IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 27, 2017, accepted October 30, 2017, date of publication November 9, 2017,
date of current version December 5, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2771486

Internet Geography and Real Estate Market

DAN KOMOSNY[1], MARTIN BULIN[2], AND PETR ILGNER[1]
[1]Department of Telecommunications, Brno University of Technology, 601 90 Brno, Czech Republic
[2]Czech Telecommunications Infrastructure, 130 00 Prague, Czech Republic

Corresponding author: Dan Komosny (komosny@vutbr.cz)

This work was supported by the National Sustainability Programme under Grant LO1401.

ABSTRACT This paper introduces a new spatial data resource for Internet geography. The idea is based on merging two up-to-now unrelated fields—the real estate market and Internet networking. The huge real estate market is present almost everywhere, and therefore, it is a valuable resource for trusted spatial data when connected to cyber space. We describe a method that gains spatial data from the real estate market and links this data to cyber space. We support the method by a real implementation. Based on the collected data, we identified the geographical scope of a set of Web servers. The new data in cyber geography may help in the Internet Web-related market competition.

INDEX TERMS Cyber geography, Internet, IP address, real estate market, Web.

I. INTRODUCTION

This paper introduces a new spatial data resource for Internet geography, also known as cyber geography. The data is available from the huge real estate market. There are extensive lists of real estates on the Internet provided by national and international real estate web portals. The geographical location is listed for each real estate available as it is the essential information for the potential buyers or renters. Further description of real estates gives information about the related estate agency. Based on our observations, there is a limited distance between real estates and estate agencies.

The number of web servers related to estate agencies is huge. The agencies are found in any larger populated place. We empirically found that estate agencies are present in cities with a population as low as 8,000. The real estate market is also geographically distributed over all inhabited areas as property trading happens globally. For example, there are about 86,000 real estate agencies in the USA with over five million homes sold each year. Particular statistical data for selected countries is available at www.nar.realtor/field-guides/field-guide-to-quick-real-estate-statistics.

Motivation for this paper is to introduce new spatial data for cyber geography. We describe a method that gains spatial data from the real estate market and links this data to cyber space. The method consists of four procedures. We first describe the procedures theoretically. Then we support the method by its implementation on the real estate market in a country. Based on the collected data, we identified the geographical scope of a large set of web servers. The results indicate an inefficient use of Internet resources. The new data in cyber geography may help in the Internet web-related market competition.

The presented analysis particularly focuses on geographical distances between the real location of web servers and their representative content. The role of distance in the Internet has recently grown in its significance as this is one of the factors that degrades networking performance and wastes Internet resources [1]. As a consequence, the future networking concept aims to limit the geographical distances by moving the content closer to the location of users [2]. On the other hand, the current concept of Cloud computing provides resources without primarily considering the proximity between content and users.

The paper is structured into seven sections: In Section II we review related work dealing with geographical mapping of Internet devices with a focus on web servers. Section III highlights the problems that are addressed in this paper. The idea of using the real estate market for cyber geography is introduced in Section IV. The real implementation and related notes are presented in Section V. Section VI describes the spatial data obtained from the real estate market. A summary of advantages and disadvantages of using the data for cyber geography is given in Section VII.

II. RELATED WORK

Research into cyber geography is mainly driven by a demand for optimization of Internet resources, providing better location-aware services, and gaining advantages in market competition. Cyber geography studies the correlation of various types of Internet devices and their real locations.

Generally, Internet devices are located by their unique IP addresses that are independent of their hardware and software features.

There are several methods for geographical mapping of Internet devices. The methods include analysis of information from network registrars (such as ARIN and RIPE NCC) [3], network data monitoring and analysis [4]–[7], detection of Points of Presence [8], [9], and other combinations that use population data and specific geographical regions [10], [11]. The methods differentiate in their accuracy and coverage. The method presented in [5] works for Internet devices in large cities (i.e. places with high population). The method introduced in [4] works for Internet devices that respond to communication latency measurements. A different method is based on data collected from mobile devices that report their geographical location (typically obtained from GPS, BTS or WiFi triangulation). The reported location is linked to the device public IP address. This method is applicable on mobile devices only and is inaccurate when NAT (Network Address Translation) is used [12].

Paper [13] studies mining the content of web servers for geographical mapping. It describes a method consisting of three parts. The most relevant part to this paper is the first one that covers extraction of positioning information from web pages of a server. The found location is assigned to the server's IP address. The position extraction algorithm is based on regular expressions for pattern matching. The location information extracted is weighted and the weights are assigned according to the relative position of the information on a page. Based on the authors' empirical observations, they conclude that on Chinese web servers the correct information is most likely situated at the bottom of a page. However, in other countries the correct location may be placed elsewhere. The authors also exclude web servers that can not be trusted for determining the location. They remove web servers containing keywords, such as 'blog, forum, bbs'. They also remove pages with a large number of positioning information. They assume that the IP addresses in blocks of /24 network size are close to each other (i.e. in one city and very unlikely to be in a different city). Also, the web hosting related to this paper is studied in [13]. When a web server is not presented locally at the company, the location obtained does not correspond with its true location. A single IP address is typically used for a number of hosted web servers and their domains. The authors detect a web hosting by checking the number of domain names for IP addresses.

