Building a Generalized DNS Resilience Tool using the Internet Yellow Pages

BARRY TER HEEGDE, University of Twente, The Netherlands

The Internet Yellow Pages (IYP) is a recently released tool on which data on the structure of the Internet can be queried. The structure of the Domain Name System (DNS) can be analyzed for resilience on multiple metrics, which depict single points of failure and Anycast usage, and can give an indication on resilience of DNS of a part of the Internet.
This research provides a way to use IYP to identify these metrics on the scale of a country, by looking at the structure of DNS of the most visited websites by the Internet users of the Netherlands. Additionally, in this work we research these metrics on the most visited websites, and compare the results to these of the DNS structure of the Dutch government.

Additional Key Words and Phrases: Domain Name System, Nameserver, Resilience, Internet Yellow Pages, DDoS

ACM Reference Format:
Barry ter Heegde. 2024. Building a Generalized DNS Resilience Tool using the Internet Yellow Pages. In . ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In the landscape of website data on the structure of the internet, there are numerous organizations collecting data. Parts of this collected data are overlapping, but often, data on what part of the internet is collected differs just a slight bit per organisation.

Projects like OpenINTEL use these openly available datasets to expand this data by actively collecting Domain Name System (DNS) data, and makes this data available for academic researchers [8].

These projects are all mapping a part of the Internet, but never give a full picture, though are they all part of the same system: the Internet.

On the 16th of January, 2023 a database tool was released by the Internet Health Report project, called the "Internet Yellow Pages"[11]. This tool, which will be referred to as IYP, is a knowledge database that gathers information about Internet resources, combining a number of openly available datasets in an easily queryable way. It does this by combining data in Neo4j [7], a graph database management system [10].

Recently, an article has been written, where this tool has mapped the Japanese Internet in different ways, with a few of them being on the structure of DNS around a website, specifically mapping the nameservers of websites [4].

This article demonstrated the possibilities of this new tool, and that it can be used to map specific parts of the internet, and with that gain insight on multiple parts of the structure of this subgraph, as the IYP combines all this data in one graph.

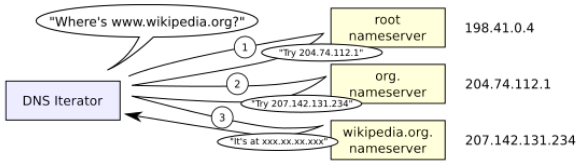

Fig. 1. Traffic flow at a DNS resolver

DNS, short for the Domain Name System, is a system introduced in 1985, to act as a phonebook for the Internet. All websites that are visible online are located on a server somewhere in the world. These servers are located at a certain IP Address, which acts similar to a house address in the real world. IP Addresses are numerically defined and can be quite hard to remember. Examples of IP Addresses can be "130.89.3.249", or "2001:67c:2564:a102::1:1", which are IPv4 and IPv6 addresses respectively. DNS gives puts a label on these addresses, which transforms the two just given IP Addresses to the easy to remember name "utwente.nl".

The system works like this:

(1) An end-user types in a domain name, such as "www.wikipedia.org".
(2) The recursive resolver, or DNS Iterator, is consulted to find the IP Address of this domain name.
(3) The DNS Iterator asks the root server where ".org" TLD server is located
(4) After receiving a response, it asks this TLD nameserver where "wikipedia.org" is located.
(5) The TLD nameserver responds with the nameservers that know where "wikipedia.org" is located, and a nameserver is consulted.
(6) This nameserver responds with the IP Address where "wikipedia.org" is located
(7) The DNS Iterator returns the IP Address of the queried domain name to the user, which then can be queried to find a web page.

If the Domain Name System fails, websites become unavailable through their domain names, which would mean users would not be able to use the websites anymore. This is why it is important that DNS is sufficiently resilient against system faults, but also cyberattacks.

