# UNVEILING INTERESTS AND TRENDS USING THE DNS

Luca Deri, Lorenzo Luconi Trombacchi, Maurizio Martinelli
IIT-CNR, Pisa, Italy
{ luca.deri, lorenzo.luconi, maurizio.martinelli}@iit.cnr.it

**ABSTRACT**

The domain name system (DNS) is a distributed database system that allows numeric IP addresses used in the Internet protocol suite to be associated with human-readable names. Often perceived as a hidden legacy service on which the Internet is funded, the DNS can instead be a rich source of data when looking at usage records.

This paper describes the design and implementation of a passive DNS monitoring system developed by the authors and used to monitor the .it country code code Top Level Domain (ccTLD). Focus of this work is not to monitor the DNS service per-se, but rather exploit the DNS system for understanding evolving trends and interests of Internet users, as well identify economical relationships.

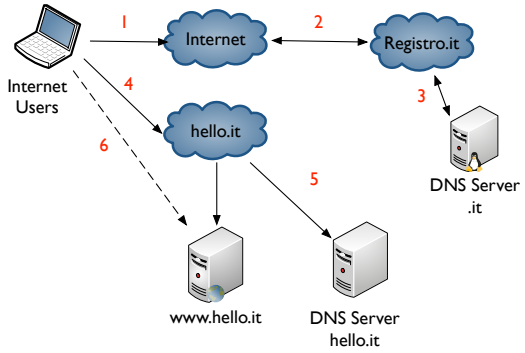
**KEYWORDS**

Domain name system, network traffic monitoring, Internet usage trends.

## 1. INTRODUCTION AND MOTIVATION

The domain name system (DNS) is a distributed database system that allows numeric IP addresses used in the Internet protocol suite to be associated with human-readable names. The DNS structure is organized as an inverted tree with the root at the top. Each node in the tree has a text label which identifies the node relative to its parent. Each node (or domain) can be further divided into additional partitions, originating in this case a new subtree (or subdomain). A Top Level Domain (TLD) is a "first level" domain, so it is a child of the root. Management of TLDs is delegated by the Internet Corporation for Assigned Names and Numbers (ICANN), which is also in charge of maintaining the root zone. The top-level domain space is mainly organized in country code Top Level Domains (ccTLDs), un/sponsored TLDs (e.g. .com, .net, .travel), and generic Top Level Domains (gTLDs). National domains are conventionally specified using the two-letter ISO 3166-1 country code, and are known as ccTLDs.

Figure 1. Iterative DNS address resolution for domain names.




The DNS protocol is based on the client/server paradigm. A DNS server stores DNS records for a set of domains for which it is authoritative (i.e. responsible), and answers to database queries that have been

performed using the DNS protocol. Every zone has a configured set of DNS authoritative servers. The client-side of the DNS is called resolver, and it is responsible for translating a domain name into an IP address or vice-versa. The address resolution mechanism is a sequence of queries used to resolve an address starting with the top level domain label. Using a file that contains the list of known root servers, the resolver first contacts a root name server, in order to obtain the address of one of the DNS servers authoritative for the TLD. Then it queries the obtained TLD server in order to obtain the address of the server authoritative for the second-level domain. This sequence is repeated until the address is resolved. Please note that in order to reduce load on the DNS servers, DNS makes use of caching. For this reason DNS responses do not last forever, but they have a time to live (TTL) set by DNS server administrators. A short TTL increases the number of queries as responses can be cached for a short period of time, whereas long TTL values delay the propagation of changes to DNS records as new queries will be performed only when cached values expire.

In order to make address resolution robust to failures, each domain must have at least two domain servers defined. In order to balance the load across them, round-robin is often the preferred policy for selecting a serves, although many resolvers cache DNS response time so they can prefer those servers that respond quicker than others.