Paper [14] introduces a three-tier methodology for IP geolocation in large cities. The authors base their methodology on two observations i) a considerable number of companies run their web servers locally and ii) relative network delay heavily correlates with geographical distances. The relevant information to this paper is that they get a list of the web servers by querying a mapping service. They specify the required area and the keywords, such as 'business or university'. The locations are taken from the results returned from the mapping service. The authors also focus on the web hosting problem. They access the webs twice using its domain name and IP address. If the content or just page heads/titles are the same for both accesses, they assume that the web server is not hosted.

III. PROBLEMS ADDRESSED

In this paper, we address these general problems from related work where the geographical locations are derived from the content of web pages:

i) When positioning information is found on the pages of a web server (or even on a single page), this location might be or might not be the true location of the web server.

ii) When more than one positioning information is found on the pages of a web server, there is an uncertainty which one should be used as the location of the web server. The first or last information on a page or their weighted combination is used. Again this location might be or might not be the location of the web server.

iii) When there is no positioning information on the pages of a specific web server, its location can not be determined.

IV. SPATIAL DATA FROM REAL ESTATE MARKET FOR CYBER GEOGRAPHY

This section describes the use of spatial data. The whole method consists of four procedures: i) crawling, ii) grouping, iii) geography, and iv) analysis. The overview of the procedures is shown in Fig. 1. We describe them theoretically first. A real implementation of the procedures is described in the next section.

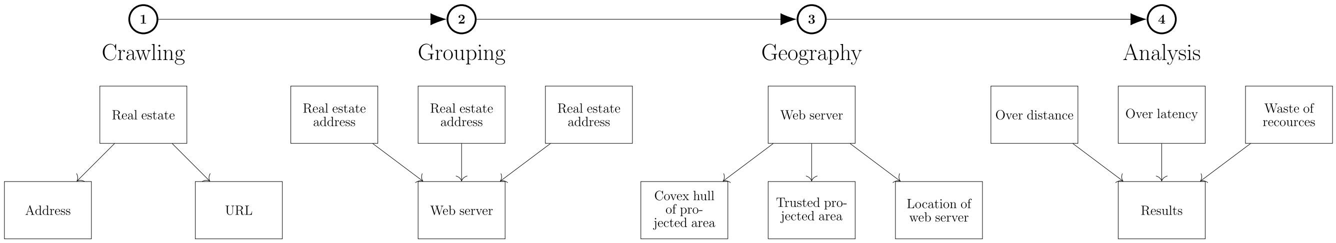


FIGURE 1. Spatial data from real estate market for cyber geography.

A. CRAWLING

The Internet is web-crawled for real estate data. There are huge web portals with extensive lists of real estates distributed worldwide or within specific countries. A first-level crawling identifies the web portals with lists of real estates. Within the domain of the found web portals, a second-level crawling identifies the particular real estates. Each real estate is looked-up for its title, postal address (geographical location) and URLs in its content. The output of the crawling procedure is a collection of web portals, real estates, postal addresses, and URLs.

B. GROUPING

The crawled real estates are grouped into sets according to the same found URLs in their content. A real estate agency web server is identified for each group by the same shared URL link. Such a URL can be placed anywhere on the page of a real estate and its specific location does not need to be known. The output of the grouping procedure is a collection of web servers. The postal addresses of real estates in each identified set are associated with the web server found for this set.

C. GEOGRAPHY

This procedure covers the geographical aspects. Four main aspects are considered: i) By a known list of postal addresses for each identified set from the grouping procedure (i.e. geographical locations representing the content of a web server) the web-content projected area is derived for each web server. This area is defined as a convex hull around the peripheral real estate locations. ii) The defined convex hull does not delimit the whole web-content projected area as only the current real estates are included at the time of crawling. Another extended area is therefore used to correctly identify the whole web-content area. This area is derived as an extension of the convex hull by a given algorithm, depending on the implementation. iii) The real location of a web server is derived depending on the web hosting status. iv) The identified location of a web server is geographically mapped to its web-content projected area.

D. ANALYSIS

This part covers processing of the spatial data gained from the previous procedure. We demonstrate the use of the data by an analysis of the current status of web hosting and its effect on cyber geography. The result is a study of distances that are outside the web-content projected area. Such geographical distributions have a negative impact on networking performance and Internet market competition, as we discuss later.

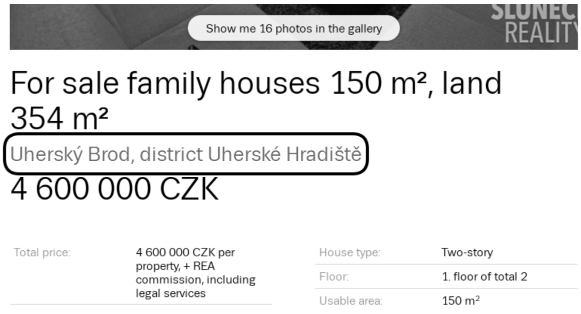
The procedures may be repeated to reflect the changes in the real estate market and web server locations. Regarding the real estates, the changes are mainly due to new estates on the market. For the web servers, changes in allocation of IP addresses and in web server hosting are the main concerns for updates.

V. IMPLEMENTATION AND REAL-CASE DATA

In this section, we describe the implementation of using spatial data from the real estate market for cyber geography. The implementation is divided according to the method procedures as defined in the previous section. Results from the *Analysis* procedure are presented separately in Section VI.