In a previously done research on the resilience of the Domain Name System (DNS) of the government of the Netherlands [6], it was shown that for a part of all domains relating to the Dutch government, redundancy to reduce critical single point of failure was subpar.

TScIT 40, February 2, 2024, Enschede, The Netherlands

For the government, it is useful to know where your critical points of failure are, and what you can do to improve this, as millions of people are dependent on the services of the government being available. But, this does not only extend to government, but also to all sorts of businesses. If a webshop goes offline, it misses out on sales, if a search engine goes down, it misses out on ad revenue, and if an news website goes offline, it misses out on readers. And one thing which is lost, that are all in common, is the sense of reliability to it's users.

1.1 Problem Statement

In this research, we will investigate whether it is possible to identify resilience in the Domain Name System of specific subparts of the Internet through the Internet Yellow Pages. Specifically, this prompts the main research question:

What insights can we gain on the resilience factors of the Domain Name System of the Netherlands through the use of the Internet Yellow Pages?

This research question can be answered with the following sub-questions:

(1) What are the factors that make DNS resilient?

(2) To what extent can we model significant parts of the Dutch Internet using the Internet Yellow Pages?

(3) What significant differences in resilience of the Domain Name System can we identify between the Dutch government websites and the most visited websites in the Netherlands?

The purpose of this paper is to investigate to what extent we can use the Internet Yellow Page to map the "Dutch Internet", and what metrics of resilience we are able to get out of this data.

In section 2, we will consider works related to the resilience of DNS, and attempt to answer RQ1. After this, we will discuss the methodology used in the research to obtain the results. Afterwards, the results are stated. These results are discussed in section 6, after which a conclusion summing the main points of this research is given.

2 RELATED WORK

Research on what makes the Domain Name System resilient has been done previously, pointing to numerous factors, varying greatly in criticality.

One of the previously mentioned researches [6], on the resilience of the Dutch government's DNS, uses a few different metrics. These metrics mainly focus on the single points of failure, that is, if one server or program fails accidentally or by malicious intent, a domain will become unreachable.
These single point of failures are identified by looking at the number of available providers a domain is dependent on, in all levels of the DNS-hierarchy, it being nameservers, autonomous systems, or even top-level domains. If only one is provided, and it malfunctions, a domain is unreachable. This becomes less likely, the higher in the DNS-hierarchy you go, as systems higher up in the hierarchy tend to be more robust [1].

Allman additionally notes that the sharing of DNS resources by parts of the Internet can make a system more vulnerable on a larger scale, and it is not uncommon for companies to outsource their DNS service to a third-party. While DNS can easily be misconfigured when attempted by yourself, and outsourcing this to a trusted company is a good way to make sure your DNS service is correctly working, it does allow for a single point of failure for not only a single domain, but for a chunk of the internet.

An example of such a failure is the 2016 attack on DNS provider Dyn, which provided a large part of the DNS services for the East Coast of the US, and with this, a chunk of the internet was offline by only attacking one provider.[9]

Another research identified the use of Anycast as an effective mechanism to enhance resilience of DNS. Along providing better latency and other benefits, Anycast enhances resilience by distributing traffic to multiple Anycast sites, being able to mitigate the impact of Distributed Denial of Service (DDoS) attacks [5].

3 METHODOLOGY

3.1 Data

The Internet Yellow Pages are not a data-measuring source itself, it is merely a tool combines a number of databases into a single queryable point. The data that it contains is currently taken from 18 data sources [11]. This data varies from research-collected data, such as projects like OpenINTEL by the University of Twente, and ASdb by Stanford University, to commercially collected data, such as the Cloudflare Radar data. IYP combines these datasets, by combining the overlapping parts of the datasets, such as domain names, AS'es and prefixes, to a single object, creating a giant graph, which effectively maps the top 1 million domains, including the infrastructure that connects these domains. This means that sites that see a decent number of visitors every day are included, but IYP does not include a small personal website with 10 visitors per day.

