# Website Hosting Data and Analysis

**Petr Ilgner** [1] | *Brno University of Technology, Brno, Czech Republic*
**Dan Komosný** [2] | *Brno University of Technology, Brno, Czech Republic*
**Saeed Ur Rehman** [3] | *Auckland University of Technology, Auckland, New Zealand*

***Abstract***

We have collected a large dataset – more than 21 000 websites – through web-crawling the public resources of the Czech Internet. The proposed method for website hosting detection along with their geographic location and software were applied on the collected data to extend basic statistical information about the Czech websites published by the national domain registrar CZ.NIC. For analysis, we divided the data into nine categories to show differences between them, for example, between the public and private sector. The procedures used in this paper may also be applied for an extended analysis of websites in other countries, for example, for verification of fulfillment of legal directives to be implemented by public sector.

***Keywords***

*Internet, web content, hosting, geographical location, Czech Republic, CZ.NIC*

***JEL code***

*C63, C80, L86*

## INTRODUCTION

Statistical data about the Internet in a country are used for various purposes. An example may be verification of fulfillment of legal directives issued by a government to be implemented by governmental institutions, public sector entities, and Internet Service Providers (ISP). From the user's point of view, the data may be also used for checking the shared resource plans as published by web hosting service providers.

In various countries, there are national domain registrars that publish statistical data about the national Internet. These data may provide information about domains, DNSSEC, DNS traffic, IPv6, registries, etc. The data are typically published as 'open access'.

Czech registrar CZ.NIC (2017, CZ Domain Hosting Statistics) presents data about domains remotely hosted by particular hosting providers (further referred to as 'hosted'). The data are divided into three sets: i) domains at the organization premises (further referenced as 'self-hosted'), ii) domains hosted at other organization, and iii) domains with unknown hosting status. If a hosting provider is not listed, the owners can report their data to the registrar. The hosting data are specifically categorized to mail hosting, nameserver hosting, and web hosting. The web hosting is of primary importance as it reflects the situation for end-users accessing the web content.

[1] Department of Telecommunications, Faculty of Electrical Engineering and Communication, Brno University of Technology, Technická 12, 612 00 Brno, Czech Republic. E-mail: petr.ilgner@vut.cz, phone: (+420)541146927.
[2] Department of Telecommunications, Faculty of Electrical Engineering and Communication, Brno University of Technology, Technická 12, 612 00 Brno, Czech Republic. E-mail: komosny@feec.vutbr.cz, phone: (+420)541146973.
[3] Department of Electrical and Electronic Engineering, Auckland University of Technology, 34 St. Paul St., Auckland, 1010, New Zealand. E-mail: saeed.rehman@aut.ac.nz.

In this paper, we present these statistics about websites: i) website hosting data, ii) specific data for the defined website categories by their content, iii) geographical related data, and iv) data about security implementations and used software.

Our data come from more than 21 000 websites that we have collected by crawling public resources of the Czech Internet, including the lists of web addresses in public company directories at www.firmy.cz, www.sreality.cz, www.toplist.cz.

The details about the data presented in this paper are the following:

i) Web hosting detection is a complex process as there is no direct (straightforward) approach to identify a website to be self-hosted or hosted. Therefore, we propose a method consisting of four particular tests. These tests aim to identify the hosting status based on information 'clues' that can be obtained from public resources. We consider these information pieces: 'reverse domain lookup', 'database of web hosting providers', 'network owner name', and 'network administrator email'. Each of this information is assigned a weight to calculate the final hosting status.
ii) The data available from CZ.NIC (2017, CZ Domain Hosting Statistics) are global numbers with no particular information about the websites and the entities they represent in their content. Therefore, we define nine entity categories as follows: banks, e-shops, hospitals, insurance companies, real estate agencies, craftsmen, government institutions, secondary schools, and universities. For each category we detect whether the entity website is self-hosted or hosted. Additionally, we show the share (or popularity) of web hosting providers across the categories.
iii) The base data do not cover the geographical distribution of the servers. Therefore, we relate the data to the location of web servers. We show the numbers for the Czech regions, cities and we list countries hosting websites with Czech content.
iv) Finally, we include the data about security implementations in each of the defined categories and give statistic about the software used for running the web.

The procedures used in the paper could be used for various purposes, including market analysis or motivation for websites hosting improvements in terms of reducing load on Internet resources. The latter one is of particular importance as previous research showed that communication latency in company web pages access has a correlation with the revenues (Sigla et al., 2014). Therefore, the decision of self-hosting or hosting, including selection of the hosting provider and its geographical location may be important. Also, the shared web hosting plans can be verified using the described method with large input data (number of websites).

The paper is structured as follows: Section 1 indicates how CZ.NIC obtains the hosting data and discusses the results they publish. Their results are compared with other sources. Related papers to this work are described, mainly considering the hosting status check. In Section 2 we describe in detail our approach to detect the web hosting status. The examples are given for each particular test, including the source of the input data. We also show how we obtained the geographical data and other related information. Description of the implementation, including web crawling, and the detailed numbers about the collected websites are given in Section 3. Section 4 discusses the results and it is divided into particular subsections according to the data type.

## 1 RELATED WORK AND DATA

Wang et al. (2011) proposed a method for IP geolocation that included identification of website hosting status. They assumed that the same IP address is used for a set of websites (domain names) hosted at the same provider (possible in the order of hundreds). On the other hand, if a specific IP address is used for a single website then it is concluded that such site is very likely to be self-hosted. They detected the web hosting status by accessing a website by its domain name and by its IP address. The returned homepage was checked according to these three suggested options i) its content, ii) head information

(<head></head>), and iii) title information (<title></title>). If this information was equal (based on the selected option), they will conclude the IP address represented a single website, i.e. the site was self-hosted. The different information may be a blank page or error message. They also stated a problem with this method when the first request is redirected. In this case, they sent an additional request for the targeted page.

Tsou et Lusher (2015) grouped the websites into twelve categories, such as 'News', 'Entertainment', 'Forum', and 'Non-profit organizations'. They compared the geographical information obtained from the pages of categorized websites to the location estimated for the IP address of the server running a site. The geographical information of web pages was taken by a manual inspection of the page content when looked for text such as 'Contact' or 'Privacy Policy/Terms of Service'. The postal address of the content creator was used. If the postal information was not found on the pages, the external information source was used for the looked-up company, such as Wikipedia. The location of the web server (given by its IP address) was obtained from the location database (Maxmind, 2017; GeoIP2 Databases). The categorized websites were compared in order to see the difference between the postal address obtained for the content creator and the location obtained for the IP address of the server running the site. The threshold for similar location identification was 50 miles as a range of a city. The most geographically accurate (smaller difference) were websites in the categories 'Educational', 'Social Media', and 'Governmental'. The least accurate websites were in the categories 'Blog', 'Special Interest Group', and 'Non-profit organizations'.