The *Crawling* procedure is implemented by a first-level crawling of the web for general URL links and analysing the content of found web pages. A real estate web portal is identified by looking for specific keywords, such as 'real, estate, property, sale, price, flat, house'. Worldwide or national real estate portals may be found. An example of a worldwide portal is www.realtor.com and national www.zoopla.co.uk. When a real estate portal is found, its web pages are second-level crawled to find the real estates advertised. These pages are analysed for specific keywords, such as 'reality, price, sale, $, USD, Euro, €, area'. To help the identification of real estates, a specific portal sub-domain may be identified. Examples of such specific sub-domains are '../property/, ../for-sale/.., ../buy/.., ../realestateandhomes/..'. We used the Selenium webdriver for the web crawling and the Python BeautifulSoup module for HTML data extraction. For each real estate found, its postal address and URLs on the page are extracted. The postal address is detected from the real estate title as the first text that can be geocoded to geographical coordinates. We used the Python Geopy module for geocoding of postal addresses.

Fig. 2 shows an example of a real estate found with the postal address highlighted that can be geocoded to geographical coordinates. The URLs are extracted from the whole content of a real estate page. Fig. 3 shows an example of the same real estate with a found URL highlighted. The found real estates, their portal domain, postal addresses, and URLs are stored in a database.



For sale family houses 150 m², land 354 m²

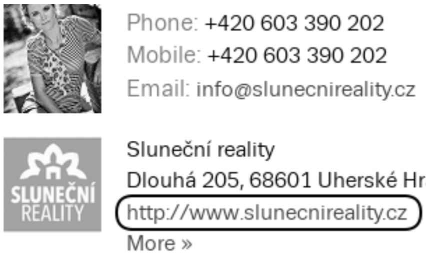
Uherský Brod, district Uherské Hradiště

4 600 000 CZK

Total price:	4 600 000 CZK per property, + REA commission, including legal services	House type:	Two-story
		Floor:	1. floor of total 2
		Usable area:	150 m²

FIGURE 2. Address of real estate.

The *Grouping* procedure is implemented by grouping the found real estates into sets based on the same URLs links in their pages. The specific position of the grouping URL link on a page may vary due to different formats used (in the previous example it is at the bottom of the page). Each URL link is verified to be a valid link to a web server. The URL links shared by all the real estates within the same portal domain are excluded as they might be misinterpreted as being



Phone: +420 603 390 202
Mobile: +420 603 390 202
Email: info@slunecnireality.cz

Sluneční reality
Dlouhá 205, 68601 Uherské Hradiště
http://www.slunecnireality.cz
More »

FIGURE 3. Shared URL link to web server; position may vary due to different page format.

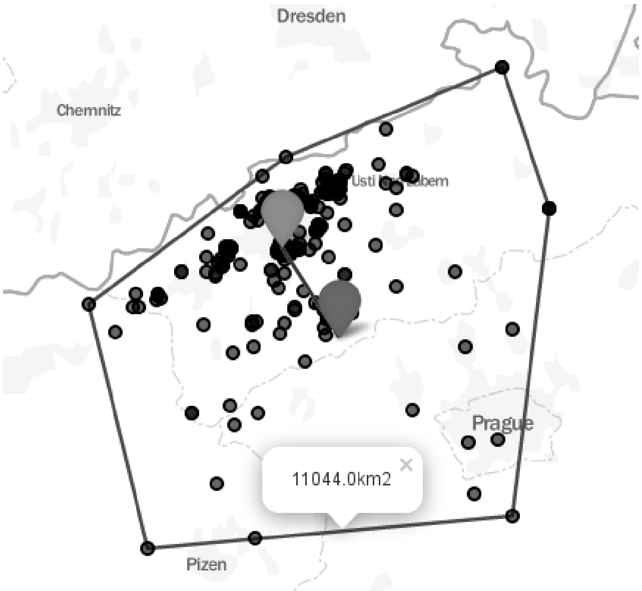
valid links, yet are most likely advertisements. Also, some real estate agencies 're-take' real estates from other agencies. Both cases would result in a very large number of wrong URL links in an identified set. We empirically found a limit of 500 shared URLs to correctly identify valid links to web servers. The postal address of a real estate agency may also be presented on a page. This address is found as the shared address in all the pages within an identified agency set.

The *Geography* procedure is implemented by association of the real estate postal addresses with the identified web servers. The addresses are geocoded into latitude and longitude. Various free geocoding services may be used for this purpose, such as those provided by the Python Geopy package that we used in the implementation. We specifically used the free OpenStreetMap Nominatim wiki.openstreetmap.org/wiki/Nominatim. The free geocoding providers limit the number of requests per time unit. The commercial providers extend these limits, such as Google geocoding API developers.google.com/maps/documentation/geocoding/usage-limits.

Fig. 4 shows geocoded postal addresses of the crawled real estates for an identified set. The dots show their locations. A convex hull is created around the found peripheral locations. We used the Python Geopandas module for this geographical-related data processing (and the other following). The area centroid is calculated and it is shown as a dark sign. The found address of the real estate agency is shown as a bright sign. The figure shows that the most real estates are within a small proximity of the agency location (bright sign). This demonstrates that the real estate agencies advertise their estates within a short distance due to their practical logistics.

