Paper

Who is Asking and for What: WHOIS Traffic Analysis

Mariusz Kamola

Research and Academic Computer Network (NASK), Warsaw, Poland
Institute of Control and Computation Engineering, Warsaw University of Technology, Warsaw, Poland

Abstract—The paper presents analysis of WHOIS requests for 13-month period. Both requestor address and the domain name being requested are analyzed, showing that WHOIS traffic can be roughly classified into systematic scanning of domain names and individual low-volume activity, mostly targeting very popular names. The comparison of requested names with standard dictionary entries reveals typical mutations for registered names, and mutations performed by scanning automata. As most popular names in WHOIS coincide with standard top website ranks, the ways of utilizing WHOIS data for the benefit of Internet community as a whole, are proposed.

Keywords*—*request map, semantic analysis, SEO, WHOIS.

1. Introduction

The term WHOIS refers in broad sense to a protocol [1] designed to query personal details related with various entities found in today's Internet. In this paper we will deal with WHOIS in more narrow and well known sense – as the technology to retrieve Internet domain name registrant's data. These data are made available to the public by appropriate servers maintaining registration databases. There are three main ways to access the registration data: a HTTP interface, a service operating WHOIS protocol on port 43, and bulk datasets obtainable from the registry. Due to the fact that the data may contain personal information as e-mail, phone number and even street address, there have been always discussions in ICANN about privacy issues, and the conflict between data openness in Internet community and privacy law imposed by local governments [2].

The issue is an important one because WHOIS data as such can serve as huge, effective, and legal directory for spamming, hacking and other socially undesirable behaviors. Tackling the matter, ICANN has come up with a series of requirements and recommendations for registries, aiming at preventing misuse of the data. Web access has been mostly equipped with CAPTCHA technology and port 43 service with rate limitations to prevent massive and automated database scanning. Such scanning is still possible on bulk data, under declaration that the results will *not* be used for marketing and alike (cf. eg. [3]). ICANN is monitoring the issues with WHOIS as DNS is evolving; see the relevant memorandum on the occasion of gTLD (Generic Top-Level Domain) release [4] where minimum set of registrant information in different domain classes has been specified; also prior related regulatory activities are mentioned therein.

Naturally, despite regulatory efforts, business wants to make money from the valuable WHOIS information – and the retail requests generated by serious or curious individuals mix with regular database scanning performed by companies. Such is the major outcome of the cursory study on WHOIS requests to NASK register. The major motivation for such a study was to gain insight into how actually the database is used, by whom and, if possible, for what purpose. Investigating business models underlying WHOIS requests made by companies thriving on added Internet services has been considered particularly important. This is not because NASK is going to compete with them; on the contrary, being a supervisor of a large part of Polish web activities, NASK is going to consider utilization of those data to stimulate healthy growth of Internet community in the country – also through educational activities, backed by sound research results reported in scientific papers.

This paper is organized as follows. Section 2 presents the data that have been worked on along with the computing equipment. Then author focuses on the part which is, in his opinion, not present in the literature: to identify requestors of WHOIS data. The sole purpose of such an investigation is to classify WHOIS users into categories and, possibly, into subcategories, based on the traffic generated by them. In Section 3 attention is shifted to the domain names requested. Their similarity and temporal pattern of requests will be examined, focusing on key commercial requestors. Section 4 comes with conclusions, possible exploitation of results and planned future work description.

2. Who is Asking

The data being subject to analysis were 180 million WHOIS requests recorded in 13 months since September 2009, stored in WHOIS database in NASK. The register covers `.pl` domain, as well as some other functional and regional domains (e.g. `.gov.pl`, `.edu.pl`, `.poznan.pl`). The selected period is long enough to get rid of any kind of seasonality if one operates on averages. However, it must be emphasized that the stable volume growth biases the results, giving more weight to latter data, cf. Fig. 1.

This section covers the analysis of the *source* of incoming requests, i.e. the IP address of the requester. From now on we will operate on IP addresses with its last byte canceled. This has been done for two reasons. The first is privacy.

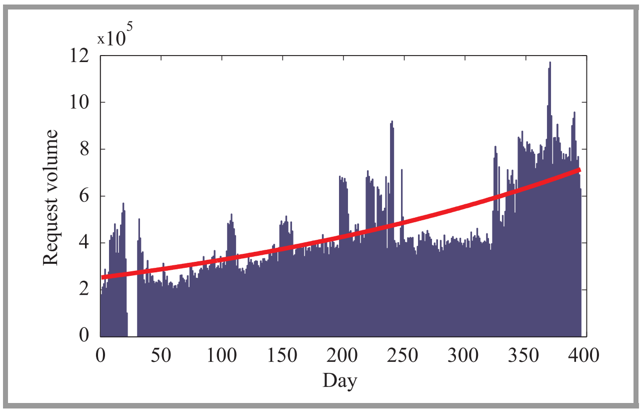

Fig. 1. Request volumes in subsequent days versus fitted exponential growth curve.