The infrastructure that IYP contains also the connections between AS'es and IXP's, and the infrastructure that provides the Domain Name System, such as nameservers.

Furthermore, in this research, two external data sources are used. These are commercially collected lists, one provided by Semrush, and one by Cloudflare. Semrush publicly publishes the Open .Trends Top 100 websites [13], which estimates the number of visitors of a website through Clickstream data [12]. The December 2023 ranking was used. Cloudflare publishes the Cloudflare Radar Top 100 domains, publicly available online on the Cloudflare website [cloudflareref]. These two lists differ in domains they contain. The list published by Semrush includes the top 100 domains by active search, so queries made by users themselves, due to their data being Clickstream data. The Cloudflare Radar top 100 is largely made of domains which are queried by machines, such as API's and Social Networks, due to Cloudflare using their public DNS-resolver as a collection point [3].

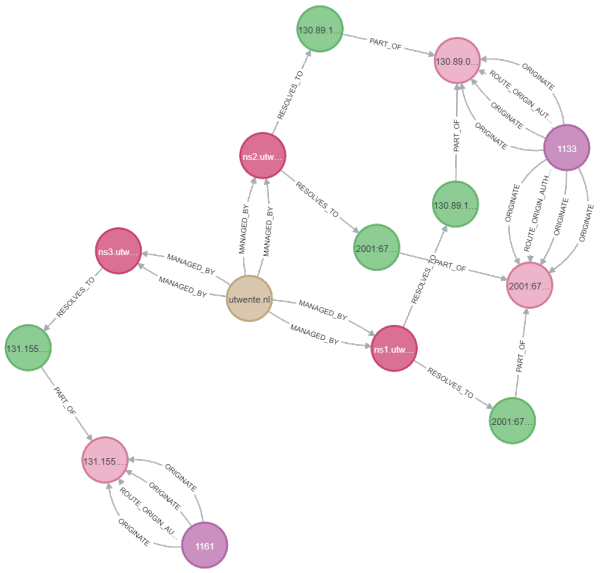

Fig. 2. Internet Yellow Pages data depicting the structure of DNS of 'utwente.nl'

3.2 Mapping the Dutch Internet

In order to answer RQ2, we first have to define what the "Dutch Internet" is. In this paper, we define it as the part of the Internet, used by people living in the Netherlands. The reasoning behind this, is that part of this paper is looking at points of failure, ways that the domains can become unreachable. So when defining what the Dutch Internet is, we decided it would fit best that it would be the part that, when unreachable, would have the largest impact on the people living in the Netherlands.

Websites, such as google.com, youtube.com, and instagram.com are hosted outside of the Netherlands by American companies, but are just as important to be accessible to the people living in the Netherlands, if not more important, than a website like nrc.nl, a typical Dutch newspaper and should be considered when mapping the Dutch Internet.

The Internet Yellow Pages contain just a bit more than 1 million domain names, as the data that it uses, is data of the top 1 million visited domains worldwide, collected by 2 different data sources: The Tranco Top 1M list, and the Cisco Umbrella Top 1M list. These are a lot of domains, and are enough to give a good image of our definition of the Dutch Internet.

To isolate the part of that list that represents the "Dutch Internet", we first considered filtering the domains on all falling under the top-level domain (TLD) ".nl". This is a domain which is intended to be used by entities connected with the Netherlands. However, this misses a significant part of the internet that is used by people in the Netherlands, as it should be including websites as "google.com" too. Furthermore, there are Dutch-only websites too, that do not fall under the .nl TLD, such as "bol.com". This means that filtering on TLD does not cover our scope.

The same problem was found when filtering domains based on the location in which their prefix is hosted. This data, provided by the Internet Health Report, say barely anything about who uses the domains. The Netherlands has a relatively high number of datacenters [14], and a high number of internationally used domains with at least one server in the Netherlands, but are barely used by Internet users in the Netherlands. Next to this, some popular Dutch domains have their hosting and DNS set up by companies such as Amazon & Google (such as "bol.com").