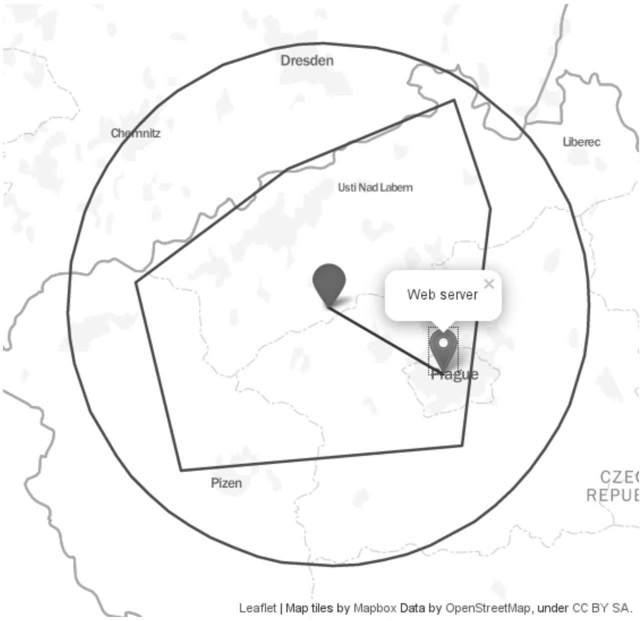
The convex hull does not delimit the whole web-content projected area as only the current real estates at the time of crawling are included. In order to define the whole web-content area, we extend the convex hull by a bounding box given by the greatest distance from minimal or maximal x,y values to the area centroid. By this value, the radius of a circle is given that defines the extended area. Fig. 5 shows such an area for the same set of real estates as in the previous example.

The next part of the geography procedure deals with the location of web servers found for each set of real estates. We differentiate two options – whether the server is hosted or not. Identification of a web server hosting is not straightforward. We therefore combine four partial tests in our implementation: i) 'Reverse domain lookup', ii) 'Known hosting



Leaflet | Map tiles by Mapbox Data by OpenStreetMap, under CC BY SA.

FIGURE 4. Example of crawled real estate data. Dots are geocoded postal addresses of real estates within identified set; area centroid (dark sign) and found address of real estate agency (bright sign) are also shown.



Leaflet | Map tiles by Mapbox Data by OpenStreetMap, under CC BY SA.

FIGURE 5. Example of extended projected area (circle). Location of web server and its relation to area centroid is shown.

provider', iii) 'Domain holder and network name', and iv) 'Contact email'. We consider a web server being hosted by default and the particular tests have to prove the opposite.

Ad i) The test 'Reverse domain lookup' compares the web server domain name and the domain name of IP address of the same web server. The reverse lookup is based on the PTR DNS record (pointer to canonical name) that maps IP address to host name. PTR records are maintained by ISPs or organizations with own IP address space. For not-hosted servers, it is common that their IP address has the PTR record set to their domain name. Such configuration is used for mail servers as many mail agents require the mail source IP address to be resolved to a domain name and this domain name to be resolved to the same IP address. If this is true, the web server is not probably hosted. Fig. 6 shows an example for a web server from our crawled data – www.slunecnireality.cz with IP address 94.177.163.27. The test result is that the server is not hosted as the domain name is the same as the name from the reverse IP address lookup.

```
[]$ nslookup www.slunecnireality.cz
Non-authoritative answer:
Name: www.slunecnireality.cz
Address: 94.177.163.27
               |
               |
[]$ nslookup 94.177.163.27
Server:    192.168.1.1
Address:   192.168.1.1#53

Non-authoritative answer:
27.163.177.94.in-addr.arpa   name = host.slunecnireality.cz.
```

FIGURE 6. Reverse domain lookup test.

Ad ii) The test 'Known hosting provider' is based on the known IP address ranges of web hosting providers. If the IP address of a web server falls in such known range, the web server is very likely to be hosted. For the same example – www.slunecnireality.cz with IP address 94.177.163.27 – the IP address falls into such a range of IP addresses (94.177.163.0 – 94.177.163.255) of a known web hosting provider. The page www.aruba.it indicates that such IP address range belongs to a web hosting company Aruba S.p.A.y. Fig. 7 demonstrates this test with a result that the web server is hosted.

```
[]$ whois 94.177.163.27
inetnum:          94.177.163.0 - 94.177.163.255
geoloc:           43.45997095884493 11.837875843048096
netname:          ARUBA-NET
country:          IT
...
abuse-mailbox:    abuse@staff.aruba.it
```

FIGURE 7. Known hosting provider test.

Ad iii) The test 'Domain holder and network name' is based on the similarity of the organization name specified in the regional registries (domain holder) and the name of the network. If these two names are similar, the web server is likely not to be hosted. The names most probably will never be exactly the same. The reason is that company names typically include a suffix specifying the company type, such as Aruba 'S.p.A.'. On the other hand, the network name may also include a prefix or suffix, such as ARUBA-'NET'. To exclude suffixed and prefixes in the names, we evaluate the similarity of the names and not the precise match. Fig. 8 shows the test for www.slunecnireality.cz. The domain name holder specified by the regional registrar RIPE NCC is '4DEVELOPMENT'. The network name for the IP address of the web server is 'ARUBA-NET'. This test indicates that the web server is hosted.

```
[]$ whois slunecnireality.cz
domain:          slunecnireality.cz
registrant:      24DEVELOPMENT
                        |
                        |
[]$ whois 94.177.163.27
inetnum:            94.177.163.0 - 94.177.163.255
geoloc:             43.45997095884493 11.837875843048096
netname:            ARUBA-NET
```

FIGURE 8. Domain holder name and network name test.

Ad iv) The fourth test 'Contact email' explores the email address found in the registrar database for a domain. This email is typically a contact to the network administrator. If the domain part of the email address is the same as the domain of the web server investigated, the server is likely not to be hosted. Fig. 9 shows a test example for the web server www.slunecnireality.cz. This test shows that the web server is hosted as the contact email is different to the domain name of the web server.

```
[]$ whois slunecnireality.cz
domain:          slunecnireality.cz
registrant:      24DEVELOPMENT
...
e-mail:          webmaster@24d.cz
```

FIGURE 9. Contact email test.

