Research Highlights

- We assess the potential global impact of network attacks on websites of public interest.
- We compare four different countries: Italy, Germany, UK, US.
- We consider dependencies from zones, nameservers, networks, autonomous systems.
- We consider also shared groups of IP addresses, networks, autonomous systems
- We assess the usage of defensive mechanisms at different abstraction levels

Robustness Analysis of DNS Paths and Web Access Paths in Public Administration Websites

Alberto Bartoli[a]

[a] *Dip. Ingegneria e Architettura, University of Trieste, Italy*

Abstract

Attacks at the naming or the routing infrastructure of the Internet have long become a reality and one single such attack has the potential of affecting access to Internet-facing services in many organizations. An important question to address is assessing the potential impact of attacks of this sort on the web infrastructure of an *entire nation*. In this work we examine the dependence of a large set of public administration websites on DNS entities and autonomous systems of four different countries: Italy, Germany, UK and US. We collected the dependencies of those websites from DNS zones, nameservers, networks, autonomous systems, and assessed the potential *global* impact of *localized* attacks on those entities. We also analyzed the prevalence of such defensive technologies as BGP Route Origin Authorization, DNSSEC and HTTPS Strict Transport Security. Our analysis highlights the structural interdependencies within the web infrastructures of public interest and illustrates the corresponding open problems, issues whose relevance can only grow.

1. Introduction

Attacks at Internet-facing services of an organization executed by abusing services of *other* organizations have long become a reality, with significant examples that have occurred at different abstraction levels. At the *routing level*, an attacker within an organization that manages an autonomous system may attempt to acquire a man-in-the-middle capability for a targeted organization by propagating malicious routing information to other autonomous systems, e.g. [1, 2, 3]. At the *name resolution level*, an attacker within an organization that manages DNS name resolutions may acquire man-in-the-middle capability with respect to other organizations by manipulating DNS records, e.g. [4, 5, 6, 7, 8]. Similarly, an attacker within an organization that manages an autonomous system may attempt to acquire a man-in-the-middle capability for the name servers themselves by means of malicious routing information, e.g., [9]. *Impersonation* attacks of this kind are attractive to attackers because they are generally difficult to detect and have the potential to hit many organizations at once, thereby

Email address: bartoli.alberto@units.it (Alberto Bartoli)

amplifying the return of attack investment. Indeed, a single attack may have a disruptive effect on a very large set of organizations including those that offer critical services, as the *denial of service* attack to the Internet infrastructure company Dyn has shown, e.g., [10].

An important question to address is assessing the potential impact of attacks of this sort *on an entire nation*. Determining whether a large fraction of, e.g., local governments or hospitals, of a nation, crucially depend on a single autonomous system company or DNS provider, is clearly an issue of crucial importance. Similarly, understanding whether a successful impersonation or denial of service attack to just a few providers could affect a large fraction of sites of public interest in a nation is extremely relevant from a strategic point of view. The fact that Internet service providers and telecommunication companies are clearly not immune from ransomware attacks and could thus be affected in their operational capabilities can only amplify the relevance of these issues [11, 12, 13, 14].

In this paper we examine the dependence of a large set of public administration websites on DNS entities and autonomous systems for four different countries: Italy, Germany, UK and US. We collected the dependencies of those websites from DNS zones, nameservers, networks, autonomous systems, and assessed the potential *global* impact of *localized* attacks on those entities. We considered a threat model in which an attacker may take control of one or more entities and then execute either an *impersonation* or a *denial of service* attack. Specifically, we examined the following research questions:

- How many websites have redundant name resolution paths? What redundancy level is used at the level of zones, nameservers, networks, autonomous systems?
- How many websites could be affected, whether in name resolution or web server access, by an attack on a single zone, nameserver, network or autonomous system?
- How many websites are replicated? How are replicas distributed across networks and autonomous systems?
- How prevalent is the usage of groups of IP addresses, networks, autonomous systems for providing redundancy while controlling the security perimeter to defend? How many websites could be affected by an attack to one of those groups?
- How prevalent is the usage of such defensive mechanisms as BGP Route Attestation Origin, HTTPS, Strict Transport Security?

We are not aware of any systematic assessment of these questions. Our analysis highlights the structural interdependencies within the web infrastructures of public interest and illustrates the corresponding open problems, issues whose relevance can only grow. Beyond the specific results for the four countries that

we have considered, we also believe that our methodology may constitute a practical and sound framework for performing similar analyses on large collections of websites of public interest, e.g., banks, energy providers, health services and alike.

We do not provide any recommendations regarding how a country should design and operate its web infrastructures of public interest, though: this important topic is beyond the scope of this work. The inherent tension between high redundancy in access paths and the need of minimizing third-party dependencies, implies that there is no single correct recipe in this respect and that a wide range of different design choices can be made. Furthermore, and perhaps most importantly, such recommendations should take into account many country-specific factors not considered here, including, e.g., ICT technology available to domestic providers, human resources available to public administrations, quality and usage of public services available on the web, digital skills of the overall population and and so on.

2. Motivations and Related Work

The main motivations for this work are associated with two events related to the COVID-19 pandemic. In April 2020, the Italian Government established the erogation of financial support to several categories of citizens, based on certain requirements and information to be provided by citizens themselves on the web site of INPS, the largest agency that manages the Italian social security system. As soon as the application period started, the INPS website collapsed and became unusable due to a combination of configuration mistakes, inadeaquate capacity planning and a distributed denial of service attack [15, 16]. In other words, breakdowns and attacks on a public administration website have constituted a major obstacle to the provision of important aid to a large part of the Italian population, precisely at a time when that aid was absolutely necessary. This event demonstrated with extreme evidence and crystal clarity that the robustness of public administration websites is no longer a sort of desirable and idealized option: it has already become a feature of fundamental importance and strategic importance.

The second event is the publication, in May 2020, of a short yet thought-provoking opinion piece by Vardi in which, based on an analysis of the huge economic damage inflicted by the COVID-19 pandemic to societies, he suggested that computing professionals should reconsider the relentless pursuit of *efficiency* that has characterized most computing applications of the last decades and emphasize more the value of *resilience* [17]. Understanding what is the robustness of the public web infrastructures that societies have built and deployed is thus a first crucial step in this direction.

In this respect, some recent works have focussed on the potential global risks that could result from localized failures or attacks. An analysis made on 63 countries accounting for more of the 80% of the global Internet population, has shown that more than 60% of all the most popular web resources accessed from a specific country depend on the submarine cable network [18]. A methodology

for assessing the resiliency to seismic forces of the Internet infrastructure in Pacific Northwest, one of the largest hubs for cloud and content provider as well as for submarine networks, is proposed in [19].

Similar concerns are raising at the DNS level, for example, an analysis of DNS traffic at one DNS root server and at two top-level domains showed that as much as 30% of all queries are originated by only 5 providers [20]. Another relevant analysis in this area showed that many websites of the Alexa Top 1 Million list share the same infrastructure for name resolution, with a significant example of 12.000 different websites that actually share all their nameservers provided by third parties [21].

Our methodology has been largely inspired by the work [22] that analyzed the robustness of the DNS infrastructure with respect to *second level domains*. We extended that methodology with a security-oriented framework for web access and focussed our analysis on public administration websites. A sophisticated analysis of *third-party dependencies* in websites is proposed in [23], which analyzed several top-ranked Alexa websites across two snapshots in 2016 and 2020 in terms of dependency on DNS, content-distribution networks, certificate revocation checking, as well as those dependencies in 200 US hospitals and 23 smart home companies. Our work shares the key motivations with the cited work, i.e., shared risks from attacks and cascading failures, but we focus on a more specific and admittedly more limited analysis of dependencies, focussing on the structure of name resolution and web access paths for websites of the public administration.

The potential global risks induced by third party dependencies in DNS have been identified and discussed for a long time [24]. Indeed, our framework for modelling dependencies between the entities of our interest could be seen as an attempt of generalizing the notion of *trusted computing base* proposed by the cited work in terms of DNS zones. This notion constituted the basis for a formal model of dependencies in the DNS and for a metric for quantifying the dependency between any pair of given domains [25]. We analyze redundancies and dependencies in name resolution and web access paths from several, complementary points of view.

We have included networks and autonomous systems in our assessment of third-party dependencies in access paths due to the increasing relevance of BGP-related security issues [26, 27, 28]. Indeed, BGP attacks have occurred even in application scenarios different from the web [29] and others have proven to be feasible [30, 31]. Interestingly, an analysis of BGP routes during the 2014 Maidan Revolution, when Russian forces took control of the Crimean Peninsula, has shown that Russian authorities and separatist forces modified BGP routes in Ukraine and forced a separation between autonomous systems consistent with the military frontline [32].

Several proposals for improving BGP security have been made [33, 34, 35], ranging from detection of BGP hijacking events [36, 37] to prevention of acceptance of fake BGP route messages [38]. We include in our analysis a small and focussed assessment of the actual deployment of *BGP Route Origin Authorizations*, a key technology for the authentication of BGP route messages [39], in a

carefully selected sample of our dataset.

The analysis of interdependencies among services of public interest is an important topic in the much broader context of *critical infrastructure protection* [40]. A methodology for quantitative analysis of those dependencies in complex systems including information and communication components is proposed in [41], along with a case study of the Italian System for Public Connectivity. A comprehensive taxonomy of metrics for quantifying such dependencies is proposed by [42]. In the proposed framework, our analysis could be seen as a collection of *shape* metrics in that we provide some quantitative assessment of dependencies and of their direction. However, we do not attempt to model outages or availability neither at the global nor at the local level. We merely attempt to gain qualitative insights from a carefully selected set of structural properties in our datasets.

Our security-oriented analysis of name resolution paths and web access paths obviously provides only a partial view of the potential security risks of the corresponding websites and organizations. For example, we do not consider software supply chain issues [43], vulnerabilities in devices exposed to the Internet [44, 45, 46], organizational policies for network access [47] and, more broadly, we do not attempt to provide any single metric for distilling the properties of the various datasets that we analyzed [48].

3. Dataset

3.1. Websites

We consider websites in several publicly available lists that we downloaded at the beginning of April 2020. For Italy, the United Kingdom and the United States we used official lists while for Germany we used an unofficial list available on GitHub[1]. The structure of the lists were heterogeneous and, in particular, they used very different categorization criteria for the websites: there were 49, 8, 2 and no categories for Italy, the United States, Germany and the United Kingdom, respectively. We preferred to not enforce a uniform categorization across the four lists due to the intrinsic differences of the corresponding institutions. However, in order to simplify the analysis, we partitioned each list in two parts, one containing websites related to the central government or to services of general interest for the whole country, the other containing the remaining websites. For Italy we executed the partitioning based on the names of the 49 categories and our knowledge of their meaning; for US and DE we considered the presence of the term "Federal" in the website description; for the United Kingdom we considered the presence of the term "Council" in the website description The names that we have chosen for the *central* categories are IT-Central, US-Federal, DE-Federal Agency, UK-Not Council. Although the resulting categorization is

[1] `https://www.indicepa.gov.it/`, `https://www.gov.uk/government/publications/list-of-gov-uk-domain-names`, `https://home.dotgov.gov/data/`, `https://github.com/robbi5/german-gov-domains`.