## 1.1 Motivation

The DNS is often perceived as a “hidden” Internet protocol, necessary for translating symbolic host names into numeric IP addresses and vice-versa, but unknown to most users. This is because the DNS is used by applications such as web browsers or email clients, and not by end-users. As most Internet applications rely on it, network administrators monitor the DNS service in order to keep it operational and able to respond to requests in a limited amount of time. Tools such as dnstop, dsc and TreeTop can be used to analyze DNS traffic, and also create reports based on the observed traffic. Others tools such as Nagios and SmokePing can be effectively used to detect name server failures, as well as monitor DNS response time and jitter. Security, performance and traffic visualization are other areas where research on DNS is currently focusing.
In the past few years, the DNS is also appealing for companies that are using it for various reasons not immediately related to its governance, including traffic redirection for non existing domains, Internet user profiling and bad ISP practices that use the naming service for increasing their profits and perhaps resell information about DNS queries performed by users.

The authors of this paper are working for the .it DNS registration authority (Registro.it), that runs the .it ccTLD. This means that users who want to resolve addresses such as www.hello.it need to contact the .it DNS servers for first figuring out the names of the authoritative DNS servers for domain hello.it and then contact one of such servers for resolving. This means that all Internet application that need to contact hosts registered under the Italian (.it) DNS tree will contact the DNS servers of the Registro.it, and thus that monitoring the DNS traffic on such servers we can produce statistics for .it regardless of the ISP used by users for connecting to the Internet and their geographical location.

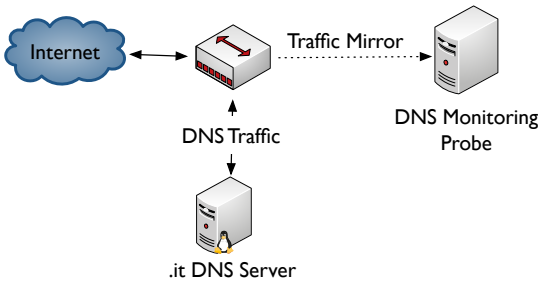
Beside monitoring the DNS protocol per-se, we have then decided to exploit the privileged position of the Registro.it in order to analyze the evolution of trends and interests of users that contact .it hosts. In essence we decided to produce detailed usage reports similar to Google Zeitgeist and Akamai State of the Internet, with the main difference that for the first time this activity is carried on by exploiting the DNS servers instead of other protocols such as HTTP. The fact that all Internet applications, and not just web browsers, use the DNS server, is very important as we can monitor not just web traffic but also other protocols including email, chat and messaging. This has been the motivation behind this work: passively identify interests and trends of a country by analyzing a limited amount of traffic (the DNS traffic is very little when compared to HTTP for instance), and monitor how these interests change over time. Due to the nature of our DNS infrastructure, we can focus only on .it Internet domain. This fact however does not limit the scope of our work as the used methodology is pretty general and it can be applies to other domains such as .com or .net.

The rest of the paper is organized as follows. Section 2 describes the design and architecture of the monitoring system. Section 3 presents the main results obtained while monitoring traffic. Section 4 highlights some open issues, future work items and extensions for the measurement architecture described on this paper.

## 2. DNS MEASUREMENT ARCHITECTURE

Due to the distributed nature of DNS, in order to monitor the whole .it ccTLD we need to monitor all .it DNS servers. For the .it this is not possible at the moment, as two (out of seven) .it DNS servers are run by third party companies that do not allow us to install DNS probes on their servers. It means that we are currently monitoring most of the .it DNS traffic but not all, even though this limitation does not affect our measurements. This is because DNS servers are selected by resolvers in a round-robin fashion with more preference to those with a lower response time, and thus it cannot happen that a resolver will contact just those DNS servers outside of our control. The widespread use of anycast addressing in DNS is also supported by our methodology as we monitor both unicast and anycast servers, and thus addressing does not affects our results.

Figure 2. Architecture of a DNS monitoring node.




For each DNS server we monitor, we have decided to create a passive monitoring infrastructure in order to avoid adding dependencies on existing components as well as load on DNS servers. This approach allows us to be independent from the DNS server type and configuration, and thus generic and suitable for monitoring even non .it servers. The DNS traffic received by a DNS server, is replicated at packet level via a traffic mirror, so that a monitoring probe (based on the open source nProbe developed by one of the authors) also receives a copy of the traffic. Each monitoring node in additions to nProbe also features a SQL database where monitoring data is stored, and a PHP-based web GUI for accessing the monitoring data. The probe associates DNS requests with replies and it also produces additional information including, geo-location of clients, response time, number of hops (TTL), and response type.

