IP ADDRESS AND AUTONOMOUS SYSTEM DIVERSIFICATION AS AN IMPORTANT FACTOR FOR BUILDING GOOGLE RANKING

Artur Strzelecki

University of Economics in Katowice, Department of Informatics,
artur.strzelecki@ue.katowice.pl

Abstract

The article contains an analysis of the influence of diversification of IP addresses on building a ranking of search results in Google search engine. The article is based on an analysis of the process of assigning addresses to website hosting providers and the possible variations of that process. It presents the process of assigning IP addresses and their operation, and it analyses the influence of autonomous systems on web search engine ranking.

Key words: ip address, hosting, autonomous system, search engine,

1   Background

For over 20 years search engines become very important services in life of many people. They have been evolved from very simple services, similar to directories, to very complex online services. Search engines each year takes into account new signals which are important to evaluate webpages.

Web search engines are an important aspect of using public Internet. They allow users to quickly find content for which they are looking. Web search engines build their ranking based on numerous signals provided by websites.

Evans [1] described insight into the techniques that ensure a page ranks highly in Google. Recognized the importance of PageRank and links as well as directories and social bookmarking sites. Moreno and Martinez [2] made a reasearch which confirmed that search engines will interpret the web accessibility achieved as an indicator of quality and will be able to better access and index the resulting web content.

The object of the article is to verify the hypothesis that signals provided by global Internet infrastructure in the form of IP addresses and autonomous systems have a significant influence on building Google search ranking.

2   IP Address

IP is an acronym of Internet Protocol. IP addresses connect all devices capable of communication into one network. The connection may be a local one through the Intranet (home/company network) or commonly known Internet (global network). An IP address is a network address through which the device connected to the network can be accessed. An IP address is the fundamental element of every computer network. Without it Internet in its current form would not be available.

IPv4 addresses are a commonly used addressing method. An IPv4 address is composed of 4 numbers separated by a period [6]. Each of the numbers is in the range from 0 to 255. An IP address is treated as a 32-digit binary number [4].

Every device connected to a network is assigned a certain IP address. If the device is connected using a DSL modem, a direct IP address is most frequently assigned to the device. However, if the device is connected through a router, the Internet provider assigns a relevant IP address to the router, which managers devices connected to it. A web hosting provider also has IP addresses at its disposal which are used to launch its own servers.

One server with an IP address may host numerous Internet domains. Internet search engines send Internet robots through the network. The robots behave in a similar manner as all computers connected to the network. The Internet robot visits domain names by sending queries to DNS servers regarding the IP address of the visited domain name. Therefore, DNSs are an important component of the Internet—the network in its current form would not exist without them.

It is commonly understood that the first three numbers of an IP address constitute class C [8]. If the first three numbers differ from one another, they create a unique class C. According to the aforementioned approach, there are 16,777,216 unique class C networks. This is the current understanding of class C.

Each URL such as www.domain.com may be assigned to 4 various categories:

1. Hostname
2. Domain
3. IP address
4. Class C IP address

Hostname is the name of the host/computer. In practice, a hostname is everything between the name of the protocol (http:// or https://) and the first slash ("/") character after the end of the domain name. It includes websites starting with the "www" prefix or any other prefix [7,9]. Domain names are composed of the name and a top-level domain [5]. Every domain name is assigned one IP address using a DNS [10]. DNS allows using easy to remember proper names instead of IP addresses while browsing the Internet. A class C IP address means considering only the first three numbers of the address.

Hyperlinks are hypertext links which join one website with another. If one counts hyperlinks leading to a single website, the values of subsequent categories (hostname, domain name, IP, class C) will always be decreasing in that order. The number of domain names will never be lower than the number of hosts. There will never be more IP addresses than domain names. There will never be more class C spaces than there are IP addresses.

According to that logic, the number of hyperlinks leading to a domain name is brought down to the smallest common denominator, i.e. various class C values, from which hyperlinks to the website are derived. Individual IP addresses may be attributed to particular hosting providers using publicly available data. This allows checking whether class C details are owned by a single web hosting company.

The aforementioned arguments indicate that it is much better to build a hyperlink network based on the attribution of an IP address to a particular hosting provider than based on class C. Numerous various class C spaces can actually belong to a single web hosting provider. It is not known how advanced Google is in terms of hyperlink devaluation. Nevertheless, one must remember that when planning to build a hyperlink network based on one's own domain names.