The primary data about web hosting in the Czech Republic are provided by CZ.NIC (2017, CZ Domain Hosting Statistics). There are 72 web hosting providers listed in total, see details in the Annex. Selected data are shown in Table 1. The table shows the first ten Czech web hosting providers. The row 'Unknown' shows that about 40% of websites could not be determined (self-hosted or hosted). The last row shows the number of self-hosted websites.

**Table 1** Share of Czech web hosting providers

| Rank | Czech hosting provider | Share [%] |
|---|---|---|
| 1 | WEDOS Internet, a.s. | 12.27 |
| 2 | FORPSI | 7.71 |
| 3 | ACTIVE 24 | 6.14 |
| 4 | ZONER software, a.s. | 2.80 |
| 5 | Cesky hosting* | 2.57 |
| 6 | Web4U, s.r.o | 2.24 |
| 7 | Gransy, s.r.o | 2.02 |
| 8 | Ignum s.r.o | 1.57 |
| 9 | ONEsolution, s.r.o. | 1.50 |
| 10 | Stable.cz | 1.49 |
| – | Unknown | 40.85 |
| – | Without hosting | 4.97 |

**Note:** * – (THINline interactive, s.r.o.).
**Source:** CZ.NIC, November 2017

Other data about web hosting are available from BuiltWith (2017). The global (world) numbers are given along with the numbers for particular countries. The Czech Republic is also listed, and the results are shown in Table 2. The table again shows the top ten hosting providers. Some hosting providers listed were

not the same as with CZ.NIC. It is probably due to the used BuiltWith methodology based on comparing data from the IP address allocation database of RIPE NCC. Most Czech web hosting providers have not assigned their own IP address block and use the block of other ICT companies, such as Casablanca INT, which is included in the BuiltWith statistics.

**Table 2** Share of Czech web hosting providers

| Rank | Czech hosting provider | Share [%] |
|---|---|---|
| 1 | Casablanca INT | 20.6 |
| 2 | WEDOS Internet | 17.7 |
| 3 | SuperNetwork | 16.8 |
| 4 | VSHosting | 15.4 |
| 5 | Internet Cz | 8.0 |
| 6 | Zoner | 6.3 |
| 7 | Active 24 | 5.2 |
| 8 | Ignum | 4.0 |
| 9 | Dial Telecom 2 | 2.0 |
| 10 | CESNET | 2.0 |

**Source:** BuiltWith, November 2017

Table 3 compares the values for the specific hosting providers listed in both sources included – WEDOS, Active 24, Zoner, and Ignum. The column 'Difference–Share' shows values ranging from 1 to 5%. The percentages are of small values and therefore, only a small difference changes the rank. The last column 'Difference-Rank' shows the change in the relative order.

**Table 3** Data about Czech web hosting providers

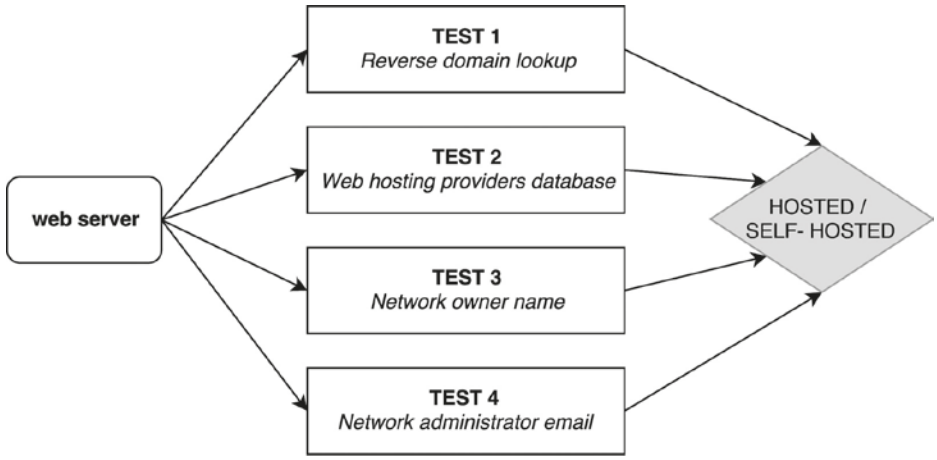
| Web hoster | CZ.NIC | BuiltWith | Difference-Share | Difference-Rank |
|---|---|---|---|---|
| WEDOS | 12.27 | 17.70 | 5.43 | 1 (1–2) |
| Active 24 | 6.14 | 5.20 | 0.94 | 4 (3–7) |
| Zoner | 2.80 | 6.30 | 3.50 | 2 (4–6) |
| Ignum | 1.57 | 4.00 | 2.43 | 0 (8–8) |

**Source:** Own construction

## 2 METHOD FOR DETECTION OF WEBSITE HOSTING

Currently, there is no simple (straightforward) method known to detect whether a website is present at the owner's premises (self-hosted) or remotely hosted at a hosting provider (hosted). Related work refers to hosting identification by a single source of information. In our method, we use a set of different information that we process in a form of tests. For each test, we empirically assign a weight. The particular tests (source of information) are 'reverse domain lookup', 'web hosting provider database', 'network owner name', and 'network administrator email', as shown in Figure 1. We use the hosting status results for further data processing described in Section 4.

**Figure 1** Web hosting status identification




**Source:** Own construction

In the following subsections we describe each test in detail, list input data (when used), and present a use-case example.

*Test 1 –* ***Reverse domain lookup***

For self-hosted sites owned and managed by an organization, it is expected that an administrator sets up the PTR records in the DNS system for each server owned. The PTR (pointer to a canonical name) records associate IP addresses with domain names and this information is used for reverse lookups (IP-to-domain). It shows if an IP address is used in a specific domain.