Figure 3. Sample DNS Probe Output.

```
#  When|DNS_Client|AS|ClientCountry|ClientCity|DNS_Server|Query|NumRetCode|
RetCode|NumAnswer|NumQueryType|QueryType|TransactionId|Answers|AuthNSs|
Cli2SrvTTL|Srv2CliTTL|NumQueryPkts|NumReplyPkts|ServerResponseTime(ms)
#
1343038802.127|xxx.xxx.xxx.xxx|16509|US|Seattle|194.0.16.215|www.xxx.it|0|
NOERROR|0|1|A|4178||ns1.xxxxx.it;ns2.xxxxxx.it|46|64|1|1|0.701
```

The DNS probe can export this data to a collector via NetFlow or dump it to a text file with the format depicted on the above figure. This raw data is not imported ‘as is’ into the node database, but it is first preprocessed in order to save only aggregated information as explained in the next section. We have decided to dump most DNS fields in order to provide monitoring applications enough information to compute all possible statistics without having to access to raw packets.

Out of the computed metrics, non-scalar values (i.e. those that have a quantity greater than one such as “top X DNS clients”) are dump on the SQL database, whereas numeric values are saved as time-series on a new database type named TSDB (Time Series DataBase) developed by the authors. The advantage of TSDB with respect to SQL databases or widespread tools such as the Round-Robin Database (RRD), is it ability to handle million of measurements with limited disk space and high efficiency during data insert and retrieval. This feature is very important as it has allowed us to implement a feature rich tool that allows us to keep track of all metrics and not just of selected ones. For instance for all DNS clients and .it domain name, we keep time-series that include, but are not limited to, the number of positive/negative queries, the round-trip

time (i.e. the network latency from the DNS server to the client), and the number of queries for AS (Autonomous System). As we have around 2.3 million .it registered domains at the time of writing, the total number of time series exceeds the number of hundred millions that it is unfeasible to handle with good performance on a SQL database.

Each monitoring node has a view of the local DNS queries, so accessing the node web interface we can access only local data. On the other hand, it is also important to have an aggregated queries view so that we can summarize queries across all monitoring nodes. We decided to avoid moving monitoring data from all nodes to a central point, as this requires a significant amount of data to be transfer daily that might be a costly activity as our DNS servers are places on IXPs (Internet Exchange Point) that usually charge per MB of data transfer. For this reason, we daily aggregate on a central point selected metrics, leveraging on HTTP for aggregating on the fly traffic reports as the user accesses the web interface. This solution has the advantage of moving data only when needed during drill-down whereas the base aggregated reports are immediately available.

We are aware that this is a very privileged position also in terms of privacy. In fact the local law does not allow network administrators to install wiretaps in order to analyze traffic without a specific authorization. Instead, monitoring traffic at the DNS servers is a legitimate activity as in order to do that we do not need to divert traffic at all as the users are themselves accessing the DNS servers. The raw data we analyze is stored temporarily for the purpose of aggregating it (we do this every hour) and it is then deleted. All data we maintain is on aggregated form, and the reports we produce do not contain raw numbers. Whenever we need to compare entities (e.g. queries to a specific domain, or registrars) we use ranking and not absolute numbers.

The DNS monitoring system described in this paper is active since more than a year and is constantly under development. In this section we have described the design of the DNS monitoring platform. The following section describes the metrics we collect and the measurements we preform on them.

# 3. MEASUREMENT RESULTS

As previously discussed, monitoring DNS traffic is a simple yet effective way to understand Internet users interests and trends. The amount of DNS traffic in fact is negligible when compared to other protocols such as HTTP thus making monitoring of high-speed links feasible using commodity servers. In our measurement we have observed that DNS traffic is less than two order of magnitude of HTTP with a typical .it DNS server handling a peak of 6 Mbit/s of DNS traffic. Although many researchers ground their studies on social networks traffic such as Facebook and Twitter, we believe that social networks cannot be used to represent all interests as most Internet users do not actively use social networks, and also because non-social trends (e.g. business and research) can be hardly measured using such approach. This as the DNS is used by all Internet protocols (e.g. P2P, HTTP), and thus regardless of the applications, the DNS is able to catch the domain names being accessed and thus, once they have been categorized, the interests of Internet users.