The second is that we suspect some organizations to use a pool of addresses while scanning WHOIS – and we did not wish to partition such activities artificially. Removal of the last byte reduces the number of unique requestor addresses from the original 3.2 million by two thirds.

The ranking of activity for 1,000 most active source addresses is presented in Fig. 2. Considering logarithmic scale, one can easily notice that the 30 most active ones account for more than 50 percent of the total requests. The first client in the ranking generates as much as 8 percent of the total traffic. Such an activity is clearly a kind of machine to machine communication. Next 100 most active requestors are also too active to be individuals; they may be commercial organizations, as well as gateways of networks with NAT – as, e.g. mobile operators. The activity of smaller requestors follows almost ideally the power law, like in social network node degrees, personal income and many natural phenomena.

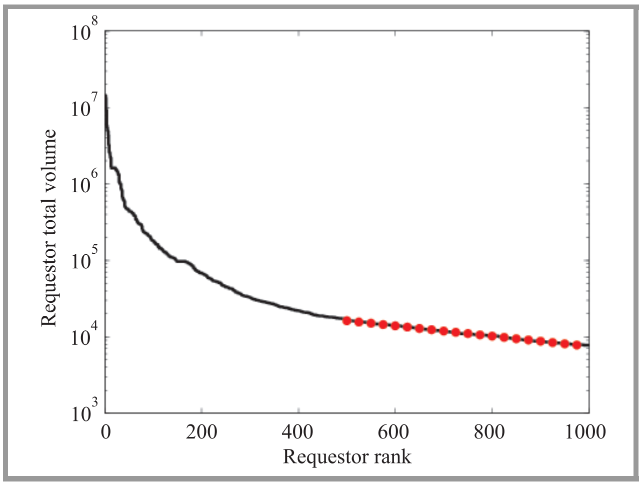

Fig. 2. Total number of requests for requestors ranked most active. Power law scaling for smaller clients shown as dotted line.

The most active in the ranking is one company providing Internet-related services as domain registration web and mail hosting, etc. It is located in one of the major Polish cities. Its traffic is generated by twelve equally loaded physical IP addresses. After stable growth in the first half of the analyzed period, its daily volume of request has stabilized around 50,000 a day. This number has not been affected at all by the rapid and stable growth of total WHOIS requests in August 2010, cf. Fig. 1.

Looking for similarities in behaviour of top requestors, let us have a look at activity of the second, third and seventh ones in the ranking (7th biggest requestor is included because of its unique location). Histograms of their daily request volumes have been compared to the overall traffic in Fig. 3. The first thing that “1” and “3” (also “2” to some extent) have in common is they exhibit quite precise limit of requests per day. If there are fluctuations in rates, they are drops, frequently reaching zero. On the contrary, the total traffic shows quite many peaks above the average. The differences in distributions are based in traffic trends, not shown here. All requestors *do not* exhibit any regular request growth in longer (e.g. monthly) window throughout the whole timespan. As regards “7”, it resembles the total trend the most, but it does not grow either. Moreover, it switches off just the moment the total number of requests grows rapidly (August 2010).

Let us now study geographical location of request origins. The results may give an idea for whom it could be valuable to have a Polish domain name – or interested in discovering who owns a name. A free geographical location of IP addresses service [5] was used for this purpose. The lookup gives the two-letter country code and a city name. It must be noted that the lookup is not reliable as it could not find the address location in 24 percent of cases. Also, the geographical location obtained is sometimes confusing, e.g. the third largest requestor, obviously registered as Polish company, the location found was Turkey. However, despite of imperfections, the request world map shown in Fig. 4 looks reasonable. Poland is evidently the leader, and the other major requesting countries are either big (China), geographically close (Czech Republic) or with big Polish diaspora (Brazil). Or all of them, as for France, Great Britain and USA. Unfortunately, the map rendering mechanism [6] is not flawless, taking USA state names for the names of the countries, and not displaying requests from Germany and Netherlands, the fifth and eighth in the ranking, respectively.