The Tranco Top 1M ranking is based on visiting numbers, but the visiting numbers themselves are not included in the IYP. This creates a problem for us, as the Dutch Internet as we define it, is based on visiting numbers.

The IYP does contain a "QUERIED_FROM" relation between a domain name and a country, which references from the Cloudflare database's "DNS Top Locations" data. This data contains for every domain name, per country the percentage of traffic that originated from that country proportional to the the total traffic received by the domain.

Through this, we were able to list the domains that received for the biggest part traffic from Internet users in the Netherlands. However, the database dump that we use only has the top 10000 domains according to the Top 1M lists, and resulted in a list of only 74 domains. However, the threshold of what amount of domains this data can be crawled for is customizable, and more domains could be pulled.

The resulting domain name list did have a problem though, as it does not include domains that are widely used by users in the Netherlands, but also have a significant user base outside of the Netherlands. The Netherlands has too little Internet users. The most used domain in the Netherlands, google.com [13], is not included in that dataset, as more than 36% of it's traffic originates from the United States, and only of the traffic 1.5% originates from the Netherlands. As this data excludes a significant part of the Internet that users in the Netherlands, we decided it did not suffice as mapping the "Dutch Internet" by our definitions.

To get a more representative image of the Dutch Internet, we used an external source for our domain list. This source provides a list of the most visited domains in the Netherlands. The two aforementioned lists of both 100 domains, combined, with the duplicates removed, result in a list of 183 domain names.

With these considerations the Internet Yellow Pages will be consulted, and with all available data on this tool, we will try to see whether it is possible to write a query in such a way that it contains all the data that is needed to get a good view of the resilience metrics defined in RQ1.

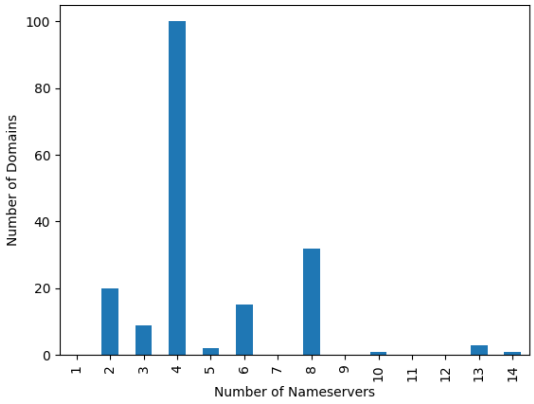

Fig. 3. Number of domains (Y) using what number of nameservers (X)

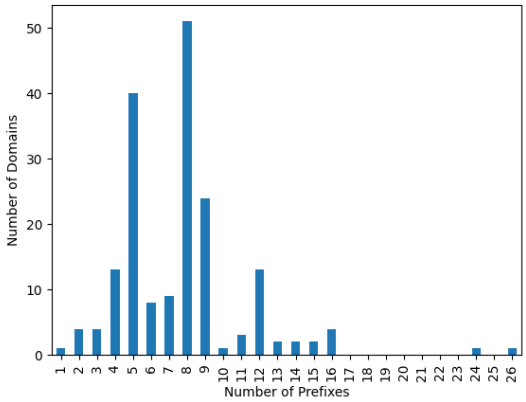

Fig. 4. Number of domains (Y) using what number prefixes for their nameservers (X)

3.3 Comparing resilience data

The third research question will be answered by combining the answers of the first two research questions, to get data on these resilience factors of the Domain Name System of the identified websites, and then comparing it with the DNS of the Dutch government.

To compare the found data with that of the Dutch government, we can compare our results with those given in the DINO Project Advisory Report. We will be comparing the number of DNS providers per domain, and the usage of anycast, by comparing the number of domains that have one or more anycasting DNS servers. We will not be comparing TLDs used by the nameservers of domains, as this is not identified as increasing resilience.