The test 'reverse domain lookup' compares a website domain name with the domain name of the IP address obtained from a DNS reverse query. Firstly, the IP address of the server is found by a DNS A query. Using the following DNS reverse query, the domain name for that IP is obtained. If the domain name for the tested web server matches the found PTR record, the website is self-hosted as shown in Listing 1.

This test is skipped if the PTR record is not set for an IP address. It may also falsely indicate hosting if the domain name of ISP is used instead of the organization's name.

Note: Our method considers Virtual Private Servers (VPS) as 'self-hosted'. They use not-shared IP addresses and a virtual server is maintained by the website owner.

**Listing 1** Example of reverse domain lookup comparison test result executed in dig

```
$ dig www.mendelu.cz AAA +short
valar.mendelu.cz
195.178.72.2
$ dig +short -x 195.178.72.2
valar.mendelu.cz
```

**Source:** Own construction

*Test 2 –* ***Web hosting provider database***

Web hosting providers typically set up a DNS PTR record for the hosted websites to point to their domain name. For example, a Czech hosting company WEDOS assigns the domain names for their sites in the following pattern `xxx.wedos.net`. If the second-level domain name `wedos.net` is included in a list of domains of known web hosting providers, the website is very likely to be hosted. For the purpose of this test, we have collected domain names of the Czech web hosting providers from DNS reverse lookups and manual verification at the provider web page. The created list is shown in Table 4.

**Table 4** List of known web hosting providers as input data for test 2. Some providers use multiple domain names, such as 'Cesky hosting' and FORPSI

| Trademark | Domain name | Trademark | Domain name |
|---|---|---|---|
| ACTIVE24 | active24.cz | Neomezeny hosting | neomezeny-hosting.cz |
| AeroHosting | aerohosting.cz | Neomezeny webhosting | neomezeny-webhosting.cz |
| Ahosting | ahosting.cz | ONEbit.cz | onebit.cz |
| Angel hosting | angel-hosting.cz | Otoman | otoman.cz |
| aspone.cz | aspone.cz | oXyShop | oxyonline.cz |
| ATTIVO | attivo.eu | Pipni.cz | pipni.cz |
| Banan.cz | banan.cz | Profitux | profitux.cz |
| Bezobav.cz | bezobav.cz | Quantasoft Hosting | qhs.eu |
| BlueBoard | BlueBoard | Rosti.cz | rosti.cz |
| Cesky hosting | ceskyhosting.cz, thinline.cz | Savana | savana.cz |
| Datahousing | datahousing.cz | Stable.cz | stable.cz |
| domeny.as | domeny.as | Station webhosting | station.cz |
| Ebola | ebola.cz | SvetHostingu.cz | svethostingu.cz |
| eBRANA | ebrana.cz | Sweb | sweb.cz |
| Endora | endora.cz | Thosting | thosting.cz |
| Eshop-rychle | eshop-rychle.cz | Tojeono.cz | tojeono.cz |
| Exo hosting | exohosting.cz | Web areal | webareal.cz |
| FORPSI | forpsi(.com, .net) | Web zdarma | webzdarma.cz |
| Gigaserver | gigaserver.cz | Web4ce | web4ce.cz |
| Gigaweb | gigaweb.cz | Web4U | web4u.cz |
| HexaGeek | hexageek(.com, .cz) | WebDum.com | webdum.com |
| Hosting 90 | hosting90.cz | Webhosting C4 | webhosting-c4.cz, skok.cz |
| Hosting Blueboard.cz | blueboard.cz | WebHosting.FM | webhosting.fm |
| HostingSolutions.cz | hostingsolutions.cz | Webnode | webnode.com, rubicus.com |
| HostingZdarma.cz | hosting(-)zdarma.cz | Webprostor.eu | webprostor.eu |
| Hukot.cz | hukot(.cz, .net) | Websupport | websupport(.cz, .sk) |
| IGNUM | ignum.cz | WEDOS | wedos(.cz, .net) |
| iSOL.cz | isol.cz | ZONER | zarea.net |

**Source:** Own construction

The test may falsely indicate self-hosting if the web hosting organization is not included in our list. It may also falsely indicate hosting if the website of the hosting organization itself is tested. Table 5 shows an example of a positive evaluation of this test.

*Test 3* – ***Network owner name***

For both the domain name and the IP address of a web server, it is possible to get the holder name from the relevant registers. National domain names can be looked-up in the WHOIS database managed by a national registrar. Regarding the IP address, this information can be acquired from he international WHOIS database managed by RIPE NCC. If the names from both sources are the same, the website is likely self-hosted.

**Table 5** Example of 'web hosting providers database' test

| | |
|---|---|
| Web address | www.uzis.cz |
| Server IP address | 178.238.37.157 |
| Domain name for IP address | yivo.onebit.cz |
| Found hosting provider | onebit.cz |

**Source:** Own construction

This test may return a false hosting result if the holder of IP address is an ISP and not an end-organization. Also, the found names may not be exactly the same. For example, the company names stored in the CZ.NIC registrar are typically listed as the name plus some suffix according to the legal form of the institution, such as 's.r.o.' In such cases, it may also indicate false hosting result. In our implementation we calculate the similarity factor of organization names obtained from the registrar and RIPE NNC to eliminate the impact of same name variants.

An example of a positive test result is shown in Listing 2.

*Test 4* – ***Network administrator email***

Large organizations typically have their own IP address space. Therefore, the relevant WHOIS registry should also contain an email to contact the holder of that IP space, in case of abuse etc. If this email address is identical with domain name of the tested website, the site is evaluated as self-hosted as being run in the organization address space. Table 6 shows an example of a positive evaluation of this test.

This test may falsely indicate hosting if the email has a different domain name from the web server.

**Table 6** Example of 'network administrator email' test

| | |
|---|---|
| Web address | czso.cz |
| Server IP address | 194.48.241.132 |
| Assigned IP addresses | 194.48.241.0–194.48.241.255 |
| Administrator email | jiri.lejnar@czso.cz |

**Source:** Own construction

## 2.1 Final hosting result

As described above, the particular tests could indicate the hosting status, but they may also fail in some cases. Therefore, we empirically assign each test result a weight as shown in Table 7.