TABLE 1. Test results for www.slunecnireality.cz.

Method	Hosted	Not-hosted
Reverse domain lookup	-	X
Known hosting provider	X	-
Domain holder and network name	X	-
Contact email	X	-

Table 1 summarizes the results for the example web server www.slunecnireality.cz as being hosted. Our data shows that the vast majority of web servers are hosted when the real estate market is considered (details are given in Section VI). Therefore, the real location of the web servers have to be considered for valid spatial data. We implemented this by the use of geolocation databases. These databases form blocks of consequent IP addresses and a geographical location is assigned with each block. These locations are obtained from various sources, such as from mobile devices with GPS, network communication analysis, or web crawling. We particularly used the MaxMind GeoIP2 database to locate the web servers by their IP addresses. For fast processing we used the database locally with API provided by the Python geoip2 module.

Fig. 10 shows the location result for the example web server www.slunecnireality.cz. The figure particularly shows the convex hull created from the real estates found for this web server, the extended web-content projected area, and

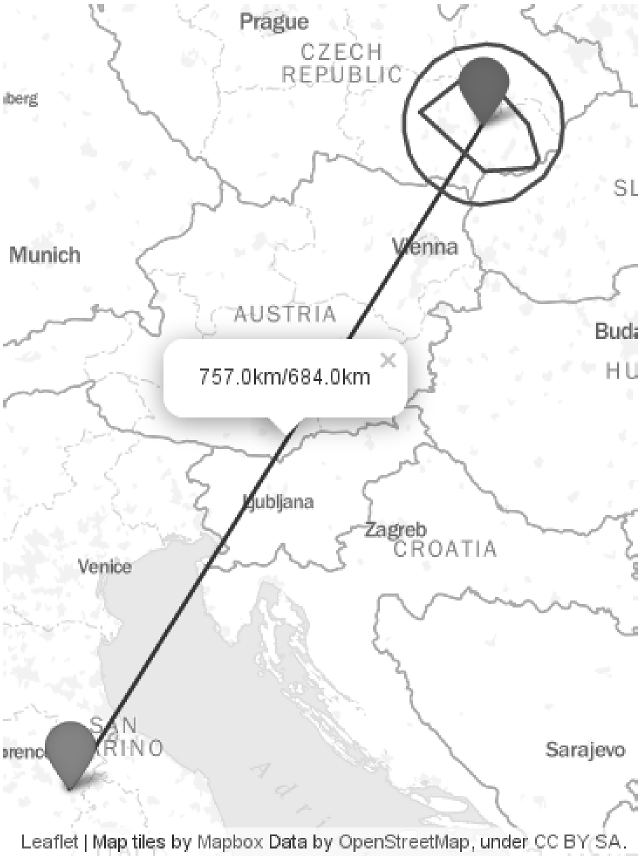


FIGURE 10. Example results for www.slunecnireality.cz – web-content projected area (polygon), extended web-content projected area (circle), found location of hosted web server, and distances to web server from area centroid and border of extended projected area.

the found server location (in Italy). The server is outside the extended area with the total distance from the centroid of 757 km and the over-distance from the area border of 684 km. We give a complex analysis of these geographical distributions in the following section.

VI. RESULTS

This section presents an application of data from the real estate market. Web-content geographical proximity is analysed and related to the real location of web servers. Such knowledge may be generally used for improving the hosting decisions [15] and study of the parts that contribute to Internet communication latency with the focus on its reduction [16]. The data comes from the method introduced in Section IV with the implementation described in Section V.

By the web-crawling procedure, we obtained more than 66,000 real estates in a country. The number of cities included in the data was around 600. The detailed numbers are given in Table 2. Fig. 11 shows the geographical distribution of the collected real estates.

By the grouping procedure, we obtained around 2,000 valid web servers with identified sets of real estates. The hosting test of the web servers showed that they are very commonly hosted. We obtained that only 2 % of the web servers were not-hosted. If we compare this data with other published numbers, we got a quite strict number. The statistical data provided by a national top-level domain name registrar [17] shows that the number of web servers definitely not-hosted has been fluctuating around 5 % in the last seven years. This number is given by inspecting the web server IP addresses against the RIPE NCC WHOIS database. The known IP address ranges of the hosting providers are also included in these statistics. ‘Hosting and Cloud Study 2015’ [18] presents a survey of about 1,500 organizations. 39 % respondents dedicate a significant portion of the application hosting spending for e-commerce website hosting and advanced functionality. The large difference between these numbers may be given by inclusion of different types of companies in the statistics. The large companies prefer to run their own web servers and carry out their own technical and administrative work. However, generally smaller companies, such as real estate agencies, prefer to outsource the technical and administrative work by having web servers hosted.

TABLE 2. Collected spatial data from real estate market.