Table 1: Dataset summary.

Country	Category	Websites	Zones	Nameservers	Networks	ASN
IT	Total	18893	18978	4204	2837	373
DE	Total	8435	8819	3154	1769	298
UK	Total	2678	1799	1589	1127	189
US	Total	4489	4928	5235	2409	481
IT	Central	139	210	254	120	41
IT	Not Central	18754	19413	5940	3542	653
DE	Federal Agency	424	296	274	90	47
DE	City	8011	8561	3011	1691	277
UK	Not Council	738	428	702	372	92
UK	Council	1940	1450	1150	845	166
US	Federal	928	780	898	583	120
US	Not Federal	3561	4582	6522	2089	542

noisy as it reflects the different administrative systems of the countries in our dataset, it is useful for gaining further insights in the comparisons. The number of responsive websites for each country and category is provided in the Websites column of Table 1.

3.2. Website names and access protocols

Let u denote the URL of a website, as appearing in a list of our dataset. When fetching u, the website may respond with zero or more HTTP redirections until a certain *landing page*. Let u' be the URL of such a landing page (it may be $u = u'$). We denote by *landing scheme* and by *landing name* the scheme and the first path segment, respectively, of u'.

We accessed each website twice, first with the `http` protocol and then with the `https` protocol, following in each case any possible redirection. In each case we recorded landing scheme, landing name and, in case the landing scheme was `https`, whether a *strict transport policy* was present. Thus, we associated each website with two tuples (landing_scheme, landing_name, sts), one for `http` access and the other for `https` access. A tuple is null if the website cannot be accessed with the corresponding scheme. A common scenario is one in which both tuples are identical, i.e., the website is available on `https` and responds to `http` requests with a redirection to `https`.

3.3. Dependency graph

Having collected the tuples described above for all the websites in our datasets, we collected all the information necessary for constructing a *dependency graph* that describes dependency relationships between entities of our interest: websites, landing names, zones, nameservers, IP /24 address ranges, autonomous systems. We consider all the IP addresses in the same /24 range as a single *network* entity. We defined such a graph as follows. A *node* represents one of the above entities. An oriented *arc* from a node n_1 to a node n_2 indicates that

the entity represented by the former *depends* on the entity represented by the latter. We defined a set of *dependency rules* aimed at capturing the nature of the security incidents of interest in this work, which operate on the mapping from service names to IP addresses and on the mapping of those addresses to a route. The full set of dependency rules is described below and is such that nodes whose inbound degree is zero are websites, while nodes whose outbound degree is zero (leaves) are autonomous systems. By ease of discussion, when we refer to a node we mean the entity represented by that node and vice versa, depending on the context. Furthermore, by type of a node we mean the type of the corresponding entity.

- A website depends on one or two landing names, one for each of its distinct and non-null landing_name attributes.
- A landing name n_1 depends on: (i) the zone that n_1 belongs to; and, (ii) either a network (when n_1 is mapped to an IP address by means of an A DNS resource record) or on another landing name (when n_1 is mapped to another landing name by means of a CNAME DNS resource record).
- Name servers are modelled like landing names, that is, let $ns_1.name$ be the name of a name server ns_1; ns_1 depends on: (i) the zone that $ns_1.name$ belongs to; and, (ii) either a network (when $ns_1.name$ is mapped by an A record) or on another name server (when $ns_1.name$ is mapped by a CNAME record).
- A zone depends on: (i) its name servers, (ii) its parent zone in the DNS tree, and (iii) the zones of the names of its name servers, with the following exceptions: a zone does not depend on itself; any zone that is a TLD is excluded from the dependency graph (analyzing dependencies in terms of TLDs does not provide any significant insights, as all or nearly all services in a country will depend on the corresponding TLD).
- A network depends on the autonomous system responsible for that address range. We mapped networks to autonomous systems based on a public database updated hourly that we downloaded in June 2020[2].
- An autonomous system is assumed to not depend on any other entity.

As an example, Figure 1 provides the portion of dependency graph containing entities and dependencies for website `www.units.it`. In Figure 1, both name servers of zone `garr.net` depend on the same network and two of the three name servers of zone `units.it` also depend on a single network. Dependency relations are transitive. Thus, all name servers of zones `garr.net` and `units.it` depend on the same autonomous system and the landing name of `www.units.it` also depends on that autonomous system.

[2] `https://iptoasn.com/`

Table 1 summarizes the number of entities from which websites in the corresponding category depend upon, either directly or indirectly. The total for each country is greater than the sum of categories because websites in different categories may depend on the same entity.

3.4. Name resolution paths and web access paths

We say that a path in the dependency graph that starts at a website and ends at an autonomous system is a *web path*. A web path that includes a name server is a *name resolution path*, otherwise it is a *web access path*. A name resolution path describes entities and dependencies involved in the name resolution procedure that a web client executes for obtaining the IP address of the website. A web access path describes entities and dependencies involved in the web access procedure that a web client executes for actually accessing the website, once the web client has acquired the IP address of the website.

We emphasize that a name resolution path is *not* a step-by-step description of the name resolution procedure: the former is merely a model of the entities and dependencies involved in the latter. Execution of the name resolution procedure usually involves further dependencies of the web client from other entities not modelled by the dependency graph, e.g., the address ranges and autonomous systems in the route between the web client and the name servers. Dependency from those entities is orthogonal to our analysis. The same remarks apply to the web access path w.r.t. the web access procedure and to the actual dependencies of web clients in web access procedures.

There are typically many name resolution paths for a given website because there are many ways for obtaining the IP address of that website, depending on the specific navigation in the DNS domain tree that the web client follows during the name resolution procedure and on the specific usage of DNS caching in that procedure. On the other hand, there are typically very few web access path for a given website—most often only one—and an execution of the web access procedure corresponds to a single web access path.

3.5. Threat model

We consider an attacker that may take control of one or more entities of the following types: zones, name servers, networks, autonomous systems. The method by which the attacker takes control of an entity is irrelevant. The attacker may control the behavior of a controlled entity in either of two ways:

- *Denial of Service*. The entity stops functioning. No procedure involving access to that entity can complete execution.
- *Impersonation*. If the entity is accessed in a name resolution procedure for a given website, then the procedure will return an IP address in control of the attacker. If the entity is accessed in a web access procedure, then the procedure will provide web content chosen by the attacker.

Let us assume that the attacker controls a node on a name resolution path for website w. The actual impact on w depends on: (i) whether there are name resolution paths for w that do *not* pass through nodes in control of the attacker; and, (ii) the number of name resolution procedure executions for w that actually pass only through those paths. For example, if the attacker controls the zone from which the landing name of w depends on *directly*, then *all* the name resolutions that cannot complete with a cached result obtained before the attack will be affected (this is the case of zone `cineca.it` in Figure 1). If, on the other hand, the attacker controls a zone from which the landing name of w depends on *indirectly*, then only those name resolutions that actually interact with that zone will be affected (zone `akam.net` in Figure 1).

We say that the *direct zones* for a website are the zones from which the website and its landing names depend on directly (e.g., zones `units.it` and `cineca.it` in Figure 1). Direct zones have a special focus in our analysis because every name resolution path and every web access path passes through a direct zone. Successful attacks on a direct zone are thus likely to have more impact on usage of the corresponding website than successful attacks on other zones.

Similar remarks can be made with respect to attacker-controlled nodes on a web access path, with the observation that redundant web access paths are infrequent and thus it is likely that most web access procedures will involve interacting with the entity in control of the attacker.

3.6. Limitations

Our approach considers a subset of the entities of interest for a website. In particular, we consider only the access to the home page and do not consider the actual content served [49, 50, 51]. Thus, there may be dependencies on other websites that are not part of our dependency graph. Furthermore, we do not consider the authentication infrastructures, which are likely an essential component of many websites. Although our framework may be extended to accommodate these missing entities easily, the analysis presented here certainly provides only a partial answer to the research questions of our interest.

As already observed in the previous section, we do not consider entities that are client-side, e.g., the address ranges and autonomous systems associated with the route between the web client and the name servers or the web servers. These entities may be important in assessing resiliency and security perimeter of a whole country.

Concerning our data collection methodology, its limitations are as follows: we considered only IPv4 addresses, i.e., we did not analyze AAAA records; we fetched records from a single location and only once (except for masking transient failures), which might fail to characterize website replication accurately and might not provide a fully accurate representation of the HTTPS configuration [52]; we did not record the additional URLs possibly involved in intermediate redirections from a website name to the corresponding landing name.

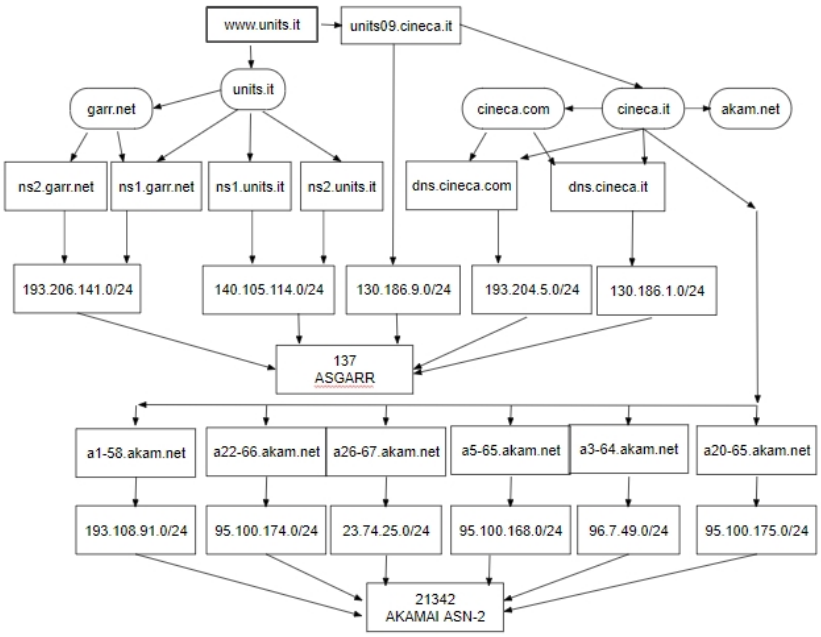

Figure 1: Portion of dependency graph containing entities and dependencies for `www.units.it`. Zones are represented as ovals for readability. Dependencies of zone `akam.net` from its name servers are omitted for simplicity.