**Listing 2** Example of testing conformity of domain and network holder name (listing is shorted)

```
$ whois www.cvut.cz
contact: SB:R1S-CES-8079-FA
org: Ceske vysoke uceni technicke v Praze
name: Ceske vysoke uceni technicke v Praze
address: Zikova 4
address: Praha 6
address: 16636
address: CZ
e-mail: neuman@vc.cvut.cz
```

```
$ dig www.cvut.cz AAA +short
cvut.cz.
147.32.3.202
```

```
$ whois 147.32.3.202
organisation : ORG -CVUT1 - RIPE
org - name : Ceske vysoke uceni technicke v Praze
address : Ceske vysoke uceni technicke v Praze
address : Zikova 1903/4
address : Praha 6
address : 166 36
address : The Czech Republic
abuse - mailbox : abuse@cvut.cz
(listings shortened)
```

**Source:** Own construction

We assign a value of –0.5 when the test 'reverse domain lookup' shows that the website is self-hosted. We assign this value as we believe its accuracy is in-between the accuracies of the last two tests. We assign the last two tests values of –0.4 (lowest) and –0.6 (highest) respectively towards self-hosting. The second test 'hosting provider database' is very firm and therefore we assign it a value of +1.

The weight of every test is included in the final score. If any test fails due to technical reasons (e.g. DNS query fails), we exclude its weight. The final hosting result is given by a sum of the weights. If the final score for a website is zero or negative value, we evaluate it as self-hosted (otherwise hosted).

**Table 7** Test results and assigned weights – negative value indicates self-hosting

| Test | Result weight | |
|---|---|---|
| | Hosted | Self-hosted |
| Reverse domain lookup | 0 | –0.5 |
| Hosting providers database | 1 | 0 |
| Network owner name | 0 | –0.4 |
| Network administrator email | 0 | –0.6 |

**Source:** Own construction

## 3 IMPLEMENTATION, DATA SOURCES, AND COLLECTED DATA

For the purpose of this work, we developed an application consisting of several modules that we categorize as source of data, data processing, and data storage. The module for source data covers crawling the Web, getting data from Internet registries, getting geographical data, and getting other related data (such as security implementations). The module for data processing is used for parsing input data, hosting status identification, and data correlation (such as geographical location and latency). The module for data storage is used for accessing SQLite database, data exporting, and data plotting on a map. For our application we used the Python 3 programming language with these main packages: `folium`, `dnspython`, `pyquery`, and `requests`.

The data used in this paper were collected through web crawling the public resources, including the lists of web addresses in public organization directories, available at `www.firmy.cz`, `www.toplist.cz`, and `www.sreality.cz`. With the first two, we crawled the public lists of organizations listed under specific categories. The data from the latter one were collected from a list of real estate agencies available at `www.sreality.cz/adresar` by Bulín (2017) and they form an additional category. The data related to IP address space primary come from regional Internet registry RIPE NCC accessed via the WHOIS database using the public server, available at `whois.ripe.net`. The data related to domain names come from the Czech domain name registry CZ.NIC accessed via `http://www.nic.cz/whois/`. For geographical data we used the free MaxMind (2017) database 'GeoLite2 City' with a local access.

In total, we have collected and processed data from more than 21 000 websites divided into nine categories. The numbers of crawled websites for each category are shown in Table 8. We selected the categories to cover both public and private sectors: i) private large sector – big companies (banks, insurance), ii) private small business sector – small companies (e-shops, real estate agencies, craftsmen), and iii) public sector (hospitals, government, schools, universities). This division is only indicative as e-shops and real estate agencies may fall into both big and small companies. Also hospitals, schools, and universities may fall in both public and private sector. We did not check the legal status and size of the entities. We rather evaluated each category independently and the numbers may further be combined based on specific needs.

Table 8 shows that most collected websites were from the private sector – small companies for these categories: e-shops, craftsmen, and real estate agencies respectively. Following was the public sector with categories of high schools and hospitals, respectively. These are the sectors for which the results

may be considered as 'strong' as a large number of websites was processed. The rest of the categories (banks, insurance companies, government institutions and universities) have smaller numbers given the size of the Czech Republic. The results are therefore only indicative and should be interpreted with the knowledge of the size of input data.

**Table 8** Collected websites divided into categories

| Sector* | Category | Websites |
|---|---|---|
| Private/large | Banks• | 36 |
| Private/small | E-shops | 11 314 |
| Public | Hospitals | 469 |
| Private/big | Insurance companies? | 26 |
| Private/small | Real estate agencies | 1 660 |
| Private/small | Craftsmen | 6 563 |
| Public | Government inst.? | 109 |
| Public | High schools | 1 192 |
| Public | Universities? | 103 |
| **Total** | | **21 472** |

**Note:** * – Indicative division; • – Indicative results.
**Source:** Own construction

## 4 RESULTS

### 4.1 Overall Data About Website Hosting

Using the tests described in Section 2 we have evaluated the hosting status of the collected websites. Table 9 shows the detail results of the test methods and the detected overall hosting status of the collected websites. The test results detect most of them as hosted. The highest numbers of websites evaluated as detected belong to e-shops and real estate agencies. The lowest number is for universities. The reason may be that larger organizations such as universities typically have their own IT departments, while those with the lowest number of detected hosted websites contain mainly smaller organizations.

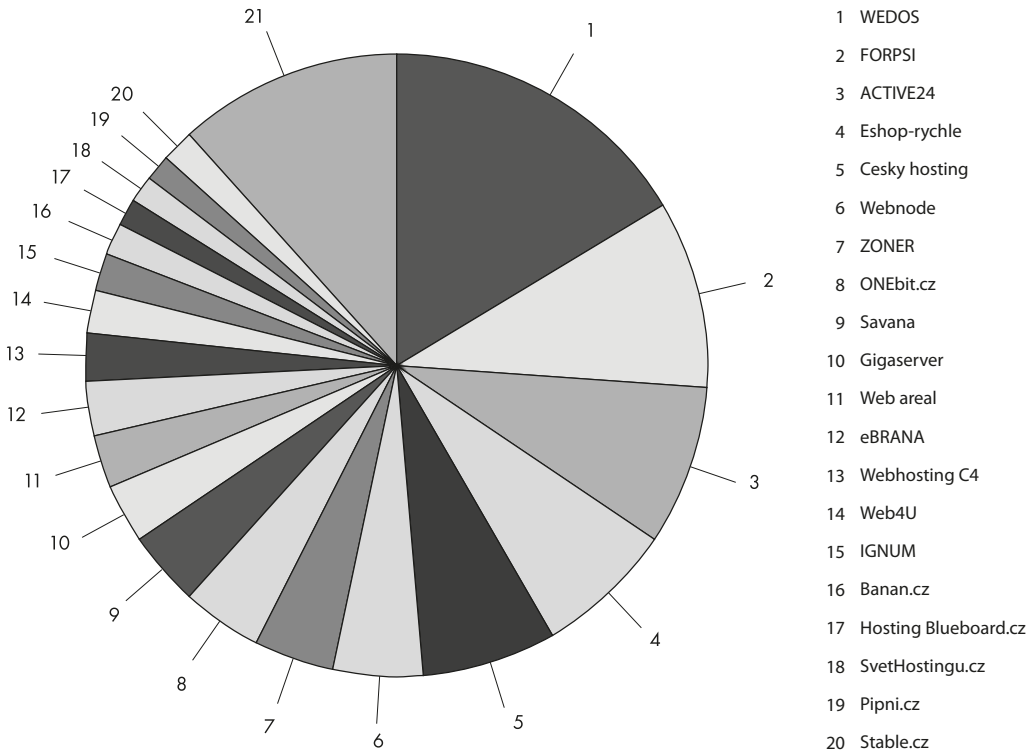
As for validation of the results, we randomly selected 400 websites from our dataset. There was the difference in the hosting status decision only in 14 of them. We may therefore state the classification accuracy of 95%.