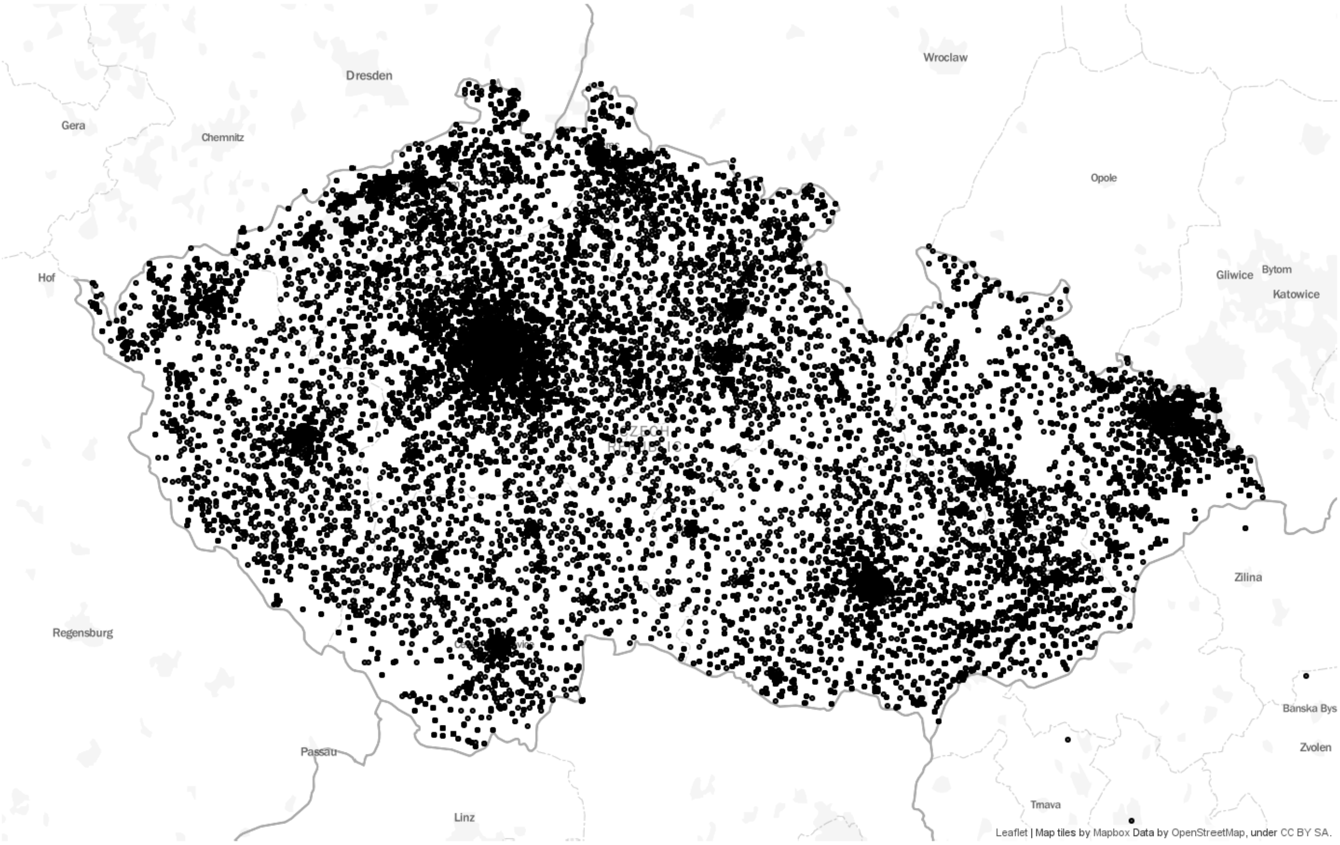
Real estates with location	66392
Real estate agencies	2488
Web servers	2094
Cities	583

As the result of the cyber geography procedure, we obtained around 1,700 valid web-content projected areas. Each area was specified as the convex hull created from the found locations of real estates for a web server. The extended web-content area was created by the bounding boxes. The area-related details are shown in Table 3. The corresponding cumulative probability for the whole range of values is shown in Fig. 12.

TABLE 3. Results for 1,733 web-content projected areas.

	Mean	Median
Projected area (convex hull) [km2]	8733	1060
Extended projected area (circle) [km2]	239354	7227

In the analysis procedure, these particular geographical aspects were evaluated: i) distance from the web-content area centroid to agency location, ii) distance from the web-content area centroid to extended area border, and ii) distance from the web-content area centroid to the real location of web server. The results are shown in Table 4. The first value of 38 km shows that real estate agencies tend to advertise real estates with a small proximity to the location of their office. The reason may be seen in the agent practical travel. It also proves that the content of their web servers has a local significance. Next value is the distance difference between ‘area centroid and web server’ and ‘extended area border and web server’. Values of 85 and 611 km respectively show



Leaflet | Map tiles by Mapbox Data by OpenStreetMap, under CC BY SA.

FIGURE 11. **Geographical distribution of collected real estate market; about 66,000 real estates.**

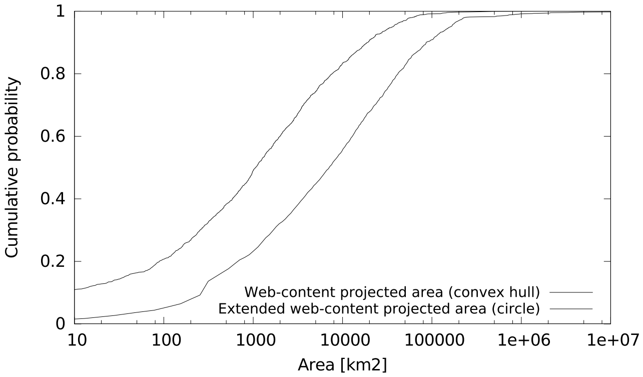


FIGURE 12. **Web-content projected area of web servers.**

TABLE 4. **Area analysis results.**

	Mean	Median
Distance proj. area centroid to agency location [km]	**38**	18
Distance proj. area centroid to extended border [km]	85	48
Distance proj. area centroid to web server [km]	611	89
Over-distance to web server [km]	**994**	125

that there are significant over-distances (outside the extended area) to web servers. An example of such an over-distance (684 km) is shown in Fig. 10. On the other hand, some web servers are located within the extended area as shown in Fig. 5. However, as the difference of the values indicates, the average over-distance of 994 km is significant. The large over-distances from some web servers to their representative content are shown in Fig. 13.