.

4. Analysis

4.1. Redundancy of name resolution paths

We analyzed the redundancy of name resolution paths in websites by following a methodology similar to that applied in [22] for assessing DNS robustness of all *second level domains (SLDs)*. Given a zone z, we denote by $\#network(z)$ the number of different networks in which nameservers of z are distributed. The DNS specification requires that each zone maintains at least two nameservers [53] and that these nameservers be both geographically and topologically diverse [54]. The extent in which this robustness requirement is actually satisfied by all SLDs has been analyzed in [22], under the assumption that name servers in different /24 IP address ranges (different networks, in our terminology) are geographically and topologically diverse. A zone z exceeds the robustness requirement if $\#network(z) \geq 3$; meets the requirement if $\#network(z) = 2$; does not meet the requirement if $\#network(z) = 1$, i.e., all nameservers of zone z are concentrated in a single network. According to the cited work the percentage of SLDs in each category were 23%, 35%, 42%, respectively (these values refer to the SLDs that were actually responsive).

In order to analyze redundancy of name resolution paths at the level of *direct zones*, we adapted the above methodology to our context by categorizing websites as follows. For each website w, (i) we determined the set of direct zones $\mathcal{Z}(w)$ that w depends upon; (ii) for each zone $z \in \mathcal{Z}(w)$ we determined $\#network(z)$; (iii) we selected the minimum of those values, denoted $\#network_w^{min}$. Then, we partitioned websites in three sets based on their respective $\#network_w^{min}$ value: a website w belongs to the *all exceed* set if $\#network_w^{min} \geq 3$; w belongs to the *all meet* set if $\#network_w^{min} = 2$; w belongs to the *some do not meet* set if $\#network_w^{min} = 1$, i.e., when all zones in $\mathcal{Z}(w)$ have the respective nameservers concentrated in a single network. The size of each category is thus an indication of the overall preparedness of websites of a given country, from the point of view of name resolution, in tolerating attacks to one or two networks (the number of websites that could be affected by attacking a specific network does not emerge from this categorization and is analyzed in the next sections).

Figure 2 summarizes the results (the figure contains also the SLD data from [22]; we consider each SLD as equivalent to a website that depends on only one direct zone, i.e., the SLD itself). In all datasets, the fraction of websites in the *some do not meet* category is significantly smaller than in SLD; and, the fraction in the *some exceed* category is significantly larger. There are important differences between datasets, though: while virtually all DE and UK websites either meet or exceed the robustness requirement, 12% of US websites belong to the *some do not meet* category; most importantly, 20% in IT-Total and 25% in IT-Central belong to the *some do not meet* category. Furthermore, in UK, DE, US, the relative size of *some exceed* is larger for central datasets than for the country as a whole, while the opposite occurs in IT. Overall, IT websites appear to be less prepared than websites of the other datasets in tolerating attacks at name resolution targeted to one or two networks.

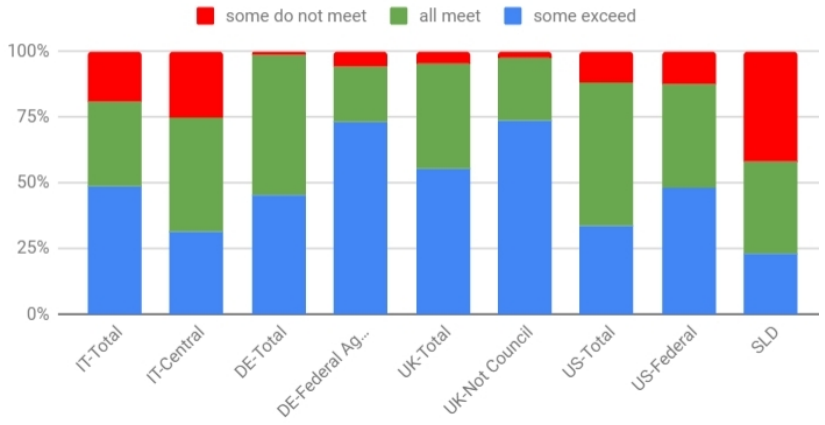

Figure 2: Website categorization based on redundancy of name resolution paths by networks.

We repeated the categorization by considering, for each website w, the set of *all* zones that w depends on rather than the set of direct zones only. The results were similar to those for direct zones, with the only notable exception that almost all UK websites belong to the *some do not meet* category. The reason for this rather surprising outcome is because almost all UK websites depend on zone `gov.uk` and this zone indirectly depends on zone `ns.uu.net` that does not meet the robustness requirement. However, since the nameservers of `gov.uk` are distributed across 7 different networks, as shown in the next section, attacks on the network where all `ns.uu.net` nameservers are placed are unlikely to be disruptive.

We analyzed redundancy of name resolution paths also at the level of *autonomous systems*, as follows. Given a zone z, we denote by $\#as(z)$ the number of different autonomous systems in which nameservers of z are distributed. We repeated the above categorization, for direct zones, in terms of $\#as(z)$ rather than $\#network(z)$ (Figure 3; the categorization of SLD in terms of autonomous systems was not performed in [22]). It can be seen that US websites and DE-Federal Agency are the two opposite extremes regarding redundancy of name resolution paths at the level of autonomous systems: almost 80% of US websites have nameservers for direct zones concentrated in a single autonomous system, while this fraction is only 18% for DE-Federal Agency. For IT and UK the corresponding fraction is around 40-55%.

Higher redundancy at the level of autonomous systems might not imply higher physical redundancy (and an enlarged security perimeter) because different autonomous systems could be managed by the same company and thus rely on the same technical infrastructure. For this reason, we performed the previous analysis also by categorizing websites in terms of the AS description field (a form of organization identifier) that describes each autonomous system. The results of this analysis turned out to be almost identical to the previous one, in which autonomous systems were identified by their AS numbers. While

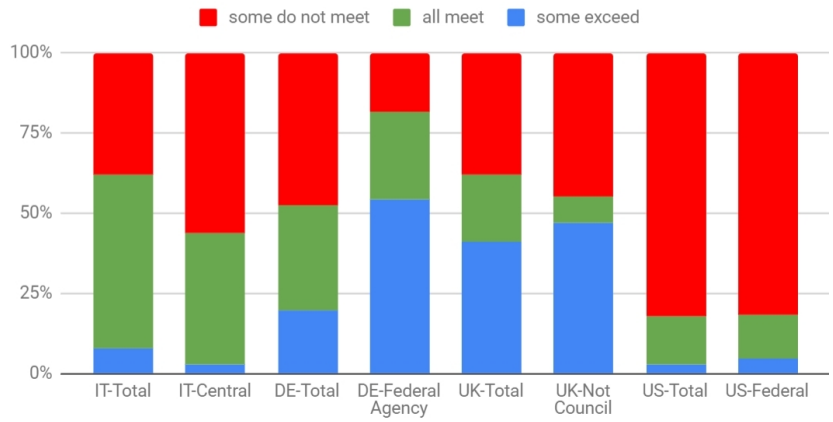

Figure 3: Website categorization based on redundancy of name resolution paths by autonomous systems.

it is not guaranteed that different values for the AS description necessarily implies a redundant infrastructure, this additional analysis confirms that diversity in terms of autonomous system number is indeed likely to correspond to real redundancy.

4.2. Distribution of name resolution paths

We analyzed the distribution of name resolution paths across zones, nameservers, networks, autonomous systems and determined the number of websites depending, in name resolution, on each of those entities. We considered only paths across zones with *direct* dependency because these are the paths where impact of an attack is higher (Section 3.4): attack on a direct zone will impact *all* the name resolution paths for the corresponding websites, whereas an attack on a nameserver, or a network or an autonomous system that depends on that zone will impact only the name resolution paths that pass across the attacked entity (we considered direct zones where all nameservers are concentrated in a single network in Section 4.2.3). This analysis allows determining the number of websites that could be affected by an attack on a single entity, as well as the entities that are potentially more critical for a country as a whole.

4.2.1. Distribution across zones and nameservers

The distribution of dependencies from zones is summarized in Figure 4, where we excluded UK websites for clarity as 96.7% and 92.1% of UK websites depend on zone `gov.uk` (full dataset and central dataset, respectively), while none of the other countries exhibits such a strong concentration of dependency on a single zone. For the UK dataset, thus, an important component of the overall security perimeter nation-wide can be identified clearly, which is an important property for focussing defense efforts. Furthermore, the fact that de-

	#nameservers	#networks	#AS	#AS countries
gov.uk.	7	7	3	3
gov.it.	2	1	1	1
bundestag.de.	3	3	3	1
bund.de.	5	3	2	1
bmas.de.	4	4	4	1
inqa.de.	4	4	4	1

Table 2: Architectural properties of top ranked direct zones.

pendency on `gov.uk` is so pervasive suggests a careful planning and coordination of activities across the numerous organizations responsible for UK websites.

Regarding the other datasets, in the full country datasets (Figure 4, left) there are no zones from which a significant fraction of websites depend on, with the only exception of two zones in the IT dataset—`edu.it` and `gov.it`. A somewhat more pronounced concentration can be observed in the central datasets, with 14% of IT central websites depending on `gov.it` and 5 zones of DE central directly responsible for 12.5%-5.2% of websites. The US dataset does not exhibit any significant concentration on any zone.

It is interesting to observe the architectural properties of top ranked direct zones (Table 2). All those zones but `gov.it.` exhibit high redundancy in terms of nameservers, of networks in which those nameservers are placed, of autonomous systems responsible for those networks. Such high redundancy is probably the result of a conscious design choice, based on the important role played by the corresponding zones. Zone `gov.it` is the top ranked zone for IT central and features only two nameservers in a single network. We find this architecture quite odd and not entirely justifiable in terms of the smaller security perimeter. Thus, it is unclear whether such a small redundancy is the result of a design choice or rather it just happened. The last column of Table 2 is based on the AS country field in the database that we used for mapping IP addresses to autonomous systems. All zones in this table depend on autonomous systems in the same country, except for `gov.uk` that also depends on the autonomous system of the Dutch research and education network and on one of a commercial US provider (Verizon). Dependency of zones on autonomous systems is analyzed in more detail in Section 4.2.2

4.2.2. Distribution across networks and autonomous systems

We analyzed the distribution of nameservers of direct zones across networks and autonomous systems and determined the number of websites depending, in name resolution, on each network and on each autonomous system. The results are summarized in Figure 5 and Figure 6, respectively. It can be seen that nearly all UK websites directly depend in name resolution on 7 networks and 3 autonomous systems. These networks and autonomous systems correspond to the nameservers of the zone `gov.uk`, that most UK websites depend on. This zone is thus highly critical for UK as a whole but, on other hand, such a

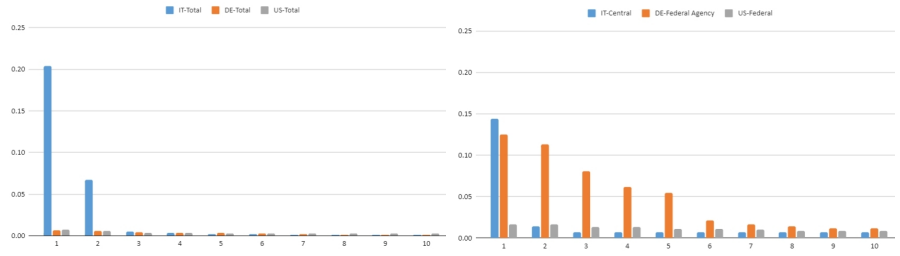

Figure 4: Direct dependency of websites from zones in name resolution. Each bar corresponds to a zone and its value is the fraction of websites that directly depends on a nameserver in that network (full country left, central websites right). UK websites are excluded for clarity, as almost all of them depend on zone `gov.uk`.