So far, the basics of IP addresses and the explanation of how to construe the number of IP addresses and class C spaces in the context of hyperlinks have been presented. Another element

which has a significant impact on the perception of IP addresses by search engines are autonomous systems. The assumed hypothesis concerning the influence of autonomous systems is supported by an analysis illustrating the influence of autonomous systems on building a ranking in Google search engine.

3   Hosting

As it was already presented above, hosting providers who host websites and domain names offer various ranges of IP addresses. The simplest option for building a hyperlink network using own domain names consists in using a shared server. Shared server accounts have a limited number of domain names which may be assigned to them. Each of the domains on the shared server account will have the same IP address assigned to it. It also applies to all other accounts created on that server and sold to other people hosting their websites using the same hosting provider. In practice, servers handling shared accounts have hundreds of domains assigned to one IP address. In some rare cases, there can be several gTLDs or ccTLDs (generic or country code top-level domains) hosted on a shared server, and all of them may be assigned to the same IP address. Hypothetically, if hyperlinks from all domains assigned to the shared server are obtained, there can be numerous connections generated, however all of them would come from the same IP address.

Another situation, which is more common among companies providing SEO services, is having a VPS (Virtual Private Server) or a dedicated server [3]. A VPS is a virtual fragment of a larger server, on which the user can use more resources. Among the numerous offers for using a VPS, there are some which offer more than one IP address at the client's disposal, however they most frequently belong to the same class C. The standard option is a single VPS being assigned a single IP address. While publishing domain names on a VPS there are 1 to 4 IP addresses available in average, but it is possible that all of them will be in the same class C. The number of domains derived from those addresses depends solely on the server owner. Hypothetically, if one were to place hyperlinks on all domains installed on a VPS, one would get 1 to 4 IP addresses.

However, it is noticeable that the origin of those domains will be less diversified. It is likely that all domain names will belong to one owner, and the possibility to define DNSs heavily depends

on where the domain name is registered. Therefore, such hyperlink network, despite being located at several IP addresses, will look less natural.

The same situation occurs when a dedicated server is used. In such case, there is only additional certainty that the utilised IP addresses are not used by third parties as well, i.e. that they are used solely by one owner. Against all appearances, such a solution causes the created hyperlink network to look even more unnatural. This is simply because there are no third-party owned websites around the website.

The third solution which is to serve as a method for circumventing aforementioned limitations are hosting providers offering numerous IP addresses attributed to a single account. Such addresses are usually offered as packages and they belong to various class C spaces. However, the efficiency of such accounts is very limited, and the hosting service owner usually makes 0.5 to 2% of server's CPU power and memory available. The possibility to install websites mainly involves simple HTML-based websites or well-optimised PHP scripts. Assuming that class C is still important, such an account gives access to several or even over a dozen class C spaces offered by hosting providers.

In general, individual class C spaces used by each hosting provider tend to be unique. If one were to buy several accounts, each from a different hosting provider, the address space would be cumulated. Such a solution seems to be a proper one and it defines the maximum possible class C address diversification. However, it is subject to significant limitations, the primary one being hardware performance Administrators of such accounts are very sensitive to server load, and an attempt to set up efficient websites which would use up more resources than it is allowed, most frequently meets with limitation mechanisms being launched. Another frequent outcome is a server response code informing the user that the server cannot handle the client's request at that moment due to being overloaded.

Having presented certain possibilities for obtaining IP addresses, let us proceed to autonomous systems. As it was indicated above, certain address spaces belong to small or large web hosting providers. The list available at ftp://ftp.ripe.net/ripe/stats/membership/alloclist.txt shows that in Poland there is a possibility to use the services of 469 Internet providers. Not all of them continue to be active, which is indicated by the "closed" status next to numerous company

names. Also, not all of them provide web hosting services allowing users to publish websites. The list includes large corporations which use the assigned pool of IP addresses for their own needs.

To put it simply, it can be assumed that the smallest possible IP network is one class C. Centralised management of a such a global computer network as the Internet is technically impossible, and, most importantly, it would be inadvisable. Thus, the Internet is managed independently at the highest level. There are numbered areas, each of which comprises numerous IP networks. Such areas are referred to as autonomous systems.

4   Autonomous systems

According to the technical specification which is defined by RIPE in Europe, RIPE is the one managing the assignment of subsequent autonomous systems to independent hosting providers. An autonomous system on the Internet is a network or a group of networks subject to joint administrative control. Clearly, it has more tasks and functions, but they are not necessarily important from the point of view of building a ranking in Google search engine.
The statistics for all autonomous systems are available at http://www.cidr-report.org/as2.0/. They indicate that as of 5 December 2016, there are 55,910 autonomous systems in operation worldwide. There are slightly more than a thousand such systems currently operating in Poland.