**Table 9** Percentage of websites detected as hosted in examined categories

| Category | Evaluated | Percentage of detected hosted websites [%] | | | | |
|---|---|---|---|---|---|---|
| | | Test 1 | Test 2 | Test 3 | Test 4 | Overall result |
| Banks | 36 | 52.78 | 5.56 | 86.11 | 77.78 | 75 |
| E-shops | 11 314 | 91.84 | 13.12 | 99.88 | 99.24 | 98.14 |
| Hospitals | 469 | 89.34 | 18.98 | 99.57 | 96.38 | 95.95 |
| Insurance companies | 26 | 76.92 | 3.85 | 96.15 | 84.62 | 84.62 |
| Real estate agencies | 1 660 | 91.27 | 16.27 | 99.64 | 98.07 | 98.13 |
| Craftsmen | 6 563 | 93.46 | 25.69 | 98.2 | 99.19 | 97.55 |
| Government institutions | 109 | 65.14 | 7.34 | 95.41 | 89.91 | 76.15 |
| High schools | 1 192 | 84.56 | 17.11 | 99.66 | 96.81 | 90.02 |
| Universities | 103 | 60.19 | 13.59 | 90.29 | 70.87 | 64.08 |
| **Total** | **21 472** | **91.35** | **17.54** | **99.19** | **98.63** | **97.01** |

**Source:** Own construction

**Figure 2** Participation of webhosting providers




**Source:** Own construction

### 4.2 Web Hosting Providers

Based on the data from reverse DNS queries we have evaluated the numbers for web hosting providers, listed in Table 10. The numbers shown include the first 20 providers with the most detected hosted webs. The rest of providers (not shown) are summarized as 'Others'. To make it clear, the data with not-detected entries excluded are shown in Figure 3. The number of web servers where no hosting provider was detected is marked as 'Not detected'. This number also includes websites hosted by less known hosting providers that are not listed in Table 4. Our number of 'Not detected' websites is comparable with the 'Unknown' result provided by CZ.NIC in Table 1.

The websites counts for the biggest web hosting providers divided into the defined categories are shown in Table 11.

These data can be compared with the data by CZ.NIC, see Section 3. The contribution of the biggest providers is comparable. Some hosting providers are not mentioned by CZ.NIC since our list includes web hosting provider trademarks instead of the company full legal names. For example, CZ.NIC lists 'Gransy s.r.o.' but the trademark is 'Station webhosting'. The second example is that CZ.NIC uses 'THINline interactive, s.r.o.' and the trademark is 'Cesky hosting'.

**Table 10** Detected websites counts for biggest hosting companies. Web hosting providers are labeled by their trademarks

| Rank | Hosting provider | Webs | Share [%] |
|---|---|---|---|
| 1 | WEDOS | 1 611 | 7.49 |
| 2 | FORPSI | 961 | 4.47 |
| 3 | ACTIVE24 | 801 | 3.72 |
| 4 | Eshop-rychle | 726 | 3.38 |
| 5 | Cesky hosting | 667 | 3.1 |
| 6 | Webnode | 460 | 2.14 |
| 7 | ZONER | 424 | 1.97 |
| 8 | ONEbit.cz | 410 | 1.91 |
| 9 | Savana | 372 | 1.73 |
| 10 | Gigaserver | 301 | 1.4 |
| 11 | Web areal | 289 | 1.34 |
| 12 | eBRANA | 264 | 1.23 |
| 13 | Webhosting C4 | 254 | 1.18 |
| 14 | Web4U | 200 | 0.93 |
| 15 | IGNUM | 192 | 0.89 |
| 16 | Banan.cz | 157 | 0.73 |
| 17 | Hosting Blueboard.cz | 147 | 0.68 |
| 18 | SvetHostingu.cz | 145 | 0.67 |
| 19 | Pipni.cz | 141 | 0.66 |
| 20 | Stable.cz | 137 | 0.64 |
| – | Other providers | 1 159 | 5.39 |
| – | Not detected | 11 693 | 54.36 |