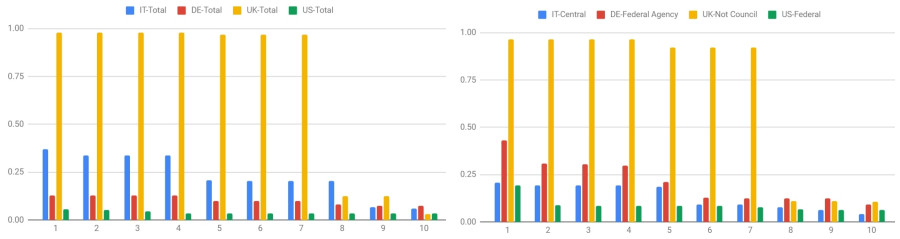

Figure 5: Direct dependency of websites from networks in name resolution. Each bar corresponds to a network and its value is the fraction of websites that directly depends on a nameserver in that network (full country left, central websites right).

strong concentration also corresponds to a small and clearly identified security perimeter to defend. Concerning the total datasets (Figure 5 and Figure 6 left), there are 8 networks and 6 autonomous systems quite critical for IT as 20%-37% of websites directly depend on them for name resolution, while the distributions for DE and US exhibit a much higher dispersion. With respect to central datasets (Figure 5 and Figure 6 right), there is a significant concentration for name resolution of DE websites on 5 networks and 5 autonomous systems, as 21%-43% of websites directly depend on those networks or autonomous systems.

A complementary and useful view of these results can be obtained by considering the autonomous system from which the top ranked networks depend on. Concerning the total datasets (Figure 7 left), it can be seen that the top 3 networks for IT, DE, UK all depend on a single autonomous system, that is thus especially critical for the country as a whole. Concentration for DE sites is particularly strong also for the top 5 networks, that all depend on a single autonomous system. Concerning central datasets (Figure 7 right), IT and DE exhibit more redundancy than with respect to the full country. On the other hand, UK and US exhibit equal or less redundancy as in the previous case. In other words, for IT and DE, the security perimeter in terms of autonomous systems tends to be larger for central websites than for the full country, while the opposite is true for UK and US.

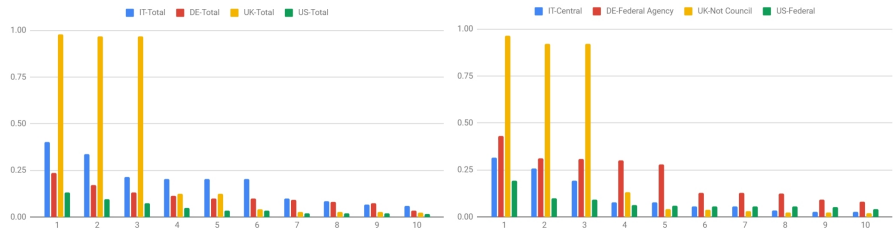

Figure 6: Direct dependency of websites from autonomous systems in name resolution. Each bar corresponds to an autonomous system and its value is the fraction of websites that directly depends on a nameserver in that network (full country left, central websites right).

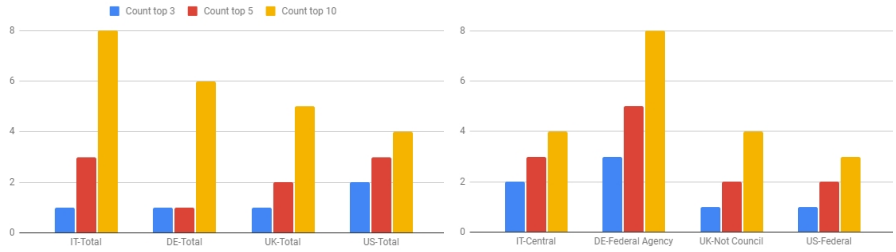

Figure 7: Count of different autonomous systems from which the top ranked networks in Figure 5 depend on (full country left, central websites right).

Further interesting insights can be obtained by looking at the details of the autonomous systems responsible for the top ranked networks (Table 3; descriptions have been obtained from `https://iptoasn.com` and slightly edited for clarity).

- Each country relies on autonomous systems associated with that country (with very few exceptions, for example the Dutch research network for UK websites).
- The role of private or for-profit companies is significantly not uniform across countries. In Germany, private companies are responsible for all the top ranked networks of the full dataset, but public organizations manage 4 of the 10 top ranked networks of the central dataset. Italy exhibits an opposite pattern: public organizations manage 3 of the top ranked networks of the full dataset while all the top networks of the central dataset are managed by for-profit companies. Not-for-profit organizations play a central role in UK, with the autonomous system of the national research and education network responsible for the 4 top ranked networks in UK, in both the full dataset and central dataset. Finally, no public organization emerges in the top ranked networks for US.
- The public research and education network infrastructure plays an important role in IT, DE, UK. The fact that those infrastructures are crucial

IT-Total		DE-Total		UK-Total		US-Total	
Aruba	1	ONEANDONE	1	JANET (**)	1	Cloudflare	2
Aruba	1	ONEANDONE	1	JANET (**)	1	Cloudflare	2
Aruba	1	ONEANDONE	1	JANET (**)	1	Akamai	4
Ktis (owned by Aruba)	2	ONEANDONE	1	JANET (**)	1	Netsolus	5
GARR (**)	3	ONEANDONE	1	MCI Verizon	2	Netsolus	5
CNR - TLD .it (**)	4	Hetzner	2	SURF - NL (**)	3	Tiggee	6
CNR - TLD .it (**)	6	Hosteurope	6	MCI Verizon	2	Tiggee	6
MIXITA	5	Hosteurope	3	Rackspace	4	Tiggee	6
Fastweb	8	Deutsche Telekom	4	Rackspace	5	Tiggee	6
Register	10	Netuse	9	Vodafone	6	Tiggee	6
IT-Central		DE-Federal Agency		UK-Not Councill		US-Federal	
Aruba	1	Hosteurope	1	JANET (**)	1	Akamai	1
Ktis (owned by Aruba)	2	ONEANDONE	2	JANET (**)	1	Akamai	1
Aruba	1	Internet AG	3	JANET (**)	1	Akamai	1
Aruba	1	InternetX	4	JANET (**)	1	Akamai	1
Fastweb	2	DFN - Research Network (**)	5	MCI Verizon	2	Cloudflare	2
Aruba	1	Knipp	7	SURF - NL (**)	3	Cloudflare	2
Aruba	1	Knipp	7	MCI Verizon	2	Akamai	1
BT Italia	4	Federal Office for Infosec (*)	8	Vodafone	4	Akamai	1
BT Italia	5	Federal Office for Infosec (*)	8	Vodafone	4	Amazon	3
Fastweb	2	Weather Service (*)	9	Vodafone	4	Amazon	3

Table 3: Autonomous system description of top ranked networks in Figure 5 (full country up, central websites down). The number next to each description is the rank of the autonomous system in the distribution in Figure 6. Asterisks indicate public organizations or not-for-profit companies; double asterisks indicate public research and education networks.

for the respective countries as a whole is somewhat surprising. Interestingly, the 6-th network of UK websites depend on the autonomous system responsible for the Dutch public research and education network, i.e., on a system of a foreign operator.

- Operators highly specialized in commercial *content distribution networks (CDN)* manage all the top ranked networks of US-central and the 3 top ranked networks of the full US dataset. These operators are particularly specialized in the management of highly scalable services and of defense from denial of service attacks. No similar operator emerges from the results for the other countries. Indeed, a recent analysis of hosting providers for 135000 government websites worldwide has recently shown that those sites tend to be privately hosted [55]. A CDN is certainly a powerful tool and the fact that major CDN operators are mostly US-based could be a key reason for the large adoption of CDNs in the US dataset but not in the other datasets. A broader analysis of the technological solutions available for defending against denial of service attacks in cloud environments and in data can be found in [56], while techniques for discriminating between denial of service attacks and sudden, massive usage of services are analyzed in [57, 58].

- The only operator with a mission explicitly tailored to national security appears to be the German Federal Office for Information Security, whose autonomous system manages the 8-th and 9-th top ranked networks of the central dataset.

4.2.3. Direct zones that do not meet the robustness requirement

We focussed on direct zones that do not meet the robustness requirement in terms of networks, i.e., on direct zones whose nameservers are all concentrated

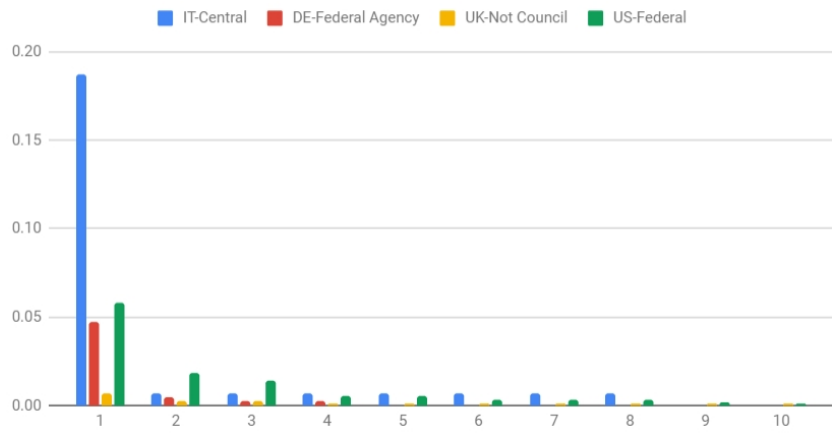

Figure 8: Direct dependency of websites from networks in name resolution (central websites that do not meet the robustness requirement). Each bar corresponds to a network and its value is the fraction of websites that directly depends on a nameserver in that network (note that for each of these websites all nameservers are in the same network).

in a single network: an attack on one of those networks would thus impact *all* name resolutions for the depending websites (Section 4.1).