4 RESULTS

Resilience factors of the Domain Name System can be identified in previously written research. To sum up what these factors are, and what we will be working with, the identified points of resilience are, in the scope of DNS, per domain:

- The number of Authoritative Name Servers
- The number of Prefixes that the Authoritative Name Servers fall under
- The number of Autonomous Systems
- Use of Anycast

Another identified point that can improve the resilience of DNS, are the time-to-live (TTL) values of DNS records in nameservers [6], which are how long records are cached, but as the Internet Yellow Pages does not provide any data on TTL values, this cannot be looked at.

Using this list of domain names as a list of domains to query in the IYP, we were able to model a significant part of the Dutch Internet.

Processing the output of the IYP, we were able to model the metrics identified in RQ1.

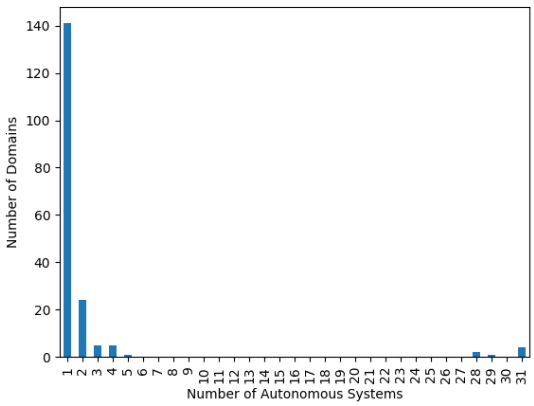

Fig. 5. Number of domains (Y) using what number Autonomous Systems for their nameservers (X)

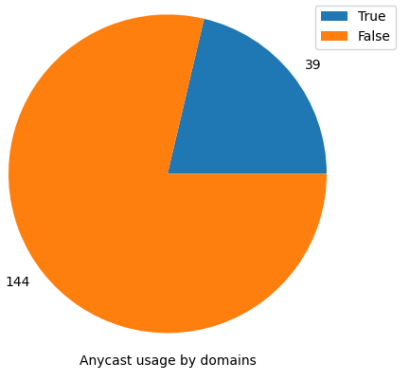

Fig. 6. Number of domains having one or more anycast DNS servers

Figure 3 shows the usage of the amount of nameservers. We see that no domain is using only one nameserver, mitigating a single point of failure. Actually, the most common option in the list of domains is to use 4 different nameservers. However, if all of a domain's nameservers are hosted on the same prefix, this still causes a single point of failure. That's why, we show of prefixes used to host nameservers per domain in Figure 4.

Only very few domains put all of their eggs in one basket, hosting on 1 prefix. For most of the domains in the Dutch Internet, the number of prefixes used seems to be equal to, or more than the number of nameservers of a domain. This means that most nameservers are hosted on one or more prefixes, which indicates good resilience.

Figure 5 shows the number of AS's identified that are hosting the nameservers per domain. An interesting observation is that the majority of domains are hosted by only a single AS, with 141 out of 183 domains (77.0%).

Another observation that can be made, is that generally the maximum Autonomous Systems a website's nameservers are hosted by is 4 to 5. But, out of the 183 domains, 7 domains have more, and use 28, 29 or 31 AS'es. These domains all use the same company for hosting their nameservers, which is NeuStar Security Services. The prefixes hosted by this company connect to the high number of AS'es, which are all AS'es with ASN's in close proximity.

Anycast usage, seen in Figure 5 does not seem to be highly adopted yet in the Dutch Internet, at least based on the available data on the Internet Yellow Pages, which is supplied by BGPtools.

We can compare these results with the results found in the DINO Project Advisory Report [6], where around 50% of all publicly available government domains were announced by a single AS. For the Dutch Internet, this was found to be 77%, and thus having a larger part with single AS usage as a critical point of failure.

When comparing Anycast data, around 85% of the Dutch government's domains had no anycast DNS servers. This is higher a higher number than the Dutch Internet, as 21.3% of the domains in the dataset has one or more anycast DNS servers, meaning 78.7% has no anycast DNS servers.

