign dataset. Of these, 15,954 domains in 113 topics were verified to be directly related to the topic at hand (*i.e.,* the relevant set). The percentage of relevant domains per topic, averaged over all topics, is 91%. Overall, these results indicate that the majority of our rules select high-quality domain names, with a small number of topics producing very general rules; often because of unrelated or non-specific Google search queries.

In order to discover the unique properties of the potentially speculated domains that our methodology selected, we examine several features and compare them to those of the background set of domains. First, we look at the distribution of registrations among the name servers and registrars within the background and relevant sets, respectively. In Figure 1, we show a log-scale plot comparing

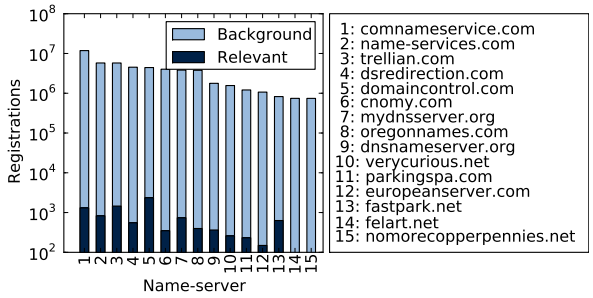


Fig. 1. Name servers in the background set

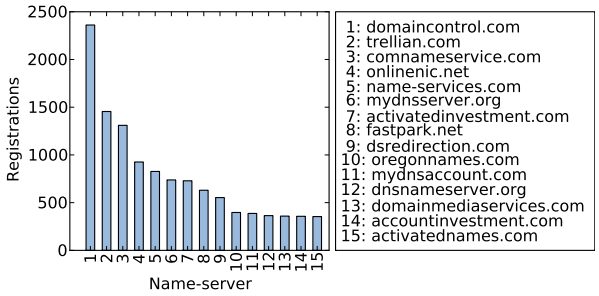


Fig. 2. Name servers in the relevant set.

the background and relevant domain registrations associated with the top fifteen name servers in the background set. For clarity, we also provide the distribution of the top name servers in the relevant domain set in Figure 2. Clearly, the distribution of registrations over these two sets is significantly different as evidenced by the ranking of name servers and the comparison plot in Figure 1. In fact, when we take a closer look at the name servers, we find that the majority of those found in the relevant set are associated with domain parking services, whereas the background set contains a much smaller fraction.

Similarly, we compared the top registrars from the background distribution to those from the relevant set. To characterize the distribution of registrars for background domains, we use the VeriSign monthly reports for the .com top-level domain (TLD) to derive the number of domains registered by the top fifteen registrars over the eight-month period examined in our study. Our analyses reveal that some registrars, such as GoDaddy and eNom, West, maintain their popularity as registrars in both sets. However, there are also some very significant differences. For example, Network Solutions drops precipitously to a rank of ten in the relevant set, and several registrars are found exclusively in the relevant set. These findings indicate that some registrars are clearly preferred by speculators, just as the name servers above were, albeit to a lesser extent.

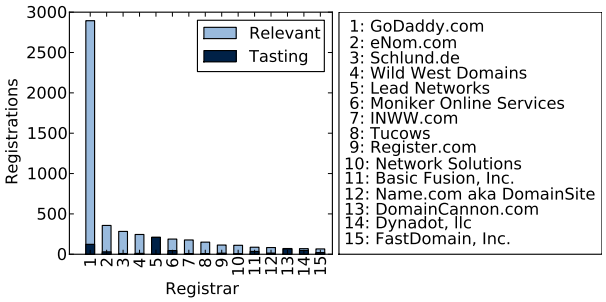


Fig. 3. Registrars in the relevant set.