The results for central websites are provided in Figure 8 (we omit the results for full countries as in that case there is no network from which a significant fraction of websites depend on). The key outcome is that one single network contains all the nameservers necessary for resolving almost 20% of central websites for IT. Controlling that single network, thus, implies fully controlling name resolution for almost 20% of IT-central websites.

4.3. Shared infrastructures for name resolution

In Section 4.2.2 we analyzed the distribution of nameservers of direct zones across single networks. In this section, we consider *sets* of networks that contain *all* nameservers of a given direct zone and determine the number of websites depending on each such set. This analysis thus allows determining the number of websites whose name resolutions could be fully controlled by an attack on a specific *set of networks*. We did not consider overlaps between different groups—i.e., a given network may belong to multiple groups. The impact of an attack on a specific group of networks, thus, could affect more websites than those resulting from our analysis.

We followed a methodology similar to that in [22], which executed a similar analysis for assessing the usage of such *shared infrastructures* for name resolution in SLDs. The cited work found a significant amount of sharing, both in terms of SLDs that share exactly the same set of nameservers (half the SLDs share the same set with at least 163 other SLDs) and of SLDs that share nameservers in the same network (half the SLDs share the same group with at least 3K

other SLDs, with 10 groups responsible for more than 20% of the more popular SLDs).

The results of our analysis in terms of groups of networks are in Figure 9. It can be seen that there is a significant usage of shared infrastructures for name resolution, although with usage pattern quite different among countries.

- Concerning IT websites, one single group of networks is fully responsible for name resolution of approximately 20% of websites, both for the full dataset and for the central dataset. The second and third top ranked groups are each responsible for more than 5% of websites in each dataset. Actual data show that the top ranked group is the same for both datasets and that the 3 top ranked groups are fully disjoint.

- Concerning UK websites, one single group of networks is fully responsible for name resolution of approximately 40% of websites, both for the full dataset and for the central dataset. The second and third top ranked groups are each responsible for more than 5% of websites in each dataset. Actual data show that the top ranked group is the same for both datasets and that the second ranked group is a superset of the top ranked one.

- Approximately 12% and 30% of DE websites (total dataset and central dataset, respectively) depend on a single group of networks. This group is different for the two datasets, though, with no overlap between the two groups.

- Usage of shared infrastructure is much less prevalent for the full US dataset than for the other countries. However, the top 5 groups are each responsible for more than 5% of websites in the central dataset. Actual data show that the top ranked group is the same for both datasets and that the 3 top ranked groups are fully disjoint.

Usage of shared infrastructure tends thus to be relatively more intensive in central websites than in the full country. This fact suggests that shared infrastructure are probably considered a useful tool for providing increased redundancy while carefully controlling the security perimeter. On the other hand, there is wide variability in the overlap between top ranked groups. In particular, it is worth emphasizing that the top ranked group is the same for the total dataset and the central dataset in IT, UK, US; the two groups are instead fully disjoint for DE.

Figure 10 provides the size, i.e., number of networks, in each of the groups of networks in Figure 10. The key outcome is the large size exhibited by the UK dataset: at least 7 different networks in each of the groups, with the only exception of the third top group in UK central that is composed of 4 networks. All the other datasets exhibit a group size in between 2 and 4 networks, with only two exceptions for US.

We repeated the previous analyses in terms of *autonomous systems*, i.e., we identified sets of autonomous systems that contain all nameservers of a given direct zone, determined the number of websites depending on each such set

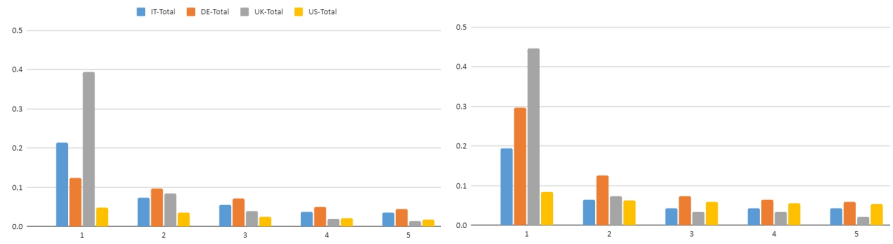

Figure 9: Direct dependency of websites from groups of networks in name resolution (full country left, central websites right). Each bar corresponds to a group of networks and its value is the fraction of websites that directly depend on a zone whose nameservers are all in that group of networks.

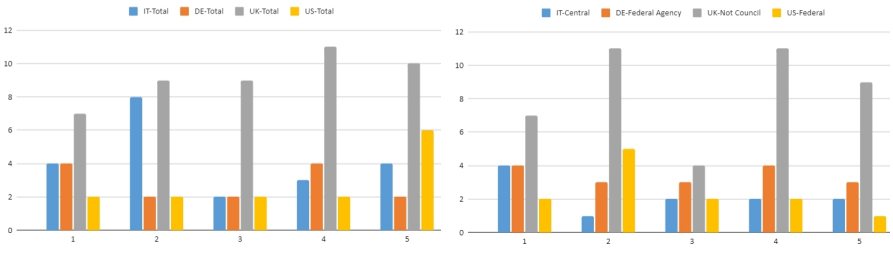

Figure 10: Size of top shared infrastructures for name resolution (full country left, central websites right). Each bar corresponds to one of the groups of networks in Figure 9 and indicates the number of networks in the corresponding group.

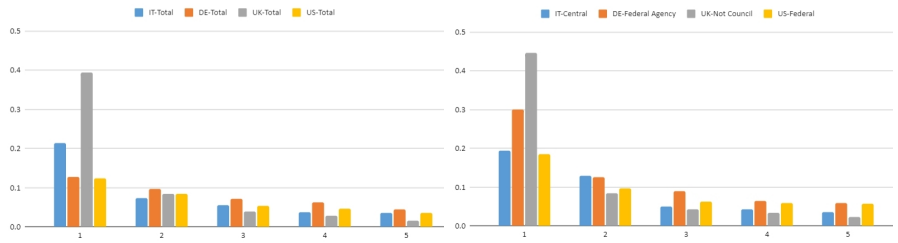

Figure 11: Direct dependency of websites from groups of autonomous systems in name resolution (full country left, central websites right). Each bar corresponds to a group of autonomous systems and its value is the fraction of websites that directly depend on a zone whose nameservers are all in that group of autonomous systems.

(Figure 11) and the number of autonomous systems in each set (Figure 12). The salient results of this analysis are:

- In the full country datasets, approximately 40% and 20% of UK and IT websites, respectively, depend on the top group of autonomous systems (Figure 11, left). The corresponding groups are composed of, respectively, 3 and 2 autonomous systems (Figure 12, left).
- A very similar pattern can be observed in central datasets of UK and IT (right figures).
- All the other datasets tend to exhibit a much smaller concentration over groups of autonomous systems, with the only exception of DE-Federal in which approximately 30% of websites depend on the top group (Figure 11, right). This group is composed of 4 autonomous systems (Figure 12, right).
- The UK websites tend to have the largest size of groups of autonomous systems while the US websites tend to have the smallest.
- The UK and DE websites tend to have largest size in their central datasets than in the respective fully country, while the IT and US websites tend to have the opposite structuring.

4.4. Redundancy of web access paths

We analyzed *redundancy of web access paths* based on the number of replicas of each website, under the assumption that each replica corresponds to an IP address. Specifically, for each website w, we determined the set of IP addresses that the landing names of the website directly depend upon (i.e., the IP addresses of the corresponding web servers). Figure 13 provides the percentage of websites with more than 2 replicas and with 2 replicas. The main difference is between UK/US and IT/DE datasets, with significant use of website replication for the former and very little use for the latter. Furthermore, for both UK

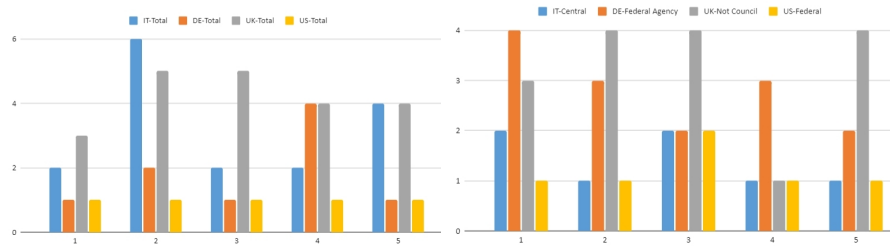

Figure 12: Size of top shared infrastructures for name resolution (full country left, central websites right). Each bar corresponds to one of the groups of autonomous systems in Figure 11 and indicates the number of autonomous systems in the corresponding group.

and US, the central datasets exhibit higher replication than the full country, with more than 30% of UK-Not Council websites associated with more than 2 replicas.

Next, we categorized websites with a methodology similar to that in Section 4.1, that is, based on whether the IP addresses of the website are concentrated in a single network, or are distributed in 2 networks, or in more than 2 networks. The corresponding categorization is in Figure 14. With respect to networks (left figure), we can observe that very few of the replicated UK websites have all replicas placed on a single network. Furthermore, the majority of replicated UK websites have the replicas distributed across more than 2 networks. In all the other datasets only ≈10% of the replicated websites have their replicas concentrated in a single network, with the only exception of IT-Central in which such a fraction is much higher (75%). Finally, while a large fraction of replicated UK websites are distributed across more than 2 networks (more than 50%), US tend to concentrate replicas mostly on 2 networks.

Distribution of website replicas across different autonomous systems is very small (Figure 14, right). While the results for IT and DE are not especially significant due to the low use of replication for those datasets (Figure 13), it is interesting to notice that some 5% of replicated websites in both the central datasets and the full country, of both UK and US, distribute their replicas across different autonomous systems.

4.5. Shared infrastructure for web access

In this section we analyze usage of *shared infrastructure for web access*, that is, we identify *sets* of single IP addresses, of networks and of autonomous systems, that contain all replicas of a given website and determined the number of websites depending on each such set.