Another important aspect of DNS monitoring is that we do not have to face with privacy concerns that instead affect other approaches. By law, wiretapping or diverting traffic toward a monitoring probe requires special security authorizations as such probes do not monitor traffic that is terminated on such probes but rather that flows from the vantage point where the probe is located. Instead in our case, we monitor the traffic of Internet users who voluntary access .it DNS servers and thus we are not artificially copying/diverting traffic making this practice legal and similar to web administrator who produce statistics on accesses to their web servers.

As previously discussed, out measurement system focuses uniquely on .it domains as this is the only traffic we can monitor at our vantage points. However the work discussed on this paper is pretty general and it does not relies on peculiarities of .it domains, thus making it suitable for a broader audience. The main goals of our measurement system are:

- Ability to produce usage records and ranking (in terms of number of queries) for each monitored domain.
- Characterize each domain assigning it a typology (e.g. media, sport, travel) so that we can collect usage trends.

- Identify Internet users interests based on the number of DNS requests for categorized .it domains.
- Associate interests with geolocation, so that we know what are the main interests of users located on each country when accessing .it domains.
- Characterize economical relationships looking at DNS traffic.
- Evaluate DNS provisioning quality in terms of response time, so that we can improve the DNS servers location in order to place them closer to places where most DNS clients are located.
- Identify Internet domain lifecycle, so that we can identity in advance emerging companies and web sites by observing how the number of queries increases with respect to domain lifetime (i.e. the amount of time passed since the Internet domain was first registered).

The following sections analyze in details all the above goals grouped into macro-categories.

### 3.1 Characterization of Internet Domains

DNS traffic records are useful to measure users activities and geolocate them both in terms of country and also autonomous system (AS) as explained in the next section. What DNS records cannot provide is the categorization of the Internet domain being searched. Companies such as Alexa (http://www.alexa.com) and voluntary-based directories such as the Open Directory Project (http://www.dmoz.org/) as well all the major search engines try to categorize (semi-)automatically Internet domains. Unfortunately this categorization is very limited for non-English/US domains, and thus we cannot use any of these domains for creating a comprehensive directory of Italian domains.

For domains registered by companies, using their VAT ID, we have planned to use the company categorization defined by the Chamber of Commerce. Unfortunately we have abandoned this approach, because defined categories were too specific and cumbersome, and too close to economics rather than to the Internet. In addition we have realized that nothing prevents a given company that does business in sector X (e.g. toys) to setup a web site on a close but different category (e.g. sexy toys) thus making this approach error-prone. While in literature there are some semi-automatic tools for web content mining, most of them are not freely available or are have used just for a specific research project. This made unfeasible for us to rely on third party web mining tools, and thus we decided to start developing our own tool for categorizing web sites. In the meantime, we have decided to start characterizing Internet domains content by using a semi-manual approach. Based on some past work, we have created a comprehensive list of categories and for each category we have identified some keywords. For instance the words bank/banca (its corresponding Italian translation) belong to the list of words for the category named “finance”. Everyday an automatic tool uses these words to characterize new domains so that humans can just double-check this work by visiting the corresponding web site. Note that sometimes this step is delayed as new domains do not have web sites or they use parking sites until the web page has been setup. Furthermore the first 1’000 domains with most queries and the emerging domains (those that increased significantly their ranking in the past weeks/month) are periodically categorized by hand in order to characterize top domains. IDN (Internationalized Domain Names) domains, i.e. domain names that contain non-latin letters, are grouped on a special category and not put on the categories we already defined, as they are mostly used to redirect users to the main site that we already classified (e.g. nestlè.it redirects users to nestle.it).

In order to further classify domains, we have also compared the number of queries with the domain lifetime, that is the number of days since the domain registration, information that can be easily collected from the domains database. We divided domains in two categories:

- Long-lasting domains that have been registered since more than 6 months. Often business domains fall into this category as they are registered by companies. Seasonal events (e.g. winter ski) also fall into this category.
- Short-lasting domains registered for events of a limited lifetime (e.g. sport events, political parties candidates, concerts and expo) and that often have a year in their name (e.g. london2012.it).