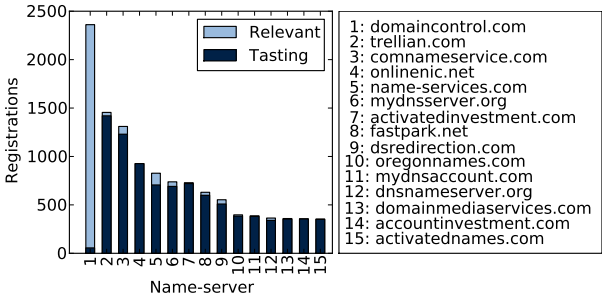
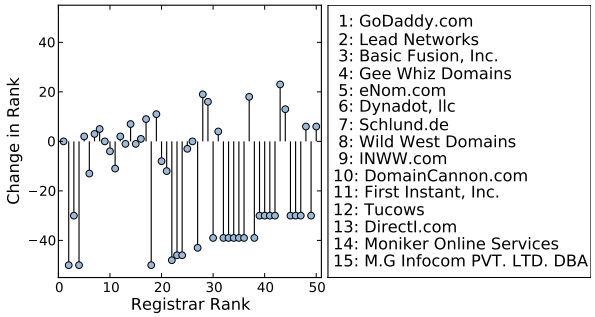


Fig. 4. Name servers in the relevant set

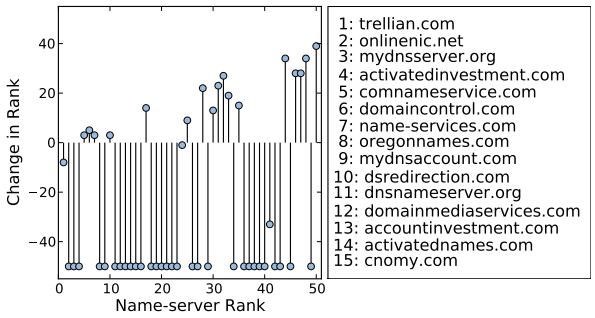
5 Domain Tasting

The second form of questionable domain registration behavior that we study is domain tasting, where a registrar is allowed to delete a domain within five days of the initial registration at no cost, also known as the add grace period. This policy can be easily abused by registrars and registrants alike in order to gain information about the value of a domain via traffic statistics taken during the grace period. To study the prevalence of domain tasting, we select all domain names from the background set of domains that were registered and then deleted within five days, which we refer to as the *tasting* set.

Results. From the full VeriSign dataset, we identified 47,763,141 (76%) distinct registrations as the result of domain tasting, with 10,576 (66%) of those occurring in our relevant set of potentially speculated domains. On average, these tasting domains were registered for 3.4 days before being deleted under the no-cost grace period policy. Figures 3 and 4 shows the comparison of registrars and name servers between all relevant domains and those relevant domains involved in tasting activity. The graphs clearly illustrate that these relevant tasting domains are not strongly connected with particular registrars. However, there appear



(a) Registrar rank in the relevant set.



(b) Name server rank in the relevant set.

Fig. 5. Changes in rank for name servers and registrars in the relevant domain set.

to be clear preferences for certain name servers by the tasters; in some cases representing all registrations for that server!

In June 2008, ICANN made changes to their policies in order to limit the practice of domain tasting. These changes took effect on July 1st, 2008, which positions us perfectly to provide an independent analysis of the impact of this policy change on the tasting of .com domains. For our purposes, we split the dataset into a *pre-reform* period and a *post-reform* period. From our background data, we find 42,467,156 pre-reform tasting registrations with an average duration of 3.4 days, while the post-reform data shows 6,290,177 registrations with an average duration of 3.8 days. For our relevant domains, we have a similar proportion of tasting registrations with 9,270 pre-reform registrations and 1,433 post-reform registrations. These relevant tasting domains were registered for an average of 2.8 and 3.7 days, respectively. In both the background and relevant tasting domains, there is a clear trend toward longer registration periods after the enactment of tasting reform.

Google Query	Association Rule	Domain Name
phillies world series	phillies ⇒ (world ⇔ series)	worldseries.com, phillieswordseries.com, worldseriesblog.com

Table 2. Decomposition of a Google query into rules and domains.

To examine the impact of the reform on the top fifty registrars and name servers in the background and relevant domain sets, we examine their change in rank after implementation of the new tasting policies. Figure 5 shows the change for names servers and registrars associated with the relevant set. Notice the substantial drops in rank for those name servers and registrars occupying the middle ranks (*i.e.,* positions 10-40 in the pre-reform data). Although several of the top-ranked name servers in both the background and relevant sets were predominantly associated with tasting domains, they are able to maintain – or even improve – their rank despite the drop in tasting registrations (*e.g.,* trellian.com).

6 Domain Front Running

Finally, we explore the extent of domain front-running activities among the top domain registrars. To do this successfully, we need to generate relevant (and presumably desirable) domain names for very timely topics, then query domain registrars for the availability of those domains in a manner that simulates widespread interest.