The results in terms of single IP addresses are in Figure 15 (datasets for full countries are not shown as even the top groups are responsible for less than 5% of the websites). A significant fraction of UK-Not Council websites (almost 25%) depend on a group of 4 IP addresses and one IP address is responsible for 5% of those websites. There are 5 IP addresses each responsible for all replicas

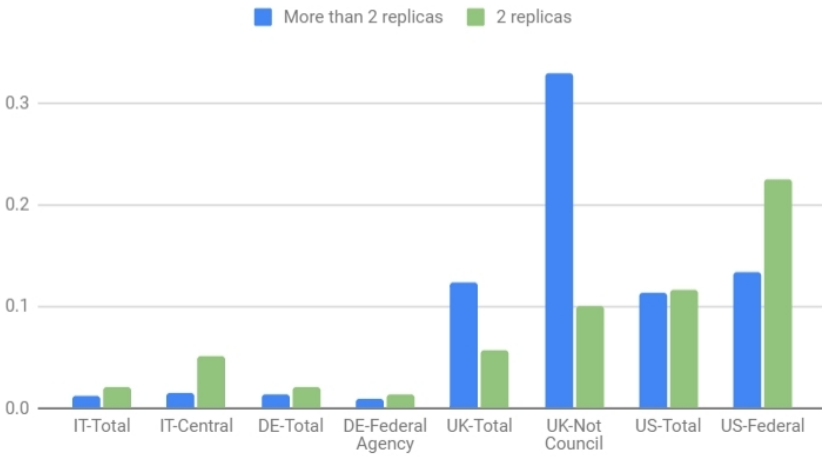

Figure 13: Percentage of websites with more than 2 replicas and with 2 replicas.

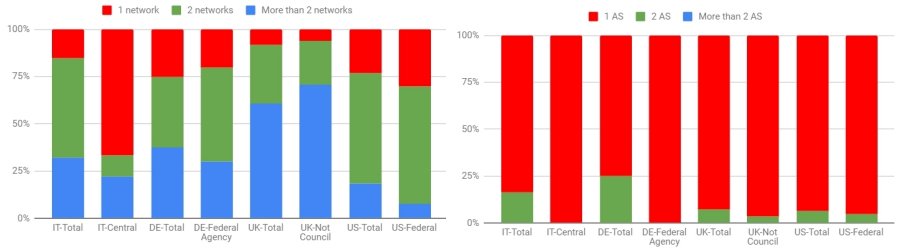

Figure 14: Categorization of replicated websites based on the number of networks (left) and of autonomous systems (right) on which the replicas are placed. Percentage values are computed with respect to replicated websites, i.e., not the full dataset.

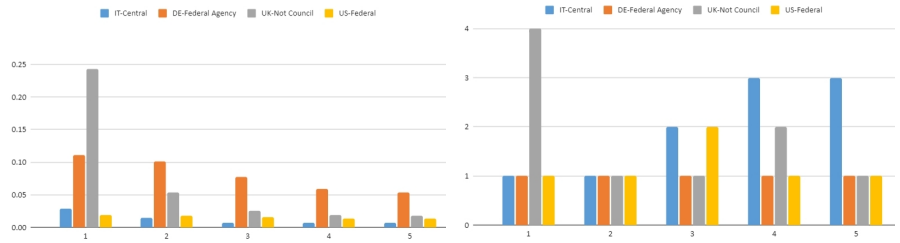

Figure 15: Dependency of websites from groups of IP addresses (central datasets only). Left graph: each bar corresponds to a group of IP addresses and its value is the fraction of websites whose replicas are all in that group of IP addresses (websites that are not replicated are considered as having only one replica). Right graph: each bar corresponds to a bar on the left graph and its value is the number of IP addresses in the corresponding group.

of 5-10% of DE-Federal Agency websites. Neither IT-Central nor US-Federal exhibits IP addresses responsible for more than 2.5% of websites each.

Figure 16 provides a complementary perspective in terms of groups of networks (datasets for full countries not shown for the same reason as above). It can be seen that:

- A group of 4 networks is responsible for almost 25% of UK-Not Council websites. By looking at raw values, it turns out that this group corresponds to the top group of 4 IP addresses previously identified in Figure 15, that is, those IP addresses are all spread in different networks. All those networks correspond to the same autonomous system.
- All the other top groups of networks are composed of one single network (with the only exception of the 5-th ranked group of UK-Not Council).
- Each of the top ranked groups for DE-Federal Agency is responsible for all replicas of 10-15% of websites.

Finally, Figure 17 provides the results in terms of groups of autonomous systems. Each group in the figure is composed of only one autonomous system. All the datasets exhibit significant dependence on single autonomous systems—e.g., the top ranked autonomous system for IT contains all replicas of more than 30% of websites. Interestingly, for all the central datasets, the top ranked autonomous system is responsible for more than 20% of websites.

4.6. Summary of architectural properties

We attempt to summarize the key insights of our architectural analysis:

- Redundancy of name resolution paths tends to be higher than that of all second level domains and, except for IT, relatively higher for central datasets than for the full country (Figure 3). Such a redundancy tends to occur also in terms of autonomous systems, with the exception of US datasets (Figure 3).

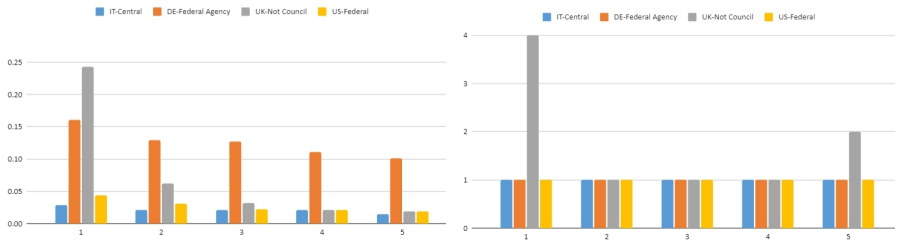

Figure 16: Dependency of websites from groups of networks (central datasets only). Left graph: each bar corresponds to a group of networks and its value is the fraction of websites whose replicas are all in that group of networks (websites that are not replicated are considered as having only one replica). Right graph: each bar corresponds to a bar on the left graph and its value is the number of networks in the corresponding group.

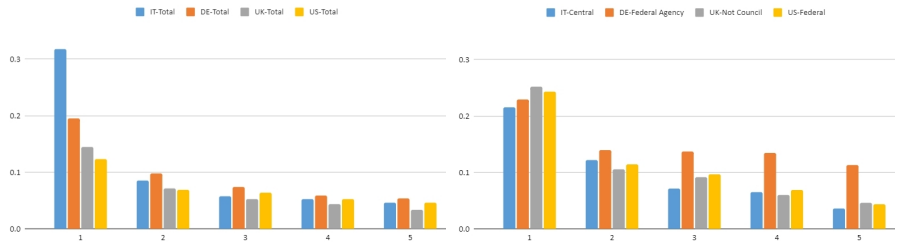

Figure 17: Dependency of websites from groups of autonomous systems (full country left, central datasets right). Each bar corresponds to a group of autonomous systems and its value is the fraction of websites whose replicas are all in that group (websites that are not replicated are considered as having only one replica). All the groups in this graph are composed of one autonomous system.

- Zones that are most critical in terms of number of depending websites are highly replicated, even in terms of autonomous systems (Table 2). The only significant exception to this observation is zone `gov.it`, that has only 2 nameservers in the same network.

- Almost all websites in UK datasets directly depend on zone `gov.uk`, that exhibits high redundancy at all levels. We believe these two properties are a consequence of a centralized design choice and a careful planning aimed at obtaining high resiliency to denial of service attacks with a clearly identified security perimeter to defend.

- Networks that are most critical for name resolution of central websites tend to be spread across different autonomous systems more for IT/DE than for UK/US (Figure 3). No single network is highly critical in this respect, with the exception of one network that contains all nameservers from which almost 20% of central websites of IT dataset depend on (Figure 8).

- Networks most critical for name resolution in a given country tend to be managed by an autonomous system of that country (Table 3). The role of for-profit companies vs public organizations is not uniform across countries: not significant in US, not significant in IT-Central but significant in IT-Total, significant in DE-Central but not in DE-Total, significant in UK. CDNs play a crucial role in US but not the other datasets.

- Usage of groups of networks or autonomous systems that contain all nameservers of a zone is significant in all datasets, with more prevalence for central datasets than for the respective full country. This structuring allows providing increased redundancy while carefully controlling the security perimeter.

- Usage of groups of networks that contain all replicas of a website is significant only in UK, where a group of 4 networks is responsible for almost 25% of websites.

- Website replication at the level of IP address is almost negligible for IT/DE and in the order of 10-20% for UK/US websites. Such a replication tends to be applied also at the level of networks but not of autonomous systems.

- Website replication across different autonomous systems is negligible in all datasets. In each country, the most critical autonomous system is responsible for more than 20% of central websites.

4.7. Deployment of defensive mechanisms

In this section we analyze the deployment of basic defensive mechanisms against impersonation attacks: *Route Origin Authorization (ROA)* in the Border Gateway Protocol (BGP), *Domain Name System Security Extensions (DNSSEC)* and *Strict Transport Security (HSTS)* in HTTP. Each of those complementary

and independent mechanisms, when applied server-side and enforced client-side, transforms an impersonation attack to a denial-of-service attack.

4.7.1. BGP Route Origin Authorization

BGP is highly vulnerable to attacks where an autonomous system announces routes for IP addresses it does not control [33, 36, 37]. The Resource Public Key Infrastructure (RPKI) addresses this issue by means of cryptographic signatures which limit the set of entities that can announce IP prefixes in route advertisements [38]. Such signatures are applied to Route Origin Authorization (ROA) objects that are published in public repositories and authorize an autonomous system to announce certain IP prefixes. Autonomous systems are supposed to implement Route Origin Validation (ROV) that consists in downloading ROA objects and use the resulting information for validating received BGP announcements.

Assessing the effectiveness of RPKI for a given client-server interaction is very hard as it depends on the actual ROA and ROV deployment on the corresponding routing path, as well as on the relative location of the autonomous system advertising malicious routing messages with respect to the other autonomous systems in the path. In this section we assess the actual ROA deployment at two carefully selected groups of networks in each dataset: the top 5 networks with highest direct dependency in name resolution (Figure 5); and, the top 5 shared networks for web access (Figure 16). The defensive effectiveness of a ROA for those networks is hard to assess, for the reasons just outlined; however, networks without any ROA do not have any such defense at the BGP level. While not exhaustive, this analysis does provide useful insights into the defensive standpoint at the BGP level for the various datasets.

We obtain ROA data for the networks of our interest from the APNIC ROA generation report tool[3], a publicly available website updated daily and reporting measurements obtained from 600 distinct vantage points (see [59] for details). For each network we determined in March 2021 whether it is classified as VLD, meaning that a ROA set is available and that ROA validated the route for that network, or UNK meaning that no ROA matched that network. To place the obtained data in some perspective, the percentage of VLD routes in Europe and US is 36% and 12%, respectively. The results are summarized in Figure 18 (each network was at least in the 90-percentile of visibility by all the 600 measurement points).

Concerning name resolution (left graph), it is evident in all datasets but the IT ones a strong attention to BGP level defense, as all bars exhibit the maximum possible value (except for one of the networks in UK). We remark that we are considering networks with direct dependency in name resolution, thus a very large fraction of accesses to the depending websites would be affected by a successful impersonation attack toward those networks. Such an attack could be defeated at the HTTPS level, i.e., an attacker that impersonates a

[3] `https://stats.labs.apnic.net/roas`.