The numbers shown in Fig. 14 indicate an inefficient use of Internet resources by generating load on intermediate devices and bandwidth depletion. We particularly focused on communication latency as it is one of the most important aspects of network communication that drives today's Internet research. Some studies have found a significant relation between communication latency and revenue of companies. For example, Microsoft and Google experimented with injecting additional latency before returning web pages from their servers to the customers. As a result, there were losses in the advertising revenue and the number of searches. When the injected delay was removed, the original numbers returned [16]. For companies offering goods and services on the web, a communication latency difference may play a role in the market competition. Work [19] analysised the partial Internet latency sources. The experiments with accessing the web servers showed that the median speed for fetching the web pages was c/32. To relate the results, we use the corresponding value of 10 km/ms. The average over-distance of 1000 km means that about

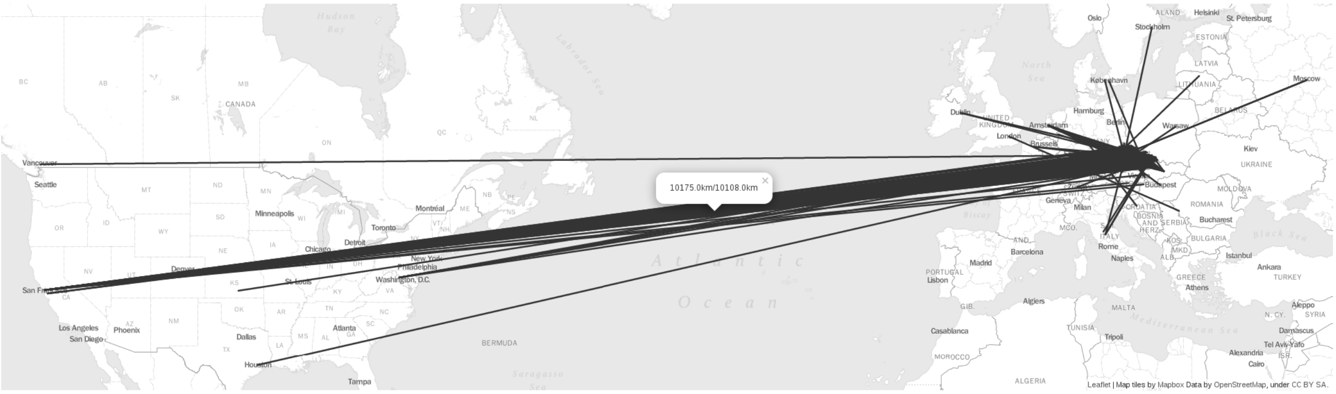


FIGURE 13. Global picture of web server location with large over-distances from their content-projected area.

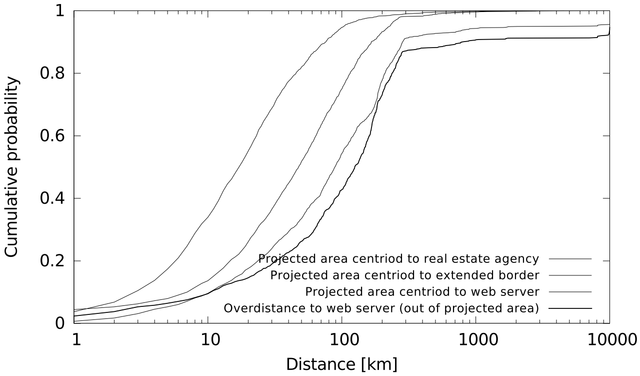


FIGURE 14. Geographical distances derived from web-content projected areas. Over-distance (bold) mean waste of Internet resources and additional delay.

100 ms delay is introduced for some web servers that are located outside their content-projected areas.

VII. PROS AND CONS OF USING DATA FROM REAL ESTATE MARKET

Advantages:

- The spatial data from the real estate market can be fairly trusted as it is directly linked to the true locations of real estates. Where the property is located is essential information for the potential buyers or renters. It eliminates errors of other spatial data sources used, such as crawling the web servers for locations on their pages.
- The real estate market is widespread in populated places, thus giving a good geographical coverage.
- Implementation of web crawling for this data is straightforward.

Disadvantages:

- A lot of geocoding is needed (converting postal addresses to geographical coordinates). There are some free geocoding services available with a limited number of requests per a time unit (typical days and seconds). The web crawling may be prolonged due these limits as geocoding is used to find the postal address of each real estate. The paid services offer higher lookup rates, such Google geocoding API developers.google.com/maps/documentation/geocoding/usage-limits, ArcGis Esri World Geocoding Service API developers.arcgis.com/features/geocoding/.
- It takes a long time to crawl the web for real estates due to the common use of JavaScript for these web sites. The web driver needs to download the whole page for further processing.

VIII. CONCLUSION

The paper was to introduce a new source of spatial data for cyber geography. The real estate market is huge, worldwide, and the spatial data can be fairly trusted as the location is essential for each real estate.