In this map of interest there is an absence of big countries: Spain, Ukraine, Greece, Lithuania – with their obvious links with Poland. To explain such disinterest and its correlation with other factors, e.g. the amount of foreign investments is, however, beyond the scope of this paper. On the other hand, some countries exhibit surprisingly high rate of requests, as Denmark and, most of all, Iran. Considerably high interest of Polish southern Czech and Slovak neighbors explains easily as they are already actively participating in Polish retain Internet trade and services.

WHOIS requests generated regionally, i.e., in Poland, have also been analysed w.r.t. its origin. Biggest requesting lo-

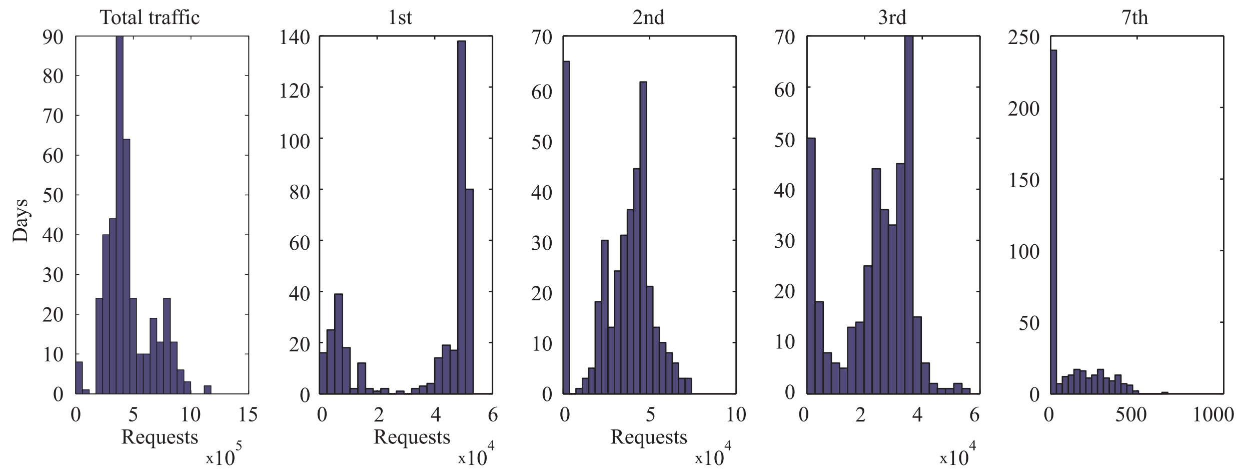

Fig. 3. Histograms of daily activity of all users, the three top requestors and the 7th requestor in the ranking.

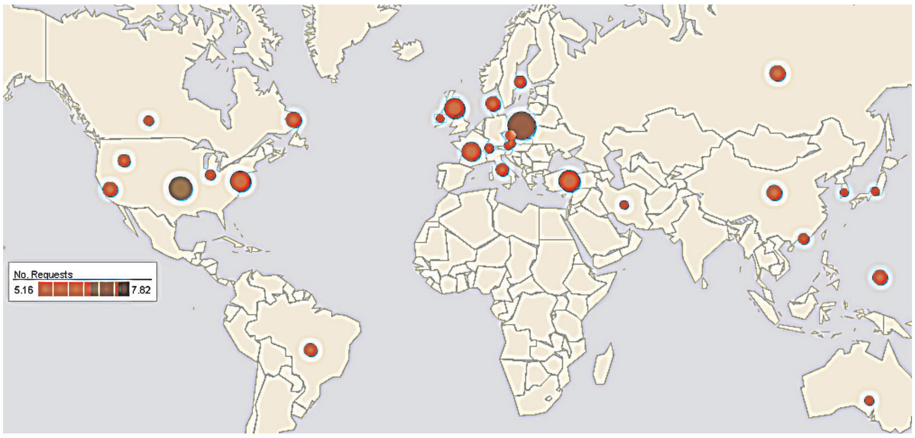

Fig. 4. Request world map.

cations are shown in Fig. 5, vs. the population of those locations. In general, the biggest sources of traffic lie in the biggest cities – which also means that the IP location database in Poland is more accurate than the global one. This rule is represented in the figure by points clustered around the straight line. Note that there appear cities rather too little active for their size ("15"), which may mean that IT there need more development. But much bigger disproportions can be seen on the other side of the straight line: there is a small town ("17") and a village ("7") generating more requests than a half-million metropolis. It is difficult to expect that the latter one is just an ordinary case. A closer look at 7's traffic history reveals that it started all of a sudden in February 2010, which mean of 1,600 requests per day and standard deviation as big as 1,500. The traffic experienced from time to pauses of several days time, and it did not show any growth trend – like for the biggest requestor, and unlike in global statistics.