This has allowed us to easily identify:

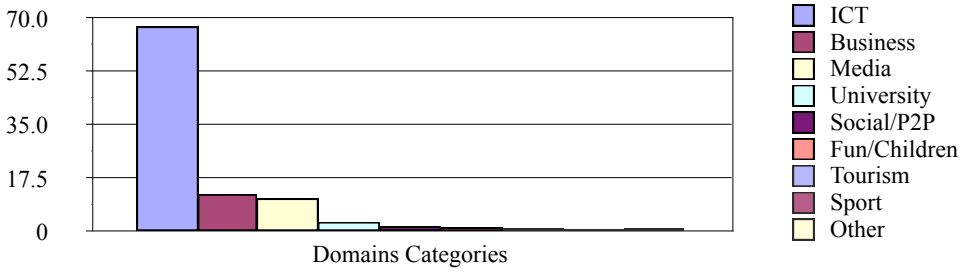
- Domains just registered (e.g. registered during the last month) that collect a large number of requests on short period of time, and then after a peak they quickly decrease of interest. These domains are mostly used for one-time, seasonal (e.g. winter ski) or periodic (e.g. political elections) events.

- New domains that are growing rapidly over a long period of time and that likely identify quick growing interests (e.g. online bet) or companies. These domains must be observed as they can represent long term trends if their growth persists over a long period of time.

Correlating the above information with the domain type, we claim that it can be used for contributing to define the sociocultural trends of a country.

Since the initial deployment of our DNS monitoring system that has happened more than a year ago, we have realized that this approach is good as in terms of DNS queries their distribution follows the Pareto distribution, thus analyzing a small portion of domains allows us to characterize most of the queries for .it domains. The following figure shows the distribution of the top 1000 .it domains (in terms of total queries) according to the categories we have identified. Note that the caching of DNS responses as specified in the TTL field of the records, prevents us from receiving all the requests that instead reach the DNS resolvers that contact us. We are aware that caching introduces some noise in the results, and thus we mitigated the results it by normalizing the number of queries per .it domain to a reference TTL of 86400 sec (1 day).

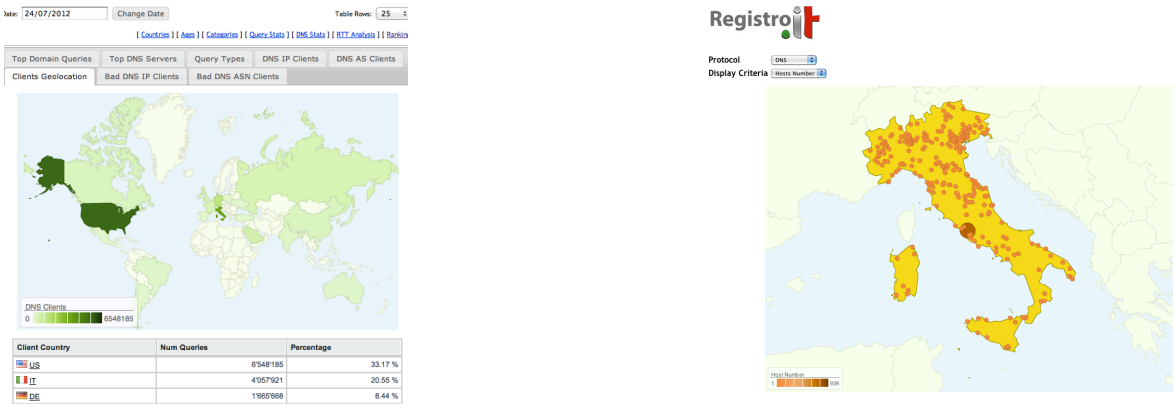
Figure 4. Percentage of Queries per Domain Category in .it




## 3.2 Geolocating Users and Interests

nProbe is able to geolocate DNS clients using a popular tool named GeoIP that includes a database for locating hosts based on their IP address. Due to the nature of the DNS we do not see the IP address of the user who performed the DNS request, but the IP of its ISP (or public DNS) from which the received the query. We have run some tests to verify the accuracy of this approach, and we have verified its accuracy for Italian networks by matching domain located data present in the whois database with GeoIP. We have verified that at country level, in 66% of the hosts GeoIP and whois data overlaps, whereas on the remain 33% the discrepancy does not usually exceeds 300 Km. Our conclusion is that looking at these data it is not possible to come up with a certain answer, as a domain registered in town X might be hosted on town Y.