ACKNOWLEDGMENT

Infrastructure of the SIX Center was used.

REFERENCES

[1] M. Kumar, "Various factors affecting performance of Web services," *Int. J. Sensor Appl. Control Syst.*, vol. 3, no. 2, pp. 11–20, 2015.

[2] K. C. Okafor, I. E. Achumba, G. A. Chukwudebe, and G. C. Ononiwu, "Leveraging fog computing for scalable IoT datacenter using spine-leaf network topology," *J. Elect. Comput. Eng.*, vol. 2017, Apr. 2017, Art. no. 2363240. [Online]. Available: https://www.hindawi.com/journals/jece/2017/2363240/

[3] V. N. Padmanabhan and L. Subramanian, "An investigation of geographic mapping techniques for Internet hosts," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 31, no. 4, pp. 173–185, 2001.

[4] S. Laki, P. Mátray, P. Hága, T. Sebők, I. Csabai, and G. Vattay, "Spotter: A model based active geolocation service," in *Proc. IEEE Conf. Comput. Commun.*, Apr. 2011, pp. 3173–3181.

[5] E. Katz-Bassett, J. P. John, A. Krishnamurthy, D. Wetherall, T. Anderson, and Y. Chawathe, "Towards IP geolocation using delay and topology measurements," in *Proc. 6th ACM SIGCOMM Conf. Internet Meas.*, 2006, pp. 71–84.

[6] B. Huffaker, D. Plummer, D. Moore, and K. Claffy, "Topology discovery by active probing," in *Proc. IEEE Symp. Appl. Internet (SAINT) Workshops*, Jan./Feb. 2002, pp. 90–96.

[7] P. Mátray, P. Hága, S. Laki, I. Csabai, and G. Vattay, "On the network geography of the Internet," in *Proc. IEEE INFOCOM*, Apr. 2011, pp. 126–130.

[8] Y. Shavitt and N. Zilberman, "Improving IP geolocation by crawling the Internet PoP level graph," in *Proc. IEEE IFIP Netw. Conf.*, May 2013, pp. 1–9.

[9] Y. Shavitt and N. Zilberman, "A structural approach for PoP geo-location," in *Proc. INFOCOM IEEE Conf. Comput. Commun.*, Mar. 2010, pp. 1–6.

[10] B. Eriksson, P. Barford, J. Sommers, and R. Nowak, "A learning-based approach for IP geolocation," in *Proc. 11th Int. Conf. Passive Active Meas.*, 2010, pp. 171–180.

[11] B. Wong, I. Stoyanov, and E. Sirer, "Octant: A comprehensive framework for the geolocalization of Internet hosts," in *Proc. 4th USENIX Conf. Netw. Syst. Design Implement.*, 2007, pp. 313–326.

[12] S. Triukose, S. Ardon, A. Mahanti, and A. Seth, "Geolocating IP addresses in cellular data networks," in *Proc. 13th Int. Conf. Passive Active Meas.*, vol. 7192. 2012, pp. 115–167. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-28537-0_16

[13] C. Guo, Y. Liu, W. Shen, H. J. Wang, Q. Yu, and Y. Zhang, "Mining the Web and the Internet for accurate IP address geolocations," in *Proc. IEEE Int. Conf. Comput. Commun.*, Apr. 2009, pp. 2841–2845.

[14] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic, and C. Huang, "Towards street-level client-independent IP geolocation," in *Proc. 8th USENIX Conf. Netw. Syst. Design Implement.*, 2011, pp. 365–379.

[15] D. Sanghi, P. Jalote, and P. Agarwal, "Using proximity information for load balancing in geographically distributed Web server systems," in *Proc. 1st EurAsian Conf. Inf. Commun. Technol. (EurAsia-ICT)*, 2002, pp. 659–666.

[16] B. Briscoe *et al.*, "Reducing Internet latency: A survey of techniques and their merits," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 2149–2196, 3rd Quart., 2016.

[17] CZ.NIC. (2017). *Webhosting Statistics*. Accessed: Sep. 2017. [Online]. Available: https://stats.nic.cz/stats/hosting/

[18] Microsoft. (2015). *Hosting and Cloud Study 2015. Beyond Infrastructure: Cloud 2.0 Signifies New Opportunities for Cloud Service Providers*. Accessed: Sep. 2017. [Online]. Available: http://download.microsoft.com/download/3/0/B/30B07B25-05CA-46B2-BCF6-1DE845383672/MSFTHostingEnd-UserFullSlideSet3.17.15Complete.pdf

[19] A. Singla, B. Chandrasekaran, P. B. Godfrey, and B. Maggs, "The Internet at the speed of light," in *Proc. 13th ACM Workshop Hot Topics Netw.*, 2014, pp. 1–7.

DAN KOMOSNY received the Ph.D. degree in teleinformatics in 2003. He is currently an Associate Professor with the Department of Telecommunications, Brno University of Technology, Czech Republic. He leads courses dealing with IP networking and network operating systems. His research is focused on cyber geography and cyber security.

MARTIN BULIN received the master's degree in teleinformatics from the Brno University of Technology. He is currently a System Administrator with Czech Telecommunications Infrastructure, Czech Republic. He focuses on data mining from public Internet resources.

PETR ILGNER received the master's degree in teleinformatics in 2017. He is currently pursuing the Ph.D. degree with the Department of Telecommunications, Brno University of Technology, Czech Republic. He specializes in Web technologies. He leads practicals of a course dealing with IP multimedia systems.