Our approach for generating domain names is similar to that of the rule generation procedure described earlier. We begin by gathering search queries from the top two popularity classifications (*i.e.,* "volcanic" and "on fire") for the current day from Google Trends, and use those searches as transactions in our rule mining process. As before, we set confidence and pruning thresholds for the rule generation for each topic separately using the threshold selection procedure described in Section 3. At the conclusion of this process we have a set of rules for each hot topic for the day.

For each association rule, we create domain names containing the keywords in the bidirectional implications of the rule in the order in which they appear in their original Google search. We then augment this domain name to generate additional names by creating all permutations of it with the keywords in the unidirectional implications. For singleton rules, we use the keyword by itself as the domain name string. Additional domains are generated by appending popular suffixes to the initial domains (*e.g.,* "2009," "blog," "online"). Table 2 provides a concrete example of the domain names generated by our methodology.

The generated domains are divided among the registrars in our study such that no two registrars receive the same domain name. The domains for each registrar are further divided into queried and held-out sets. This division of domains allows us to examine the increase in the rate of registration for those domains that were sent to registrars over those that were not, and pinpoint

the increase in registration rate for certain domains to a particular registrar. Furthermore, in order to ensure that our queries appear to emanate from a diverse set of locations, we make use of the PlanetLab infrastructure to distribute domains for each topic to between two and four randomly selected nodes, which then query the registrars for availability of these domains via the registrars' web site. Lastly, each day we check `Whois` records to determine if any of our queried domains were subsequently registered. In this experiment, we assume that a statistically significant increase in registration rate between queried and held-back domains by a particular registrar is related to front-running activities.

Results. In our study, we issued queries as described above to the nineteen most widely used registrars, accounting for over 80% of the market share according to RegistrarStats.com. Over the period spanning December 1^{st} 2009 to February 1^{st} 2010, we generated 73,149 unique domains, of which 60,264 (82%) were available at the time of generation. Of those available at the time of generation, 16,635 were selected for querying and distributed to the PlanetLab nodes, leaving 43,629 domains in the held-back set. A total of 23 of the queried and 50 of the held-back domains were registered during this period.

To examine the significance of our results, we perform statistical hypothesis tests for each of the registrars in isolation. Specifically, we model the rate of registration in both the queried and held-back case as a binomial distribution with probability of success equal to the unknown rate of registration. The Fisher-Irwin exact test is applied instead of the standard z-test since it avoids approximation by a normal distribution and explicitly calculates the probabilities for the two binomials given the numbers of queried and held-out domains. Our analysis indicates that none of the registrars are associated with a statistically significant ($p < 0.05$) increase in the registration rate of queried domain names.

7 Summary

In what follows, we discuss the implications from our empirical analyses, and examine the relationship among tasting, front-running, and speculation activities.

On the Quality of the Generated Rules. Based on the results of our speculation and front-running experiments, we argue that the rule mining and threshold selection methodologies worked surprisingly well given such noisy data. For our speculation experiments, we found that an average of 91% of the selected domains were related to their respective popular topics, and many of our automatically generated domains were indeed registered. For those rules that generated non-relevant domains, the primary cause can be attributed to incoherence in the related search terms provided by Google. Nonetheless, we believe that our techniques show significant promise in taking unstructured keywords and returning general rules that can be applied to a variety of problems.

Incentives for Misbehavior. A natural question that arises when considering these abusive domain registration behaviors is: what are the incentives that drive them? To begin to answer this question, we performed a cursory analysis of the contents of potentially speculated domains selected by our methodology, along with an examination of potential ad and resale revenue associated with the topics in our study. Most of the domains that we examined contained significant pay-per-click ad content, and our analysis showed that many of these sites were hosted by known domain parking firms. Based on data gathered from Google's AdWords Traffic Estimator, we found that the average cost-per-click for the topics in our study was $0.76 per click, and many of these topics have expected click rates in the 300-400 clicks per day range. Beyond ad revenue, some domain names associated with the topics we studied were resold for an average price of $1,832, with the largest of these being $15,500 for `obama.net`.