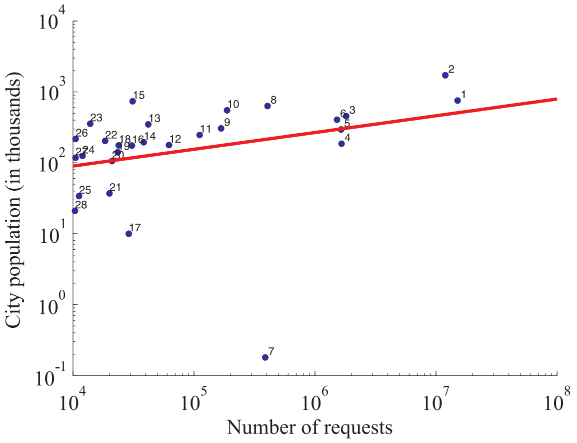

Fig. 5. Top Polish requesting cities versus their population. Mark numbers represent city rank w.r.t. generated traffic. Linear regression in log-by-log domain is shown by the red line.

Summing up, we may observe that the biggest traffic is generated by commercial scanners. These are located mainly in big cities, with few exceptions. Commercial requestors usually maintain their daily rate of requests, insensitive to overall WHOIS traffic growth. They use a single address or a pool of addresses, thus distributing their machine loads.

3. Requested names analysis

To get a reference point, let us confront our top 20 requested WHOIS names with top 20 visited `.pl` addresses by Alexa [7] as shown in Table 1 . Although Alexa ranking was done a year after the WHOIS data end, one can still see that entries in both columns are similar, especially for the top 10. Therefore WHOIS statistics may serve as a decent measure of company or website popularity – at least this applies for big fish. This interest is risen by individuals who, certainly, are not going to buy such domain names; it is rather curiosity that drives users to check extra info about companies. Further entries are not so well matched: those WHOIS names that are not present in Alexa top 20 are given in italics. Such discrepancies may be due to the lag of Alexa ranking, but it may be also different kind of interest driving users to WHOIS and to the wepage itself. Take for example `platformaobywatelska.org.pl` which is the political party governing right now: it is not shown in Alexa top 20[1] but appears for WHOIS. The reason could be that requestor is interested in who is *personally* involved in domain registration, which amounts to looking for reliable extra information about the party.

Table 1
Polish WHOIS vs. Alexa: top 20

	WHOIS	Alexa
1	google.pl	google.pl
2	onet.pl	onet.pl
3	wp.pl	allegro.pl
4	nk.pl	wp.pl
5	allegro.pl	gazeta.pl
6	nasza-klasa.pl	interia.pl
7	gazeta.pl	nk.pl
8	interia.pl	mbank.com.pl
9	home.pl	o2.pl
10	*test.pl*	pudelek.pl
11	o2.pl	sport.pl
12	demotywatory.pl	otomoto.pl
13	*tpnet.pl*	goldenline.pl
14	*wrzuta.pl*	kwejk.pl
15	*platformaobywatelska.org.pl*	demotywatory.pl
16	*nazwa.pl*	ceneo.pl
17	pudelek.pl	home.pl
18	*peb.pl*	tvn24.pl
19	*blox.pl*	filmweb.pl
20	*ropa.pl*	chomikuj.pl

However, a domain name in WHOIS estimates interest in smaller companies, ideas or activities as well. Take for example `tiny.pl`, which does not appear in the above list, but is second last frequently asked domain on the 30th Sept. 2010, i.e., the last day of the period analyzed. The service accomplishes domain name abbreviation, like many other ones – and with similar business model behind. Figure 6 illustrates the rapid growth of interest in the name, preceded by a long period of rather poor interest (the name was registered as early as 2004). About 50% of those requests have been made from unique IP addresses, which means in this specific case (as well as for more popular names) most of the traffic is generated by curious individuals rather than commercial scanners.

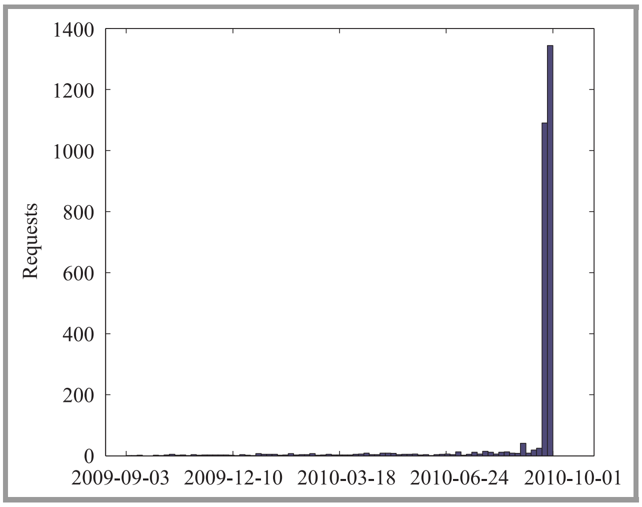

Fig. 6. Number of WHOIS queries for `tiny.pl` name, grouped in 5-day periods.