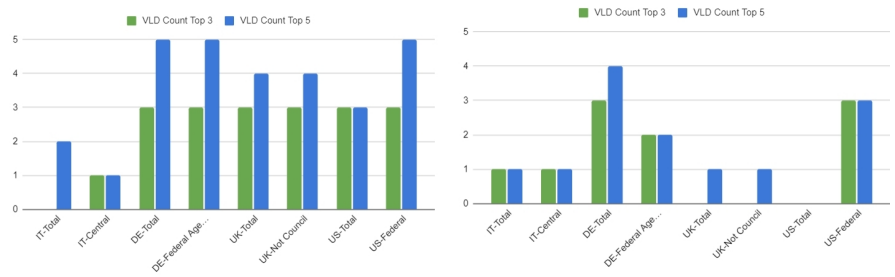

Figure 18: Count of networks with BGP Route Authorization Origin. Left graph: top networks for direct dependency in name resolution (Figure 5). Right graph: top groups of networks for web access (Figure 16; groups with more than one network are counted as a single network without a valid ROA, as it turned out that all networks in each group fall in this category).

name server by means of a BGP attack could provide a response pointing to an attacker-controlled website and the subsequent impersonation attack could be detected by an effective HTTPS defense. However, presence of a BGP level defense is an excellent application of the defense in depth principle and there are many possibilities for circumventing HTTPS defenses, e.g., phishing. Thus, the results for IT datasets appear in this respect disappointing.

Concerning shared networks for web access, i.e., networks hosting all replicas of a website (right graph), ROA deployment is much less pervasive: optimal or near-optimal results are exhibited only by DE-Total and US-Federal (top 3 networks only). As observed in the previous paragraph, absence of ROA implies a missed opportunity for a strong defense in depth.

4.7.2. DNS Security Extensions (DNSSEC)

DNSSEC is a DNS security extension aimed at providing authentication and integrity of DNS data (as well as authenticated denial of existence). Such guarantees are provided by means of public key cryptography: a DNSSEC-signed zone publishes its public key and signs all the zone data with the corresponding private key. A DNSSEC-enabled client may thus detect and discard fake DNS data provided by an attacker that impersonates the nameserver of the zone of those data (or that provides those fake data in the form of cached responses).

DNSSEC deployment and maintenance is complex [60, 61, 62] and many operators are reluctant to adopt this technology because its overall costs tend to be perceived as greater than its benefits (a concise yet highly useful summary of this issue can be found in [63]). DNSSEC is currently deployed at a small percentage of all the DNS zones and its deployment is uncommon even in zones highly ranked in terms of user traffic [64]. Furthermore, DNSSEC support client side is not ubiquitous: end systems usually do not validate DNSSEC responses while only a fraction of DNS resolvers worldwide do so. It follows that a client may be tricked in using a fake DNS response even for a DNSSEC-signed zone because the corresponding DNS query happened to follow a path among DNS resolvers in which DNSSEC fetch and validation is not fully implemented.

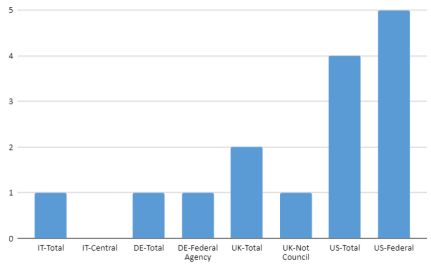

Figure 19: Count of DNSSEC-signed zones at the top 5 zones with highest direct dependency in name resolution, for each dataset (Figure 4).

We determined the actual DNSSEC deployment at the top 5 zones with highest direct dependency in name resolution, for each dataset (Figure 4). We obtained the corresponding data from the Verisign Lab DNSSEC debugger[4] The results are summarized in Figure 19. To place these data in perspective, we remark that only $\approx$ 1% of all second-level domains in the DNS subtrees `.com`, `.org` and `.net` is DNSSEC-signed [64]. It can be seen that nearly all the top zones in the US dataset are DNSSEC-signed. Concerning UK, nearly all the zones directly depend on `gov.uk` (Section 4.2.2) and this zone is DNSSEC-signed. Usage of DNSSEC in the IT and DE top zones is rare.

4.7.3. HTTPS support in web access paths

As described in Section 3.2 in detail, we accessed each website with the `http` protocol, with the `https` protocol and determined whether a *Strict Transport Security policy (HSTS)* was present in HTTP responses. HSTS is a security enhancement specified by a web application through a dedicated HTTP response header: when a browser receives this header, that browser will prevent any communications to that web application from being sent over HTTP and will instead send all communications over HTTPS. Modern browsers have a preloaded list of websites for which an HSTS policy is applied by default, even without having contacted the website earlier. We have not checked the presence of websites in our datasets in such lists and assumed that HSTS policies can only be acquired by contacting a website.

The threat model of HSTS is impersonation attacks executed when the browser accesses a website available on `https` with the `http` protocol. In those attacks, either the attacker redirects the browser to another attacker-controlled (and possibly https) website, or the attacker forces the browser to use http for serving attacker-controlled content (SSLStrip attack). HSTS is effective toward those attacks because a browser with an HSTS policy for a website will never access that website with `http`—execution of an impersonation attacks would thus correspond to a denial of service.

[4] `https://dnssec-debugger.verisignlabs.com`.

We summarized the results by categorizing websites as follows:

- *secure+almost* the website is available on `https` and has a HSTS policy; either the website is only available in `https` or `http` access is redirected to `https`. The name of this category reflects the fact that a browser is vulnerable only when it does not have any HSTS policy in place for the website, i.e., only before the first visit.
- *https_strip* like secure+almost but the website does not have any HSTS policy.
- *no_redirect* the website is available on `https` but `http` access is not redirected to `https`.
- *http_only* the website is not available on `https`.

Figure 20 provides the composition of the resulting categories. To place HSTS data in perspective, in March 2021 W3Tech estimated HSTS usage at 19% of all websites, with a growth of 7% over the past year[5]. We observe what follows.

- Usage of HSTS is very high in the central datasets of UK (approximately 60%) and especially US (more than 80%). DE datasets, UK and US full countries are around 25%. IT datasets are well below those values. After our data collection, usage of HSTS for US datasets has probably increased even more [65].
- The fact that central datasets of UK and US exhibit values much higher than the respective full country is a signal of careful planning and correct technical administration.
- DE, UK and US all exhibit a small prevalence of *no_redirect* at the full country level and an almost negligible one on the respective central datasets, which is signal of more careful administration. IT datasets instead exhibit 29% and 24%: values sufficiently large to not convey a positive signal in this respect.
- Websites that do not support `https` are negligible in US Federal and approximately 10% in US full country; 9% and 19% in UK, central and full country respectively; approximately 15% in DE and 25% in IT. To place these results in perspective, in 2020 `https` usage was above 80% for the top-ranked Alexa websites [23] and around 40% for a dataset of 135000 government websites worldwide [55].

The recent decision of the Chromium developers to use `https` as default access protocol rather than `http` [66] will provide users of Chromium-based browsers of an effective defense against the impersonation attacks considered here, even for websites that do not use HSTS (https_strip), for those that are

[5] `https://w3techs.com/technologies/details/ce-hsts`

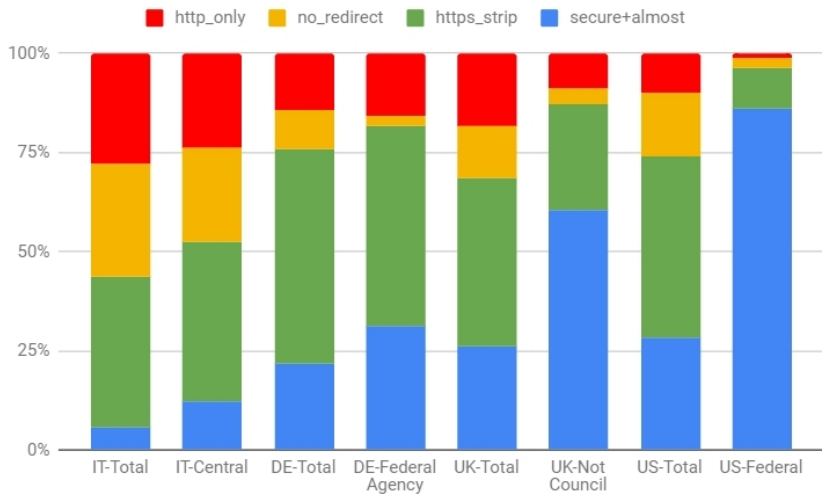

Figure 20: HTTPS support in web access paths. Each bar corresponds to a network and its value is the fraction of websites that directly depends on a nameserver in that network.

not configured correctly (no_redirect), for those that do use HSTS but are not preloaded in browsers (secure+almost). The results of our analysis demonstrate that there are plenty of such websites in our datasets.

4.8. Summary of deployment of defensive mechanisms

Regarding the deployment of the defensive mechanisms analyzed, the key outcome is as follows:

- Usage of BGP ROA in the most critical networks for name resolution is very high in all datasets, except for IT. Its usage in the most critical networks for web access is instead much lower, except for DE-Total and UK-Federal.
- Deployment of DNSSEC in the zones with highest direct dependency is almost ubiquitous in the US and UK datasets, while it is almost negligible in IT and DE.
- Usage of HTTPS is very high, except for IT datasets. Usage of HSTS is very high in US/UK central datasets and higher than average, except for IT datasets. Overall, there are significant fractions of websites without any specific defense against impersonation attacks.

Although the defensive mechanisms considered provide important security guarantees, the fact that they are not deployed everywhere should not be interpreted necessarily as a negative feature. Every defensive mechanism comes with a cost and it is not possible to ascertain, from our viewpoint, whether investing in one of those mechanisms is a rational usage of the defensive budget

available to a given organization. Furthermore, occasional misconfigurations of those mechanisms could result in a denial of service even without any actual impersonation attack—e.g. due to an expired HTTPS certificate or to a zone-signing key updated incorrectly. Even by taking these considerations into account, though, it seems reasonable to expect at least an ubiquitous deployment of HTTPS in the websites of interest in this work.