Clearly, there is significant financial incentive to both resell popular domains and to use parking services to generate advertising revenue. In fact, as long as the average revenue among the domains owned by the speculator exceeds the hosting and registrations costs, the speculator is better off retaining as many domains as possible and only serving ad content. As a concrete example, we note that the keywords associated with the automatically generated domains from our front-running study would have produced revenue in excess of $400 per day, while domain parking services can be purchased for as little as $3.99 per domain each month. This represents a net profit of approximately $11,700 per month from ad revenue alone for the 73 registered domains in our front-running study! Furthermore, the strong connection between domain popularity and revenue provides insights into the use of tasting and front-running behaviors as a mechanism for determining the true market value for domains without having to invest capital.

Difficulty of Measurement. Another surprising lesson learned from our study is that many of these questionable registration behaviors are particularly difficult to definitively measure. In the case of speculation, we attempted to use several metrics to distinguish those domains registered due to speculation from those registered for legitimate use, including the length of registration, the timeliness of the registration after the increase in search popularity, the rate at which hosting changes, and manual inspection of web page content. Of these, only inspection of the content yielded any significant results, and even in this case there were several instances where it was difficult to identify the true purpose of the page (*i.e.,* to deliver legitimate content, or to serve ads). In our cursory examination, 60% of sites redirected users to parking web pages that contain only ad content, while in the remaining cases the pages contained a non-trivial number of ads in addition to seemingly legitimate content.

In regard to front-running, while we found no statistically significant evidence of misbehavior by registrars, during the course of this study we uncovered the fact that many registrars have several subsidiaries that also performed registration duties on their behalf. The connections among these entities are exceedingly difficult to discover and, unfortunately, little information exists in the public

domain that can be used to confirm them. Therefore, if some registrars were involved in front-running behaviors, it is entirely possible that they could hide questionable activities by routing registrations through subsidiaries or partners. As a whole, these results call into question overhasty statements by ICANN that front running is not occurring. Moreover, from what we can tell, these relationships frequently change, underscoring the difficulty in detecting misbehavior by dishonest registrars.

Regarding ICANN Policy. While it is clear that policy changes instituted by ICANN have had an appreciable impact on the practice of domain tasting, reforms aimed at curtailing speculation and front-running appear to be nonexistent. One obvious, if drastic, solution would be to eliminate the conflict of interest that arises when registrars are allowed to sell domain names. Other potentially effective approaches, including offline domain availability checks, have also been put forth. However, these approaches have all been rejected outright by ICANN, even after seemingly acknowledging the threat of domain speculation as the reason for postponing any new applications for generic top-level domains (gTLDs) [7]. At the very least, we hope that our results shed light on the challenges inherent in detecting such malfeasance, and that they will spur constructive dialog on relevant public policy.

Acknowledgements This work was supported in part by the U.S. Department of Homeland Security under Contract No. FA8750-08-2-0147, and the National Science Foundation under award numbers 0831245 and 0937060.

References

1. R. Agrawal, T. Imieliński, and A. Swami. Mining Association Rules Between Sets of Items in Large Databases. In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, May 1993.
2. D. Arthur and S. Vassilvitskii. k-Means++: The Advantages of Careful Seeding. In *Proceedings of the* 18^{th} *Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035, January 2007.
3. A. Banerjee, D. Barman, M. Faloutsos, and L. N. Bhuyan. Cyber-Fraud is One Typo Away. In *Proceedings of the* 27^{th} *Conference on Computer Communications*, pages 1939–1947, 2008.
4. T. Holgers, D. E. Watson, and S. D. Gribble. Cutting Through the Confusion: A Measurement Study of Homograph Attacks. In *Proceedings of the* 24^{rd} *Annual USENIX Technical Conference*, pages 261–266, 2006.
5. D. Kesmodel. *The Domain Game: How People Get Rich From Internet Domain Names*. Xlibris Corporation, 2008.
6. T. Moore and B. Edelman. Measuring the Perpetrators and Funders of Typosquatting. In *Proceedings of Financial Cryptography and Data Security*, 2010.
7. Michael Palage. New gTLDs: Let the Gaming Begin. Part I: TLD Front Running. In *The Progress & Freedom Foundation*, volume 16, Aug. 2009.
8. Y. Wang, D. Beck, J. Wang, C. Verbowski, and B. Daniels. Strider Typo-Patrol: Discovery and Analysis of Systematic Typo-Squatting. In *Proceedings of USENIX SRUTI*, pages 31–36, 2006.