Each autonomous system must have its own number assigned by RIPE. It is a 32-digit number—the Autonomous System Number (ASN)—which is later used during the exchange of information about Internet data package routes. The numbers are unique to prevent looped routing.

A single Internet provider, hosting provider or a corporation may have more than one autonomous system. The exactness of addresses assigned at the level of autonomous systems is a good compromise between being realistic and the degree of complication of acquisition of data for analysis. Each autonomous system has at least one broadcast prefix. The broadcast prefix is a sequence of IP addresses assigned to the individual autonomous system. An example of a prefix broadcast by SILWEB-AS-EDU, number AS8508, by an Internet provider for the University of Economics in Katowice is 213.227.80.0/20. Information on prefixes broadcast by autonomous systems is public. Therefore, the remaining data contains the following information:

1. Number: AS 8508

2. The number of prefixes: 8

3. The number of IP addresses: 160,000

4. IP/prefix: 20,000

5. AS Name: SILNET-AS-EDU

6. AS description: Silesian University of Technology, Computer Centre

7. Country: DE

8. Assigned: 19971013

9. Prefix: 213.227.80.0/20

Information presented above is easy to read. The first one is the date of assigning of IP addresses to the provider, and the second one is the broadcast prefix. Each autonomous system may broadcast a certain number of IP addresses grouped into prefixes. The exact significance of that number is as follows: from the point of view of building a ranking in Google search engine, one can obtain 2,228 various class C addresses from a German company—Hetzner. Hetzner broadcasts 16 prefixes, which give a total number of 570,368 IP addresses to be used.

The autonomous system number is also very significant. In the beginning, it was written using 2-byte numbers, which allowed launching up to 65,536 autonomous systems. However, in 2008 the numbers were changed to 4-byte numbers and from then on there could be many more autonomous systems created, yet the assigned numbers are still lower than 65,536. The ASN also allows a determination of whether the prefixes being broadcast, despite the significant differences in the first two octets in the address, belong to the same autonomous system.

The largest autonomous system in Poland, TPNET, was assigned the number AS5617. According to publicly available data, TPNET, i.e. Orange Polska S.A., broadcasts only 38 prefixes, even though it has twice as much of them registered for broadcasting. Practice shows that the list of records in RIPE's database is an administrative item, therefore it must always be compared to the actual state. The size of an autonomous system is not defined by the number of broadcast prefixes, but rather by the range of assigned IP addresses. Obviously, there are providers who broadcast more prefixes, however the aggregate of IPs at their disposal is much smaller than in case of TPNET. TPNET has close to 5.5 million IP addresses at its disposal.

That information can be correlated with the findings concerning web hosting providers offering numerous IP addresses, i.e. the third option for acquiring addresses, by pointing out that each of the latter make several or even dozens of class C IP addresses available. The companies' offers clearly indicate which class C spaces are concerned. Users who use their services may check which IP addresses are available through their administration panel. After looking at several of them, one may conclude that none of the hosting providers offering several IP addresses is an autonomous system.

Their IP address pools are leased from other autonomous systems. Having checked lists containing addresses offered by hosting companies one may find that each of the hosting providers belongs to an autonomous system. The most common autonomous system is AS16276, i.e. OVH from France. The diversification of IP addresses and class C address space may prove inefficient, if publicly available data shows that all of them are within the same autonomous system.

Autonomous system diversification affects Google search engine's website ranking. The data is publicly available and it leads to aforementioned conclusions, if one were to compile a fragment of Internet data and use it to show who owns certain ranges of various address classes. It is possible that Google search engine has known about it for a long time and considers autonomous systems while evaluating the structure of a hyperlink network. They are no longer mere hyperlinks from various domain names based on various IP addresses—they are also hyperlinks from IP addresses belonging to various autonomous systems. While building a hyperlink network one should select such hosting providers, which provide a maximal autonomous system diversification.

Therefore, before purchasing a hosting service to build a hyperlink network one should verify to which autonomous system the hosting provider's domain name belongs. The most frequent case is that the IP address of a server purchased at a hosting company will belong to the same autonomous system as the company's website.