Such general curiosity about big companies never shows for big requestors. The objects of interest of big requestors are (in the order of their frequency):

- names of objects (not being trademarks or proper names): *shawl*, *office*, *room*, etc.,
- expressions (with words mostly being written together): *new photography*, *stairs from Poland*, *property valuation*, etc.,
- the above names, but located in functional or regional `.pl` subdomains,
- proper names and trademarks.

It is worth noting that verbs in names are rare and, contrary to widespread opinion, names referring to sexuality are few.

It is interesting to examine how requested domain names are related to words commonly used in written or spoken language. In particular:

- which dictionary words are most popular as domain names,
- which dictionary words are not interesting, and why.

The problem is what should be considered the reference dictionary. Language corpora contain lots of words and their variations according to specific language grammar

[1] Municipal and presidential elections in 2011 and parliament election in 2012 were equally good reasons for `platformaobywatelska.org.pl` to appear in Alexa top 20 ranking, however, such domain name did not appear there altogether.

and morphology that are not appealing as domain names (e.g., non-infinitive forms of verbs) or anachronisms. To make our reference dictionary a contemporary one, we decided to use frequency list for Polish words as offered by `wiktionary.org` [8]. It contains only 10,000 Polish words, and probably other alternative sources of data [9] could have turned out to be more useful and suitable.

To measure word popularity, we have to take into account not only exact requests for this word, but also similar requests. The most widespread technology for fuzzy word matching, used also in spell checking and correction, is based on Levenshtein metric [10]: the minimum number of original word elementary transformations that make a dictionary word out of it. Elementary transformations come in four kinds:

– *deletes* – removal of a single letter at position i,

– *transposes* – swap of letters at positions i and $i+1$,

– *replaces* – replacement of a letter at position i with another letter,

– *inserts* – insertion of a single letter at position i (shifting letter at $i+1$ one position behind and so on).

Let $A = 'a', 'b', \ldots, 'z'$ be the alphabet and w be a word $w = (w_1, \ldots, w_n)$ where $w_{i \in 1..n} \in A$. Let us denote the set of all words being result of a single transformation of w by $U^{(1)}(w)$. Let $D = d_1, .., d_m$ be the dictionary of words and $F = f_{w_1}, \ldots, f_{w_m}$, $f \in \Re$ their respective frequencies. Then the distance-one fuzzy word matching procedure chooses

$$M^{(1)}(w) = \arg \max_{q \in U^{(1)}(w) \cap D} f_{w_q},$$

i.e. the most frequent dictionary entry reachable by one modification of w. Analogously,

$$M^{(2)}(w) = \arg \max_{q \in U^{(2)}(w) \cap D} f_{w_q},$$

is the distance-two matching procedure, where

$$U^{(2)}(w) = \bigcup_{z_i \; U^{(1)}(w)} U^{(1)}(z_i)$$

defines the set of words available by two elementary modifications to w. The matching routine proposed in [11] combines distance-zero, distance-one and distance-two algorithms, returning the best match regardless of the number of modifications (0, 1 or 2, respectively) needed.

The above routine has been employed, with further modifications, for matching requested domain names with dictionary entries. The needed modifications are requested domain name preprocessing rules, before the actual matching takes place:

- Cutting domain name to 15 initial characters. This is because matching algorithm performance decreases rapidly for longer words.
- Removing the domain name part after first dot. This makes `.pl` itself and its all functional and geographical subdomains equally important, being the simplest way to detect interest in names registered in subdomains.
- Removing all dashes in names as irrelevant.
- Replacing all numerical substrings with single ‘0’ (zero) character. This way, popular numerical pre- or suffixes to names are easily detectable.