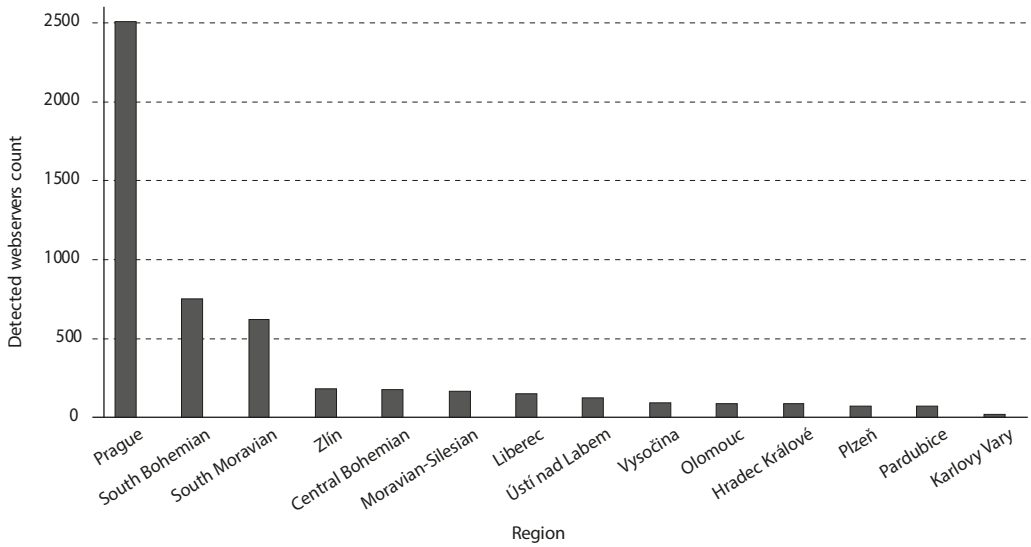
**Source:** Own construction

**Table 11** Share of most common Czech web hosting providers in the examined categories

| Category | Webhosting provider share [%] | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | WEDOS | FORPSI | ACTIVE24 | Eshop rychle | Cesky hosting | Webnode | ZONER | ONEbit |
| Banks | 0 | 0 | 2.78 | 0 | 0 | 0 | 0 | 2.78 |
| E-shops | 6.61 | 2.93 | 2.67 | 6.21 | 3.25 | 0.96 | 1.92 | 1.8 |
| Hospitals | 9.17 | 6.4 | 4.26 | 0 | 3.84 | 3.41 | 3.41 | 2.77 |
| Insurance companies | 0 | 0 | 3.85 | 0 | 3.85 | 0 | 0 | 0 |
| Real estate agencies | 6.99 | 4.94 | 3.86 | 0.12 | 2.59 | 1.33 | 1.45 | 2.11 |
| Craftsmen | 9.23 | 6.92 | 5.53 | 0.32 | 3.08 | 4.46 | 2.19 | 2.03 |
| Government institutions | 1.83 | 0.92 | 2.75 | 0 | 0.92 | 0.92 | 1.83 | 0.92 |
| High schools | 7.63 | 4.95 | 3.78 | 0 | 2.85 | 1.43 | 1.59 | 1.76 |
| Universities | 3.88 | 3.88 | 0.97 | 0 | 0 | 1.94 | 1.94 | 0.97 |

**Source:** Own construction

**Figure 3** Web server location in the Czech regions

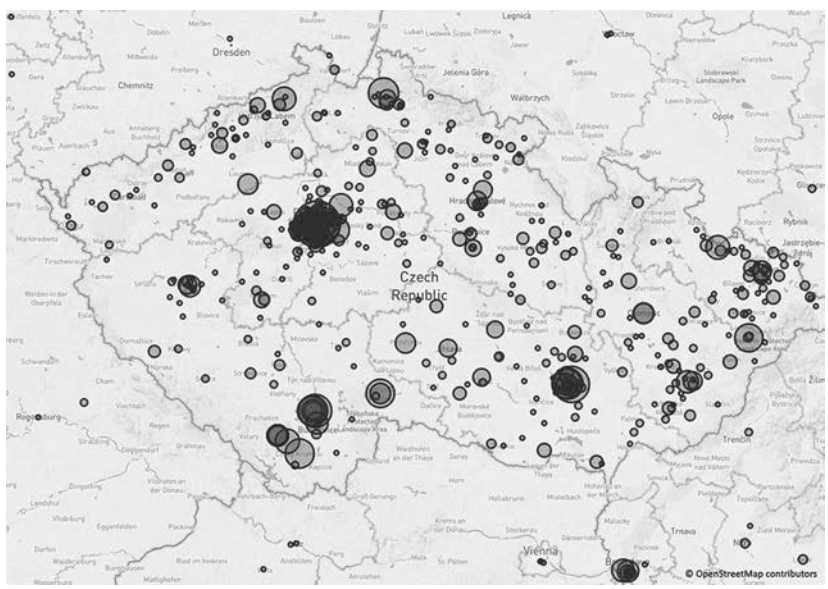



**Source:** Own construction

### 4.3 Geographical related data

The Czech Republic is a small country and the intra-country distance does not have any serious effect on web page loading delay. Almost half of the tested websites were hosted in the Prague region. The second one was the South Bohemian region, where servers of a large provider are situated. A map showing the location of servers in cities is shown in Figure 4.

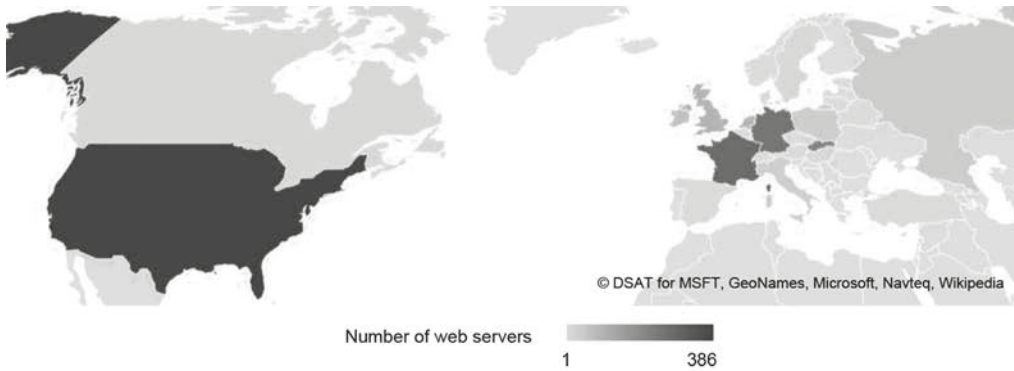
**Figure 4** Web server location in Czech Republic




**Source:** Own construction

Although we focused on Czech domain zone, we detected 1 532 web serves hosted outside the Czech Republic. Most of these websites were hosted in the United States (386), France (274), Germany (247), and Slovakia (205). A map with major countries with the Czech websites hosted is shown in Figure 5.

**Figure 5** World Czech website hosting server location




**Source:** Own construction

The median geographical distances from our university server in Brno to servers with a Czech domain in selected countries are listed in Table 12. For these distances, we estimated the minimal additional round-trip delay caused by data transmission in optical links over these distances. We used a simplified value of 5 us delay per 1 km (Coffey, 2017). As we calculate the minimum additional delay, we omitted the cable links inflation over distances and actual routing paths.