5 DISCUSSION

In the process of generating the results, limitations of the Internet Yellow Pages can be noted.

The first is that IYP lacks data on visiting per country. The modeling of the Dutch Internet, as to our definition, is not possible. Visiting data per domain per country is not included in any of the datasets the IYP uses, so an external source has to be used for mapping parts of the internet by visiting numbers. The reason that this is not included could be because such data would be very computationally expensive to collect, and while it can be done, it is not done by research institutes, but rather by commercial parties that would like to sell this data.

The found data on anycast usage might be a wrong depiction and limitation of the IYP. This is because the data of BGPtools is used. BGPtools admits on their website, that the way they collect Anycast data causes for quite some false negatives [2]. A prefix is only identified to be anycasting, if it is anycasting in Western Europe, West Coast US, and East Coast US. For a large part of the Dutch Internet, domains may be only anycasting in Western Europe, as the largest number of users are only from the Netherlands anyways [2].

While the two figures on the number of nameservers used by domains, and the number of prefixes used by these nameservers, show that generally more prefixes are used than nameservers, hinting at redundancy in the number of prefixes used by nameservers, it may still have a significant number of domains with critical points of failure, or not be as resilient as it seems. There is still a possibility that even when a domain has multiple nameservers, all but one nameserver are linked to the same prefix, and that the last nameserver is linked to numerous prefixes. Though it is unlikely and would be illogical to set a system up in this way, and it still having some form of resiliency, it should be noted that it would not appear any different in our charts than highly resilient systems, having multiple nameservers and multiple prefixes per nameserver.

The number of domains used to generate the results is low. This was due to being forced to use an external source for the list of domains, and a larger list with the visiting data per country per domain is only commercially available, with the current list being a free sample.

The results also show a few domains using between the 28 and 31 AS's for hosting their nameservers. All the nameservers of these domains are advertised through the AS'es of 'Neustar'. These AS'es are independently registered, but the infrastructure between them is unknown, and it is unknown why Neustar does this. Because of this, we cannot conclude whether this achieves actual resilience: the different AS'es may be all using the same infrastructure, which results in a single point of failure, and lacking any resilience.

5.1 Reproducibility

To facilitate reproducibility and to build further on used queries, or usage for a different part of the Internet, we published the codebase used in this research. This can be found at https://github.com/barryth/iyp-metrics-tool.

5.2 Future Work

While this research lays down an example of what the Internet Yellow Pages can be used for, and in what way considering the limitations, further research could be done on DNS, but also on the resilience of the Internet in other aspects.

Mainly, further research can be done on the Dutch Internet, if more visiting data per country was available, as it would possibly give a better insight on the resilience of a greater part of the Dutch Internet.

Furthermore, with IYP it is possible to query the number of prefixes per nameserver, and give insight on how this is generally structured.

Furthermore, there is more to the Domain Name System and the Internet in general than prefixes and Autonomous Systems. AS's and are connected to other AS's through a web of perring connections and connections to IXP's, who on itself are interconnected with each other. This data is available on IYP, and further resilience by redundancy and single point of failures can be investigated using IYP.

And lastly, if in the future data such as visiting data per domain per country, accurate anycasting data, and TTL data is added to IYP, it is worth repeating this study to include those factors.

6 CONCLUSION

The Internet Yellow Pages is a tool that groups datasets by different organizations in a useful and easy-to-access way. It can be used to test certain resilience metrics on a part of the Internet, all through a single portal, though it does have certain limitations. Specific subparts of the Internet, based on visiting data, need external sources to define a subgroup, but other parts may be possible by solely using IYP. Next to this, Anycast data is available through IYP, but appears incomplete, and this metric currently needs a better, external data source, to be able to be correctly shown.