Figure 5. Demographic Distributions of Queries for .it domains: International and Domestic.

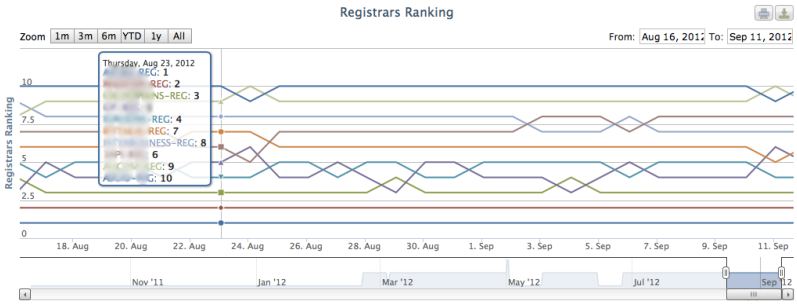


Seen that geolocation at town level might not be too accurate using GeoIP, we wanted to verify the accuracy of the GeoIP database at country level. Instead of using the whois approach we used previously that did not lead to a certain conclusion about host location, we decided to use a different approach. The goal is not to find out whether a certain host located by GeoIP in country X is really there, but to identify those situations where such hosts are definitively on a country other than X. We have then created a tool named *quickping*, that exploiting the same algorithm used by the traceroute tool, allows us to ping hosts much quicker than the ping tool that comes with the operating system while, in case of no host response, returns the round-trip-time (RTT) from the host closer to the target that sent a response back. This way we can ping ~1 million hosts (i.e. the number of DNS clients that in average contact each monitoring probe every day) in a couple of hours, providing a round-trip-time value for each client. By using GeoIP and some math we compute the distance in Km of the place where the host is located with respect to our monitoring system. As the data propagation speed in optical fibers is about 2/3 of the light speed, we compute the minimum RTT that a given host can have. Comparing the RTT with the minimum RTT, we identified some hosts (less than 10% of the DNS clients that contacted our servers) that have a RTT less than the minimum RTT. We marked those hosts as suspicious as we need to further investigate case-by-case, so that such hosts are not considered on out reports.
Geolocation allows us to produce various reports including the distribution of queries per country, overall, per .it domain, and per category. In other words we know what are the countries from which we see queries for a given .it domain or category (e.g. media). For instance we produce reports that leveraging on domain categorization and geolocation, allow us to know what are the main interests of Internet users located in country X when accessing .it, or the query distribution per country of DNS queries for domain xxx.it. This information is interesting for many people including for the local administration, that uses that for identifying how this country is perceived abroad both in terms of interests and also of companies, thus business types, associated with DNS requests. We want to emphasize that this information can be gathered easily in our privileged monitoring location, whereas it would be much more complicated to obtain when using other means, this unless we are a large search engine or public DNS provider.

## 3.3 Using the DNS as Economical Indicator

We believe that the DNS traffic can also be used as economical indicator. Using geolocation and domain characterization we can collect statistics about the type of businesses that are located on a given geographical area, as well identity what are the geographical regions that generate most DNS queries for a given domain type (e.g. tourism). This information can also be used for identifying areas of digital divide or places where the Internet penetration, in terms of domains registration and availability, is limited as depicted on figure 5. Furthermore, using the number of queries as economical indicator, we have created a registrars (i.e. the companies that register .it domains under authorization from the .it DNS registry) ranking based on the number of queries for domains registered by such registrars.

Figure 6. Registrar Ranking as Economic Indicator




This is a novel idea, as currently registrars are ranked based on the number of domains they registered, that gives a better score to registrars that registers domains at low costs but that often are visited very seldom by Internet users. During this analysis we have also used the HHI (Herfindahl-Hirschman Index) index, a widespread indicator of competition among companies, has been used to evaluate the concentration ratio of registrar. The result has confirmed that for .it, the first 10 registrars register more than 50% of domains making this market tough for small registrars, whereas most of these registered domains are not hosted by

such registrars (HHI index of 280). This means that the domain registration business is concentrated on few players hands, whereas domain hosting is quite open and competitive for many companies.