**Table 12** Median distance and minimum additional RTT for servers running websites hosted outside the Czech Republic, rows are sorted by counts of hosted websites in each country

| Country | Median distance | | ExRTT* |
|---|---|---|---|
| | [km] | [miles] | [ms] |
| United States | 7 489 | 4 653 | 75 |
| France | 1 039 | 645 | 10 |
| Germany | 579 | 360 | 6 |
| Slovakia | 122 | 76 | 1 |
| United Kingdom | 1 212 | 753 | 12 |
| Netherlands | 893 | 555 | 9 |
| Ireland | 1 648 | 1 024 | 16 |
| Russia | 1 748 | 1 086 | 17 |
| Poland | 361 | 224 | 4 |
| Italy | 689 | 428 | 7 |

**Source:** Own construction

### 4.4 Other Related Data

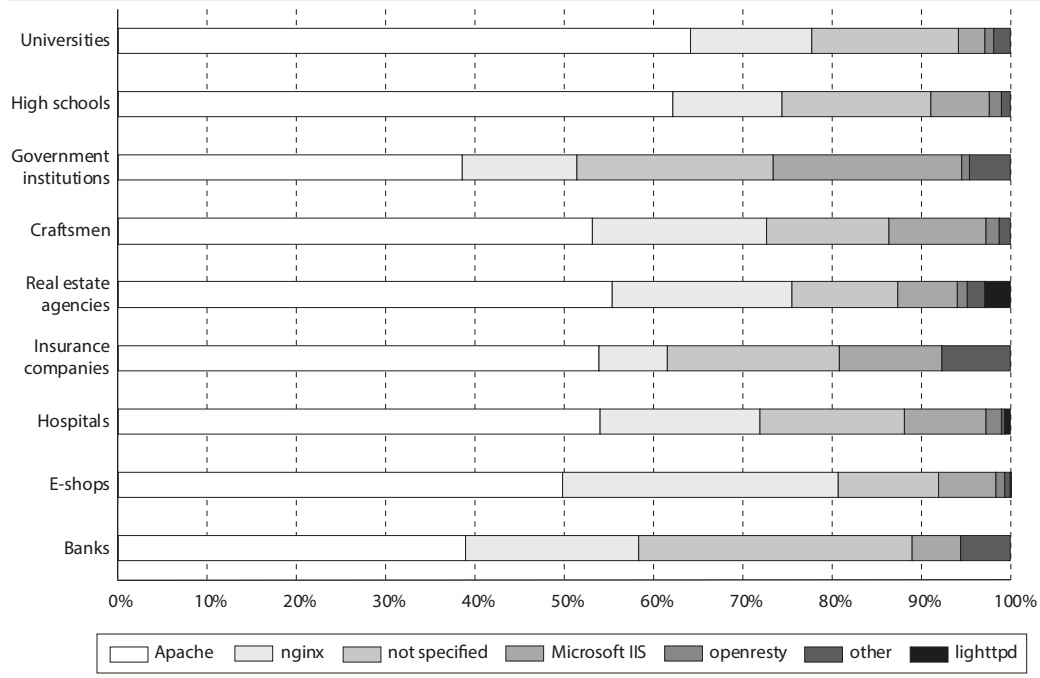
For other related data, we focused on the use of IPv6 and security implementations inthe Czech web servers. We detected the use of IPv6 by checking the existence of a DNS AAAA record and server HTTP-availability. For security implementations, namely HTTPS and DNSSEC, we considered a web server as HTTPS-compliant if there was a positive response for an HTTPS request and, also, a verified certificate was

present. For certificate verification, we used a list of certificates provided in the Debian `ca-certificates` package (2017). We obtained the DNSSEC data (DNSSEC keyset) from the WHOIS domain registry managed by CZ.NIC (2017, CZ domain registry). We considered a web server as DNSSEC-compliant if the keyset was present. We also found the web server software by inspecting the 'Server' field of the HTTP response. We again divided the results into nine organization categories.

The share of web server software is shown in Figure 6. We observed that the use of the free Apache is lower with government institutions and banks. Also, the data show that with the bank category, many web servers hide the information about used software for security reasons. We found that the use of the NGINX software is noticeably higher in the e-shop category.

We also detected version for some of the used software. The oldest detected version was for Apache '1.3.27', released in 2002. The use of such an old version can be dangerous due to known unfixed serious vulnerabilities (CVE Details, 2017).

**Figure 6** Use of web server software




**Source:** Own construction

The use of IPv6 and security implementations is listed in Table 13. We included the numbers for IPv6 as its use is given by support of the hosting provider and, also, by DNS settings maintained by the administrator of the server domain. One could expect that universities would be the leading entity for the use of IPv6. However, the table shows that government institutions have the biggest number – 40%. This is probably given by implementation of the Czech government resolution (2009) about the use of IPv6. Banks and insurance companies have the highest percentage for the use of HTTPS as the entities offer secure services. The use of DNSSEC is the most significant with government institutions (72%, second is 50%). This is again probably given by implementation of the Czech resolution about the use of DNSSEC (2013).

**Table 13** Use of IPv6 and secure implementations by organization categories

| Category | Technology support [%] | | |
|---|---|---|---|
| | IPv6 | HTTPS | DNSSEC |
| Banks | 5.56 | 66.67 | 30.56 |
| E-shops | 29.17 | 32.64 | 39.89 |
| Hospitals | 33.26 | 12.15 | 44.14 |
| Insurance companies | 3.85 | 69.23 | 42.31 |
| Real estate agencies | 24.88 | 14.82 | 49.34 |
| Craftsmen | 33.17 | 12.48 | 42.3 |
| Government institutions | 40.37 | 28.44 | 72.48 |
| High schools | 26.68 | 15.52 | 44.38 |
| Universities | 20.39 | 33.01 | 37.86 |

**Source:** Own construction

## CONCLUSION

The paper presented a method for website hosting detection. The method consists of four partial tests with assigned weights. Large data were collected from the Czech Internet and processed for website statistical analysis. The data came from more than 21 000 websites that we have collected by web-crawling the public resources. We analyzed the websites in nine defined categories according to organizations they represent in their content. We also processed the data geographically to analyze the locations of the web servers. In addition, we focused on HTTPS, DNSSEC, and IPv6 protocols support. The used procedures applied in a country may be of use for marketing purposes, verification of fulfillment of legal directives, and for assessment of claimed web hosting plans.