Having analysed the list of 1,000 Polish hosting providers and those operating on the Polish market it was found that the most popular autonomous system is the one belonging to a French hosting provider—OVH Systems—which is used by as many as 183 hosting companies, the

second place is taken by German HETZNERAS Hetzner Online AG RZ with 106 hosting companies, and the third place is taken by Polish TPNET Orange Polska S.A. with 33 companies. During the analysis, autonomous systems of 939 companies were recognised, while the attempts to determine those used by the remaining 61 companies have failed (an error message was displayed).

5   Evaluation

The aforementioned hypotheses have been verified using the following analysis. Using the resources available in Poland on the Internet, the most one can do is to create a network of hyperlinks leading to websites within address ranges assigned to several autonomous systems. Such a situation is faced by anyone wanting to establish such a network. The hypothesis was verified using a domain name which did not have any hyperlinks leading to it yet. It is a commercial website. The website was not modified in any way prior to the analysis. The key word which was used did not result in the website being displayed anywhere among the first 1,000 Google search results.

A sitewide hyperlink to the website was prepared with the key word, which was then inserted on websites built using two various scripts. A sitewide hyperlink is a fixed hyperlink displayed on all subpages under a particular domain name. The first script is Webmini catalogue which is used to generate content which is rapidly indexed by Google, the other script is WordPress. In case of the first script, the hyperlink was placed in the side panel, while in the second case it was either located in the side panel or in the website footer. The hyperlinks' appearance and their placement on the website was not favourable in any way [3]. They were not surrounded with content and they were located in sections which tend to be depreciated by Google. What was important, was that the domain names with the scripts were located in more than 20 various autonomous systems.

The hypothesis was verified using 38 domain names at 28 various IP addresses, which were located within 28 various class C spaces, which belong to 20 various autonomous systems. There are no ready-made, dedicated tools yet, which would be able to provide such pooled data. There are only tools which provide such data for individual source domains, such as a plug-in for Firefox browser called ASNumber or scripts written using Linux OS. The list of autonomous

systems used to verify the hypothesis does not include autonomous systems which are common in Poland, which has possibly given the hyperlinks an additional attribute of uniqueness.

After 14 days, the website appeared among the first ten Google search results for the analysed key word. The conducted analysis shows that the Google ranking algorithm considers hyperlinks from various autonomous systems as an important factor. The research hypothesis was positively verified.

The article involved conducting an analysis of influence of IP addresses and autonomous systems on building Google ranking. The results of the analysis confirm the hypotheses. One should diversify IP addresses used while building one's own network of hyperlinks which are to be used to build one's Google ranking based on servers or leased accounts. One should remember to maintain the diversification at all levels and, at the same time, diversify class C spaces and autonomous systems providing the addresses.

6   Conclusion

The importance of the analyses from the point of view of economics or management is that websites, which have links from different autonomous systems are more visible in search engines. This way websites have more traffic from users using search engines and can realise more their own goals like having contact enquiries or purchases.

References

[1] M.P. Evans, Analysing Google rankings through search engine optimization data. Internet research, 17(1), pp. 21-37 2007.
[2] L. Moreno and P. Martinez, Overlapping factors in search engine optimization and web accessibility. Online Information Review, 37(4), pp. 564-580, 2013.
[3] S. M. Beloussov, S. S. Protassov and A. G. Tormasov, Virtual private server with enhanced security. U.S. Patent No. 7,461,144. 2 December 2008.
[4] K. Egevang, and P. Francis, The IP network address translator (NAT). No. RFC 1631, 1994.
[5] M. A. Islam and M. S. Alam, Webometric study of private universities in Bangladesh. Malaysian Journal of Library & Information Science 16.2, pp 115-126, 2011.
[6] K. Nichols, S.Blake, F. Baker, and D. Black, Definition of the differentiated services field (DS field) in the IPv4 and IPv6 headers. No. RFC 2474, 1998.
[7] B. Shuster, J. Leong, M. Price, B. Lam, and D. F. Johnson, System apparatus and method for hosting and assigning domain names on a wide area network. U.S. Patent No. 6,687,746. 3 February 2004.
[8] P. Srisuresh, and K. Egevang, Traditional IP network address translator (Traditional NAT). No. RFC 3022, 2000.

[9] F. J. Stanbach Jr, D. G. Hoffman and B. R. Keiser. Internet hosting system. U.S. Patent No. 6,449,657. 10 September 2002.

[10] Y.F. Tan, M. Y. Kan and D. Lee, Search engine driven author disambiguation, Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, ACM, 2006.