In practice, precise and reliable operation of the above matching approach increases with dictionary word length. Obviously, distance-two modifications of short words give plenty of other dictionary words, obscuring the conclusions. Otherwise, for longer (e.w. 10-letter) words the algorithm detects most of the word’s variations that are requested. For instance, the most popular (by means of related WHOIS requests) 10-letter dictionary word, ‘fotografia’ (*photography*), has 619 distance-two similar domain names. The most popular ones have been listed in Table 2. Even for readers not familiar with Polish language all those names appear as loose variations of the basic term ‘photography’: they point to websites with similar functionality, too. Interestingly enough, the basic form, ‘fotografia’ appears only on the 5th position. These variations are not created by adding random infixes; they all are meaningful: diminutive, imperative, plural, bearing typical suffixes for village names, indicating continuous service. This is also an indication that those domain names are registered and alive.

Table 2
Requests similar to dictionary entry ‘fotografia’ (*photography*), in order of their frequency in WHOIS requests

fotografuj.pl
fotografijka.pl
fotogratis.pl
fotografika.pl
fotografia.pl
fotograma.pl
fotografie.pl
ek-fotografia.pl
fotografow.pl
foto-grafi.pl
fotografie.org.pl
fotografia24.pl

If we rank popularity of 10-letter or longer dictionary words referred to in WHOIS requests, it will be as in Table 3. The number of domain names similar to a dictionary word does not apparently depend on the word rank – but it is always substantial (at least 62, which means that those domain names are valuable). The average Levenshtein distance can be as small as 0.57, meaning that regional or

Table 3
Top 15 dictionary words,with their number of relevant WHOIS requests, number of distance-two similar domain names, average Levenshtein distance and actual Alexa Rank value

Dictionary word	English translation	Number of requests	No. of similar domains	Average L. distance	Alexa Rank
fotografia	photography	21090	619	1.0388	17,298,998
apartament	suite	19797	465	1.4710	0
akademicki	academic	17202	68	1.1912	0
gospodarka	economy	15821	79	1.7595	9,578,089
nieruchomość	real estate	15303	137	1.8978	1,089,090
certyfikat	certificate	10906	239	1.1172	0
biblioteka	library	9251	334	0.6317	7,299,153
elektronika	electronics	8855	261	1.1533	0
bezpieczeństwo	safety	7442	169	0.6331	0
dziewczyna	girl	7438	178	1.4663	12,629
autostrada	motorway	7295	144	1.6181	897,172
elektroniczny	electronic	7269	243	1.6626	0
architektura	architecture	7247	285	1.4421	3,966,672
administracja	administration	5576	156	1.1731	14,118,094
budownictwo	construction	5474	154	0.5714	97,740
encyklopedia	encyclopedia	5197	156	0.9423	175
dominikana	Dominican Republic	5093	62	1.2581	12,207,332
astrologia	astrology	5081	143	1.4336	1,147,848
administrator	administrator	4557	149	1.3624	0
elektrownia	power plant	4355	93	1.3656	0

functional domains are preferred than variations of the base name (cf. *library, safety, construction*). On the other end there are domains with big distance: *real estate, economy, electronic*, i.e. presumably denoting services with national range.

The last column of Table 3 gives *current* Alexa Rank, made available by one of SEO (search engine optimization) services [12]. The first observation is that this rank is much incomplete, missing highly interesting domains (other metrics: Page Rank and link popularity provide even more sparse data). The second is that Alexa Rank is quite inadequate to our rank of the dictionary word. The probable reasons are:

- comparison ignores the 2 years that passed since the end last analyzed data,
- bigger number of name alternatives decreases value of the name `dictionary_word.pl` itself.

Contrary to ‘fotografia’ (*photography*) keyword case, there are quite interesting domain names that apparently are targeted by scanners. Taking, for example the word ‘wyjazd’ (*trip*), confronting Table 4, we can see that the names requested are mutations of the base word. Mutation operations include swapping and doubling of pairs of letters, thus following common typographical errors made while entering the domain name. Therefore aim of the activity could be finally to register names similar to existing ones to intercept http requests containing typos and, for example, redirect them to competitive websites. Alternatively, the domain owner may resell the name to the owner of the “correct” domain name. Definitely, none of healthy competitors of the `wyjazd.pl` owner would like to run her business under a name containing a typo. Regarding the time pattern, we see that requests are made at equal 7-second intervals. Discontinuities of this schedule are due to the fact that some of the requests names had too big Levenshtein

Table 4
Scanning activity for the name `wyjazd` performed in a single day from a single IP address

Time	Name
16:07:37	e–wyjazd.pl
19:17:07	e-wjazd.pl
19:17:14	e-wjyazd.pl
19:19:15	e-wwyjazd.pl
19:19:29	e-wyajzd.pl
19:19:43	e-wyjaazd.pl
19:20:10	e-wyjazdd.pl
19:20:17	e-wyjazzd.pl
19:20:23	e-wyjjazd.pl
19:20:30	e-wyjzad.pl
19:20:37	e-wyjzd.pl
19:22:26	e-wyyjazd.pl
19:31:35	e-ywjazd.pl
23:13:41	ee-wyjazd.pl
00:32:23	eewyjazd.pl

distance to ‘wyjazd’ (*trip*) to be detected by our analytic software. Taking into account the geographical location of the requestor, it is located abroad and makes also regular requests for the proper domain name from neighboring IP addresses, but in much longer timescale.