REFERENCES

[1] Mark Allman. 2018. Comments on DNS Robustness. https://www.icir.org/mallman/pubs/All18a/All18a.pdf.
[2] bgp.tools. 2024. How bgp.tools detects anycast addresses. bgp.tools/kb/anycatch.
[3] Cloudflare. 2024. About Cloudflare Radar. https://radar.cloudflare.com/about.
[4] Romain Fontugne. 2023. Understanding the Japanese Internet with the Internet Yellow Pages. https://blog.apnic.net/2023/09/06/understanding-the-japanese-internet-with-the-internet-yellow-pages/.
[5] Remi Hendriks. 2023. Improving anycast census at scale. https://essay.utwente.nl/95878/1/Hendriks_MA_EEMCS.pdf.
[6] Sommese R. Jonker M. Moura, G. 2022. How resilient is the Netherlands government DNS. https://www.ncsc.nl/binaries/ncsc/documenten/rapporten/2023/maart/14/best-practices-for-resilience-of-authoritative-dns-servers/dino-mngt-report-v2-EN-annotated.pdf.
[7] Inc. Neo4j. 2024. Neo4j Graph Database Analytics. https://neo4j.com/.
[8] OpenINTEL. 2023. OpenINTEL - Active DNS Measurement Project - Data Access. https://openintel.nl/data-access/.
[9] Nicole Perlroth. 2016. Hackers Used New Weapons to Disrupt Major Websites Across U.S. https://www.nytimes.com/2016/10/22/business/internet-problems-attack.html.
[10] Internet Health Report. 2024. Github - Internet Yellow Pages Console. https://github.com/InternetHealthReport/internet-yellow-pages.
[11] Internet Health Report. 2024. Internet Yellow Pages Console. https://iyp.iijlab.net/.
[12] Semrush. 2021. What Is Clickstream Data and How Do We Use It at Semrush .Trends? https://www.semrush.com/blog/what-is-clickstream-data/.
[13] Semrush. 2024. Top websites in the Netherlands (All industries). https://www.semrush.com/trending-websites/nl/all.
[14] Statista. 2024. Number of data centers worldwide in 2023, by country. https://www.statista.com/statistics/1228433/data-centers-worldwide-by-country/.

Appendix A QUERIES

Query getting the top 500 domains by visiting data, which are hosted in the Netherlands:

```
MATCH (c:Country {country_code: 'NL'})-[cx:COUNTRY]-(px:Prefix)-
[p:PART_OF]-(ip:IP)-[r:RESOLVES_TO]-(dn:DomainName)-[ra:RANK]-
(tranco:Ranking{name: 'Tranco top 1M'})
WHERE ra.rank < 20000
MATCH (dn)-[m:MANAGED_BY]-(ans:AuthoritativeNameServer)-
[rt:RESOLVES_TO]-(ip2:IP)-[p:PART_OF]-(px:Prefix)-
[o:ORIGINATE | ROUTE_ORIGIN_AUTHORIZATION]-(a:AS)
RETURN DISTINCT(dn), ra.rank as rank, COUNT(DISTINCT(ans)),
COUNT(DISTINCT(ip2)), COUNT(DISTINCT(a))
ORDER BY rank ASC
LIMIT 500
```

Query used to generate resultdata:

```
MATCH (dn:DomainName)
WHERE dn.name in """ + str(query_domain_list) + """
MATCH (dn)-[m:MANAGED_BY]-(ans:AuthoritativeNameServer)-
[rt:RESOLVES_TO]-(ip:IP)-[p:PART_OF]-
(px:Prefix)-[o:ORIGINATE | ROUTE_ORIGIN_AUTHORIZATION]-
(a:AS)
RETURN dn.name, COUNT(DISTINCT(ans)), COUNT(DISTINCT(px)),
COUNT(DISTINCT(a)),
EXISTS((:Tag{label: 'Anycast'})-[:CATEGORIZED]-(:Prefix)-
[:PART_OF]-(:IP)-[:RESOLVES_TO]-(dn)) AS Anycast
```