The particular results detect 97% of the websites as hosted by another organization. The most used software was Apache followed by NGINX. Furthermore, 30 % of the crawled websites were available via IPv6, most of them in the category of government institutions. 24% of websites were available via HTTPS protocol, most of them in the categories of insurance companies and banks. The DNSSEC protocol was supported by 41.8% of the tested domains.

## ACKNOWLEDGMENT

This work was supported by grant NSP LO1401 and the SIX Research Center.

## *References*

BUILTWITH. *Web Hosting Usage Statistics* [online]. Sydney: BuiltWith, 2017. [cit. 5.1.2018]. <https://trends.builtwith.com/hosting>.

BULIN, M. *Analysis of Real Estate Market Using Information on Internet* [online]. Master thesis, Brno: Brno University of Technology, Faculty of Electrical Engineering and Communication, 2017. <http://hdl.handle.net/11012/65811>.

COFFEY, J. *Latency in optical fiber systems* [online]. White Paper, Hickory: Comm-Scope, 2017. [cit. 12.1.2018]. <https://www.commscope.com/Docs/Latency_in_optical_fiber_systems_WP-111432-EN.pdf>.

CVE DETAILS. *Apache 1.3.27 Security Vulnerabilities* [online]. CVE Details, 2017. [cit. 14.1.2018]. <https://www.cvedetails.com/vulnerability-list/vendor_id-45/product_id-66/version_id-8205/Apache-Http-Server-1.3.27.html>.

CZ.NIC. *CZ Domain Hosting Statistics* [online]. Prague: CZ.NIC, 2017. [cit. 2.1.2018]. <https://stats.nic.cz/stats/hosting>.

CZ.NIC. *CZ domain Registry (WHOIS)* [online]. Prague: CZ.NIC, 2017. [cit. 2.1.2018]. <https://www.nic.cz/whois/>.
DEBIAN. *Package: ca-certificates* [online]. Debian, 2017. [cit. 2.1.2018]. <https://packages.debian.org/en/stretch/ca-certificates>.
GOVERNMENT OF THE CZECH REPUBLIC. *Usnesení vlády České Republiky ze dne 8. června 2009 č. 729 ke Zprávě o přechodu na internetový protokol verze 6 (IPv6)* [online]. Prague: Government of the Czech Republic, 2009. [cit. 2.2.2018]. <https://kormoran.vlada.cz/usneseni/usneseni_webtest.nsf/0/6BFDE5B071A154C5C12575E5004024F1/$FILE/727%20uv090608.0727.pdf>.
GOVERNMENT OF THE CZECH REPUBLIC. *Usnesení vlády České Republiky ze dne 18. prosince 2013 č. 982 ke Zprávě o zavádění technologie DNSSEC a o plnění usnesení vlády ze dne 8. června 2009 č. 727 ke Zprávě o přechodu na internetový protocol verze 6 (IPv6)* [online]. Prague: Government of the Czech Republic, 2013. [cit. 2.2.2018]. <https://apps.odok.cz/attachment/-/down/VPRA9EVEBJYC>.
MAXMIND. *GeoIP2 Databases* [online]. Waltham: MaxMind, 2017. [cit. 2.1.2018]. <https://www.maxmind.com/en/geoip2-databases>.
SINGLA, A., CHANDRASEKARAN, B., GODFREY, B., MAGGS, B. The Internet at the Speed of Light. In: *Proceedings of the 13th ACM Workshop on Hot Topics in Networks,* New York: ACM Press, 2014, pp. 1–7.
TSOU, M. AND LUSHER, D. Mapping Web Pages by Internet Protocol (IP) Addresses: Analyzing Spatial and Temporal Characteristics of Web Search Engine Results. In: *Proceedings of the 17th International Symposium on Cartography in Internet and Ubiquitous Environments,* Tokyo: International Cartographic Association, 2015, pp. 1–20.
WANG, Y., BURGENER, D., FLORES, M., KUZMANOVIC, A., HUANG, C. Towards street-level client-independent IP geolocation. In: *Proceedings of the 2011 USENIX Annual Technical Conference,* Portland: USENIX, 2011, pp. 365–379.

## ANNEX – List of web hosting providers considered by CZ.NIC

ACTIVE 24; AERO Trip PRO, s.r.o.; AIVision, s.r.o.; Amazon.com, Inc; Angel hosting; Axfone, s.r.o.; Banan, s.r.o.; Basefarm, AS; BEST-NET, s.r.o.; Blueboard.cz, s.r.o.; Bodis, LLC; business communication, s.r.o.; Bydzovsky, s.r.o.; Casablanca, Int.; Cesky hosting (THINline interactive, s.r.o.); Cesky server.cz, s.r.o.; CESKY WEBHOSTING, s.r.o.; CZOL media interactive, s.r.o.; Datahost, s.r.o.; DOMENY, s.r.o.; EBOLA Czech, s.r.o.; Explorer, a.s.; FlyNetwork, s.r.o.; FORPSI; Fortion Networks, s.r.o.; Gigaserver.cz; Google, Inc.; Gransy, s.r.o; Group NBT, plc; Happy Technik, s.r.o.; HEXAGEEK, s.r.o.; HOSTING90 systems, s.r.o.; HostingSolutions s.r.o; HUMLNET CREATIVE, s.r.o.; IglooNET, s.r.o.; Ignum s.r.o; ISOL Int., s.r.o.; IT Host.CZ, o.s.; KRAXNET, s.r.o.; LTweb s.r.o.; Luvenex plus, s.r.o.; Nethost, s.r.o.; NETIO Solutions, s.r.o.; Netlook, s.r.o.; Next Dimension, Inc.; Nodus Technologies, s.r.o.; OBSIDIAN, s.r.o.; ONEsolution, s.r.o.; OVH; PIPNI, s.r.o.; savana.cz s.r.o.; Savvy, s.r.o.; SecurityNet.cz, s.r.o.; Stable.cz; SuperNetwork, s.r.o.; SvetHostingu.cz; TELE3, s.r.o.; Telefonica O2 Czech Republic, a.s; Topweby; Trellian, Ltd.; UNIHOST, s.r.o.; united-domains, AG; Vas-hosting.cz; Web4ce, s.r.o.; Web4U, s.r.o; Websupport, s.r.o.; Web zdarma s.r.o.; WEDOS Internet, a.s.; W HOSTING, s.r.o.; WinSoft Company, s.r.o.; XHOSTING.CZ group, s.r.o.; ZONER software, a.s.