If we consider least frequently queried names that are similar to dictionary entries, we will find that there are surprisingly many being the vocabulary entry itself, with no mutations. For example, in a reverse popularity ranking complementary to Table 3 the first mutation occurs only on the 95th position. Out of 30 first dictionary words least used, 15 are nouns (*vaccine*, *agreement*), 8 are adjectives (*southern*, *fixed*) and 5 are verbs (*apply*, *happen*). The words given in parentheses, which are examples of domain names queried only once, seem not to be uncommon at all – yet they did not gain much popularity. Therefore there is an measurable reason for them to gain popularity (high frequency in dictionary), and at the same time there exist technical possibility to achieve that (registering word mutations as domain names). Making such statistics available to the public may stimulate further growth of registered domain names.

4. Conclusion

Analysis of the source of WHOIS requests and their content prove that at systematic domain names scanning activity is commonplace, and that it has a considerable share of the overall WHOIS traffic. Certainly, there must be the business case for that: it may be the detection of unregistered and attractive domains, monitoring of availability of the popular names for registration, retrieving e-mail addresses to send commercial offers or any other reason. Explanation of reasons of scanning would require joint analysis of the domain registration and name querying processes, which lies outside of the scope of this paper. Such analysis would be definitely interesting and worth effort – but, most of all, it needs to have a well defined social purpose.

On the other side, we observe big volume of requests for very popular domains that correlate well with Alexa ranking (cf. Table 1). Therefore we can spot at least two basic classes of requestors: commercial scanners and private users. As for the latter, we may suppose they place requests out of curiosity for a company that is behind a domain name: its real name, location, entry date. In this regard, a WHOIS record is equal to a economic press release, or better, an official register of companies.

Regardless of requestors’ motives, WHOIS activity for a domain name can be perceived as reliable metric of the domain importance and – maybe – its true value. WHOIS statistics, when used skillfully, may contribute to overall domain market growth. We believe that such growth is good for country’s economy as such – regardless of the benefit of companies already profiting from domain registration or trade processes. Having big number of domains means that Internet users appreciate their Internet identity and – since DNS itself should not be commercial – their freedom. It is only that such identity and freedom should come at adequate and affordable prices. So, we can exploit WHOIS statistic in two ways: 1) to sanitize names trade and 2) to suggest unused domain names that are meaningful and otherwise valuable. As regards the first idea, comparing prices of domains sold at stock (cf. Table 5, with data retrieved from [13]) we see that often quoted prices are strikingly inadequate to the number of registered WHOIS requests. Examples of such names are: `licencje24.pl` (*licences24*), `green-age.pl` or `we-love.pl`[2]. As the mentioned names do not denote trademarks but common terms, their rank should not be affected by 2-year time difference in dates of WHOIS statistic recording and domain stock exchange query. Anyhow, their prices seem to be far exaggerated; we hope that making this sort of comparison publicly available may restore some order and sanitize prices[3]. Suggesting unused domain names for sale is considered now an idea for further discussion.

Table 5
List of currently most valuable names for sale vs. historical WHOIS number of requests

Domain name	Stock quote (PLN)	No. of WHOIS requests
msza.pl	200,000.00	91
jedwab.pl	130,000.00	81
icrm.pl	100,000.00	58
cov.pl	100,000.00	213
ddw.pl	92,000.00	279
goracezrodla.pl	60,000.00	61
pc.com.pl	50,000.00	216
dobrarobota.pl	50,000.00	62
e-kontakt.pl	40,000.00	148
licencje24.pl	27,000.00	7
najwiekszy.pl	24,000.00	43
e-sprzedawca.pl	10,000.00	52
forumeo.pl	10,000.00	53
sciag.pl	10,000.00	77
green-age.pl	10,000.00	0
sondeo.pl	10,000.00	22
highspeed.pl	10,000.00	57
zyjmyzpasja.pl	10,000.00	0
we-love.pl	10,000.00	3
naprawa-serwis.pl	10,000.00	27

Utilizing WHOIS as reliable register of domain values has been addressed already in a patent [14] in a complementary context: the author proposed to enrich WHOIS data with a record describing its value, computed on the base on a number of external metrics. The article proves that internal WHOIS statistics themselves can also be considered

[2] Names that appear to be registered trademarks and those registered with Polish diacritic signs are excluded from analysis because they might not exists at time when WHOIS samples were registered.