## 4. OPEN ISSUES AND FUTURE WORK

As described earlier on this paper, we are now focusing on an automatic domain classification tool that would enable us to characterize most, if not all, .it domains in a semi-automatic fashion. This is very important because it would allow us to create broader reports instead of focusing only on top domains. The tool leverages on a web crawler and a classification system developed for a smaller project that we will extend and make it suitable for classifying the whole .it DNS tree.

In addition we are plan to develop more sophisticated techniques for detecting DNS scanners (i.e. those that scan the .it tree and resell the list of .it domains mostly for illegal purposes such as spamming), and typo squatters. Domain squatting, also known as typosquatting, is the act of registering or using a domain name with bad faith so that it is similar to a popular name or trademark not belonging to the registrant. We plan to use our existing tool, for identifying the domain names that can be classified as typosquatting, and understanding what is the ratio between traffic sent to the original site with respect to typosquatting sites.

## 5. CONCLUSION

This paper presented a novel approach to DNS traffic analysis, whose goal is to understand the trends and interests of a country by analyzing queries to ccTLD domain servers. To the best of out knowledge, this is the first attempt to use the DNS for this purpose when measuring requests at a country level, and not limited to a specific organization such as an ISP or search engine. This work has demonstrated that although the DNS is not an ideal protocol when compared to HTTP for understanding trends and interests, it is a solid approach alternative to similar attempts often based on the analysis of social networks whose data might be biased by unfair players that create artificial interests and vogues (“like” in the FaceBook parlance, or “followers” on Twitter).

## REFERENCES


A. Bonaccorsi, A. Del Soldato, M. Martinelli, C. Rossi, and I. Serrecchia, 2002. Measuring Internet Diffusion in Italy. *In Proceedings of IFIP Workshop on Internet Technologies, Applications, and Societal Impact 2002*, Wroclaw, Poland.

B. Huberman, D. Romero, and F. Wu, 2008, Social Networks that Matter: Twitter Under the Microscope. *In Social Science Research Network*, http://dx.doi.org/10.2139/ssrn.1313405.

F. Wang, G. Agrawal, R. Jin, and H. Piontkivska, 2007, SNPMiner: A Domain-Specific Deep Web Mining Tool. *Proceedings of BIBE 2007*, pp. 192 - 199.

L. Deri, 2003, nProbe: an Open Source NetFlow Probe for Gigabit Networks. *Electronic proceedings of Terena Networking Conference (TNC) 2003*, Zagreb, Croatia.

L. Deri, S. Mainardi and F. Fusco, 2012, tsdb: A Compressed Database For Time Series. *Proceedings of TMA 2012*, Vienna, Austria, pp. 143-156.

T. Hardie, 2002, *Distributing Authoritative Name Servers via Shared Unicast Addresses*, RFC 3258.

R. Kosala and H. Blockeel, 2000, Web Mining Research: A Survey. *In ACM SIGKDD Explorations Newsletter*, 2(1), pp. 1-15.

R. Wilson, S. Gosling, and L. Graham, 2012, A Review of Facebook Research in the Social Sciences. *In Perspectives on Psychological Science*, 7(3), pp. 203-220.

R. Govindan and H. Tangmunarunkit, 2002. *Heuristics for Internet Map Discovery*. Proceedings of ACM SIGCOMM, Stockholm, Sweden.

J. M. Pierre, 2001, *On the Automated Classification of Web Sites.* Linköping University Electronic Press.

J. Lindholm, 2003, Experiences of Harvesting Web Resources in Engineering using Automatic Classification. *Ariadne*, No. 37, http://www.ariadne.ac.uk/issue37/.

Google Inc., 2011, *Google Zeitgeist 2010*, http://www.google.com/zeitgeist.

Akamai Technologies, 2011, *State of the Internet: Q4 2011 Report*, http://www.akamai.com/stateoftheinternet/.