[3] Obviously, such action may result in fake requests made in order to inflate domain popularity; we believe such activity can be filtered out easily.

a good estimate of domain true value. Both approaches share the same point of view that WHOIS database is currently underutilized as a source of reliable information on domain names. It can be improved that without getting involved in privacy issues.

Such preliminary analysis opens way for future activities. The most needed is repeating the research for current WHOIS records to avoid lags between WHOIS and external data like domain market pricing, cf. Table 5. The most powerful of them and the most complicated at the same time is joint analysis of WHOIS requests and domain registration process. Substantial part of temporal [15] and semantic [16] analysis of DNS registration have been already performed in NASK, the latter one focusing on malbehavior detection. Both tasks are strongly related to the approaches presented here; also, the three publications constitute a strong basis for future joint study on WHOIS requests and domain registration, in economic and safety aspects. Other useful directions are providing requestor classification criteria, mastering the algorithm for word match and promoting registration of domains related to popular dictionary entries.

References

[1] L. Daigle, "WHOIS Protocol Specification", RFC Draft Standard, 2004 [Online]. Available: http://tools.ietf.org/html/rfc3912

[2] M. Mueller and M. Chango, "Disrupting global governance: the Internet whois service, ICANN, and privacy", *J. Inform. Technol. Polit.*, vol. 2, no. 3, 2008.

[3] "Obtaining Bulk Whois Data from ARIN", 2012 [Online]. Available: https://www.arin.net/resources/request/bulkwhois.html

[4] "Thick vs. Thin Whois for New gTLDs", ICANN Memorandum, 30 May 2009 [Online]. Available: http://archive.icann.org/en/topics/new-gtlds/thick-thin-whois-30may09-en.pdf

[5] "My Address Lookup and Geo Targeting", 2012 [Online]. Available: https://www.hostip.info

[6] "Visualization: Geomap – Google Chart Tools", 2012 [Online]. Available: https://developers.google.com/chart/interactive/docs/gallery/geomap

[7] "Alexa Internet – Our Technology", 2012 [Online]. Available: http://www.alexa.com/company/technology

[8] "Listy frekwencyjne", 2012 [Online]. Available: http://pl.wiktionary.org/wiki/Kategoria:Listy_frekwencyjne

[9] "Frequency Word Lists", 2012 [Online]. Available: http://invokeit.wordpress.com/frequency-word-lists/

[10] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals", *Soviet Phys. Doklady*, vol. 10, pp. 707–710, 1966.

[11] "How to Write a Spelling Corrector", 2012 [Online]. Available: http://www.norvig.com/spell-correct.html

[12] "Alexa Rank", 2012 [Online]. Available: http://seotracker.pl/alexa-rank.html

[13] "Domain Stock Exchange", 2012 [Online]. Available: http://www.nazwa.pl/gielda-domen/najciekawsze-domeny/top-lista-domen

[14] "Publishing Domain Name Related Reputation in WHOIS Records", US Patent Application Publication No. 2006/0095459, 4th May 2006.

[15] P. Arabas, P. Jaskóła, M. Kamola, and M. Karpowicz, "Analysis and modeling of domain registration process", *J. Telecom. Inform. Technol.*, no. 2, pp. 63–73, 2012.

[16] K. Lasota and A. Kozakiewicz, "Analysis of the similarities in malicious DNS domain names", in *Proc. Sec. Trust Comput., Data Manag. Appl. STA 2011*, Loutraki, Greece, 2011, Springer, vol. 187, pp. 1–6.

Mariusz Kamola received his Ph.D. in Computer Science from the Warsaw University of Technology in 2004. Currently, he is an Associate Professor at Institute of Control and Computation Engineering at the Warsaw University of Technology. Since 2002 with Research and Academic Computer Network (NASK). His research area focuses on technologies supporting modern IP network design, monitoring and management, social networks and other complex systems.

E-mail: Mariusz.Kamola@nask.pl
Research Academic Computer Network (NASK)
Wąwozowa st 18
02-796 Warsaw, Poland

E-mail: mkamola@ia.pw.edu.pl
Institute of Control and Computation Engineering
Warsaw University of Technology
Nowowiejska st 15/19
00-665 Warsaw, Poland