5. Concluding remarks

We have examined the robustness of access paths to websites of public interest, in realistic threat models, at the level of a full country. We have not compared the outcome for each analyzed country against any ideal set of design rules that should be enforced country-wide—the formulation of such rules is beyond the scope of this work, as pointed out in the introduction. The inherent tension between high redundancy in access paths and the need of keeping the defense perimeter as small as possible, implies that there is no single correct recipe in this respect and that a wide range of different design choices can be made. Indeed, significant differences between countries have emerged from this point of view. The numerous properties that we have extracted from the various datasets should not be summarized in any synthetic score or ranking, we believe, for several reasons: (i) the relative relevance of the two conflicting objectives high redundancy vs small perimeter depends on the type of entity to be replicated —e.g., a network vs an autonomous system; (ii) low redundancy in access path is not necessarily a signal of low resiliency to attacks—e.g., a single network or autonomous system could be equipped with state of the art defense against denial of service attacks; (iii) the ability to defend an entity with low redundancy may strongly depend on the technology and human resources available in a country—hence a given design choice could be rational for a certain country and not rational for a different one.

We believe that our survey provides useful and important insights into the reality of the web infrastructures of public interest, including in particular the structural interdependencies at the level of a full country, an issue whose relevance can only grow. Our methodology may constitute a practical and sound framework for performing similar analyses on large collections of websites of public interest.

[1] C. C. Demchak and Y. Shavitt, "China's maxim – leave no access point unexploited: The hidden story of china telecom's BGP hijacking," *Military Cyber Affairs*, vol. 3, no. 1, p. 7, 2018.

[2] C. Cimpanu, "Russian telco hijacks internet traffic for google, AWS, cloudflare, and others." `https://www.zdnet.com/article/russian-telco-hijacks-internet-traffic-for-google-aws-cloudflare-and-others/`. Accessed: 2021-3-31.

[3] D. Goodin, "How 3ve's BGP hijackers eluded the internet—and made $29m," *Ars Technica*, Dec. 2018.

[4] C. Cimpanu, "DNS hijacks at two cryptocurrency sites point the finger at GoDaddy, again." https://therecord.media/two-cryptocurrency-portals-are-experiencing-a-dns-hijack-at-the-same-time/, Mar. 2021. Accessed: 2021-3-31.

[5] W. Mercer, "DNS hijacking abuses trust in core internet service," *Talos Intelligence*, Apr. 2019.

[6] C. Cimpanu, "Hackers breached greece's top-level domain registrar." https://www.zdnet.com/article/hackers-breached-greeces-top-level-domain-registrar/. Accessed: 2021-3-31.

[7] M. Hirani, "Global DNS hijacking campaign: DNS record manipulation at scale." https://www.fireeye.com/blog/threat-research/2019/01/global-dns-hijacking-campaign-dns-record-manipulation-at-scale.html. Accessed: 2021-3-31.

[8] P. Rascagneres, "DNSpionage campaign targets middle east," *Talos Intelligence*, Nov. 2018.

[9] D. Madory, "BGP / DNS hijacks target payment systems," *Oracle Internet Intelligence*, Aug. 2018.

[10] B. Kreb, "Hacked cameras, DVRs powered today's massive internet outage," *Krebs on Security*, Oct. 2016.

[11] D. Goodin, "Foul-mouthed worm takes control of wireless ISPs around the globe." https://arstechnica.com/information-technology/2016/05/foul-mouthed-worm-takes-control-of-wireless-isps-around-the-globe/, May 2016. Accessed: 2021-3-31.

[12] C. Cimpanu, "Ransomware gang demands $7.5 million from argentinian ISP." https://www.zdnet.com/article/ransomware-gang-demands-7-5-million-from-argentinian-isp/. Accessed: 2021-3-31.

[13] C. Cimpanu, "Hackers breached A1 telekom, austria's largest ISP." https://www.zdnet.com/article/hackers-breached-a1-telekom-austrias-largest-isp/, Mar. 2021. Accessed: 2020-6-11.

[14] "Tiscali: attacco hacker provoca disservizi web e rete fissa," tech. rep. Accessed: 2021-3-31.

[15] A. Bannister, "INPS hack: Italy's social security website back online following cyber-attack claims." https://portswigger.net/daily-swig/inps-hack-italys-social-security-website-back-online-following-cyber-attack-claims, Apr. 2020. Accessed: 2021-9-17.

[16] J. Jay, "DDoS attacks took down italy's social security website." `https://www.teiss.co.uk/ddos-attacks-italy-inps-website/`, Apr. 2020. Accessed: 2021-9-17.

[17] M. Y. Vardi, "Efficiency vs. resilience: what covid-19 teaches computing," 2020.

[18] S. Liu, Z. S. Bischof, I. Madan, P. K. Chan, and F. E. Bustamante, "Out of sight, not out of mind: A User-View on the criticality of the submarine cable network," in *Proceedings of the ACM Internet Measurement Conference*, IMC '20, (New York, NY, USA), pp. 194–200, Association for Computing Machinery, Oct. 2020.

[19] J. Mayer, V. Sahakian, E. Hooft, D. Toomey, and R. Durairajan, "On the resilience of internet infrastructures in pacific northwest to earthquakes," in *Passive and Active Measurements - PAM*, 2021.

[20] G. C. M. Moura, S. Castro, W. Hardaker, M. Wullink, and C. Hesselman, "Clouding up the internet: how centralized is DNS traffic becoming?," in *Proceedings of the ACM Internet Measurement Conference*, IMC '20, (New York, NY, USA), pp. 42–49, Association for Computing Machinery, Oct. 2020.

[21] L. Zembruzki, A. S. Jacobs, G. S. Landtreter, L. Z. Granville, and G. C. M. Moura, "Measuring centralization of DNS infrastructure in the wild," in *Advanced Information Networking and Applications*, pp. 871–882, Springer International Publishing, 2020.

[22] M. Allman, "Comments on DNS robustness," in *Proceedings of the Internet Measurement Conference 2018*, IMC '18, (New York, NY, USA), pp. 84–90, ACM, 2018.

[23] A. Kashaf, V. Sekar, and Y. Agarwal, "Analyzing third party service dependencies in modern web services: Have we learned from the Mirai-Dyn incident?," in *Proceedings of the ACM Internet Measurement Conference*, IMC '20, (New York, NY, USA), pp. 634–647, Association for Computing Machinery, Oct. 2020.

[24] V. Ramasubramanian and E. G. Sirer, "Perils of transitive trust in the domain name system," in *Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement*, IMC '05, (Berkeley, CA, USA), pp. 35–35, 2005.

[25] C. Deccio, C. Chen, P. Mohapatra, J. Sedayao, and K. Kant, "Quality of name resolution in the domain name system," in *2009 17th IEEE International Conference on Network Protocols*, pp. 113–122, Oct. 2009.

[26] P.-A. Vervier, O. Thonnard, and M. Dacier, "Mind your blocks: On the stealthiness of malicious BGP hijacks," in *Proceedings 2015 Network and Distributed System Security Symposium*, (Reston, VA), Internet Society, 2015.

[27] P. Moriano, S. Achar, and L. J. Camp, "Incompetents, criminals, or spies: Macroeconomic analysis of routing anomalies," *Computers & Security*, vol. 70, pp. 319–334, 2017.

[28] K. Sriram and D. Montgomery, "Resilient interdomain traffic exchange:: BGP security and DDos mitigation," tech. rep., National Institute of Standards and Technology, Gaithersburg, MD, Dec. 2019.

[29] P. Litke and J. Stewart, "BGP hijacking for cryptocurrency profit," *Dell SecureWorks Counter Threat Unit*, Aug. 2014.

[30] M. Apostolaki, A. Zohar, and L. Vanbever, "Hijacking bitcoin: Routing attacks on cryptocurrencies," in *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 375–392, May 2017.

[31] H. Birge-Lee, Y. Sun, A. Edmundson, J. Rexford, and P. Mittal, "Bamboozling certificate authorities with BGP," in *27th USENIX Security Symposium (USENIX Security 18)*, pp. 833–849, 2018.

[32] F. Douzet, L. Pétiniaud, L. Salamatian, K. Limonier, K. Salamatian, and T. Alchus, "Measuring the fragmentation of the internet: The case of the border gateway protocol (BGP) during the ukrainian crisis," in *2020 12th International Conference on Cyber Conflict (CyCon)*, vol. 1300, pp. 157–182, ieeexplore.ieee.org, May 2020.

[33] G. Huston, M. Rossi, and G. Armitage, "Securing BGP — a literature survey," *IEEE Communications Surveys Tutorials*, vol. 13, no. 2, pp. 199–222, 2011.

[34] A. Mitseva, A. Panchenko, and T. Engel, "The state of affairs in bgp security: A survey of attacks and defenses," *Computer Communications*, vol. 124, pp. 45–60, 2018.

[35] Y. Sun, M. Apostolaki, H. Birge-Lee, L. Vanbever, J. Rexford, M. Chiang, and P. Mittal, "Securing internet applications from routing attacks," *Communications of the ACM*, vol. 64, no. 6, pp. 86–96, 2021.

[36] P. Sermpezis, V. Kotronis, P. Gigis, X. Dimitropoulos, D. Cicalese, A. King, and A. Dainotti, "ARTEMIS: Neutralizing BGP hijacking within a minute," *IEEE/ACM Trans. Netw.*, vol. 26, pp. 2471–2486, Dec. 2018.

[37] S. Cho, R. Fontugne, K. Cho, A. Dainotti, and P. Gill, "BGP hijacking classification," in *Network Traffic Measurement and Analysis Conference (TMA)*, June 2019.

[38] T. Chung, E. Aben, T. Bruijnzeels, B. Chandrasekaran, D. Choffnes, D. Levin, B. M. Maggs, A. Mislove, R. van Rijswijk-Deij, J. Rula, and N. Sullivan, "RPKI is coming of age: A longitudinal study of RPKI deployment and invalid route origins," in *Proceedings of the Internet Measurement Conference*, IMC '19, (New York, NY, USA), pp. 406–419, Association for Computing Machinery, Oct. 2019.

[39] K. Kirkpatrick, "Fixing the internet," *Commun. ACM*, vol. 64, pp. 16–17, July 2021.

[40] R. Bloomfield, L. Buzna, P. Popov, K. Salako, and D. Wright, "Stochastic modelling of the effects of interdependencies between critical infrastructure," in *Critical Information Infrastructures Security*, pp. 201–212, Springer Berlin Heidelberg, 2010.

[41] A. De Nicola, M. L. Villani, M. C. Brugnoli, and G. D'Agostino, "A methodology for modeling and measuring interdependencies of information and communications systems used for public administration and egovernment services," *Int. J. Crit. Infrastruct. Prot.*, vol. 14, pp. 18–27, Sept. 2016.

[42] E. Casalicchio and E. Galli, "Metrics for quantifying interdependencies," in *Critical Infrastructure Protection II*, pp. 215–227, Springer US, 2008.

[43] D. Geer, B. Tozer, and J. S. Meyers, "For good measure: Counting broken links: A quant's view of software supply chain security," *USENIX ;login:*, vol. 45, no. 4, 2020.

[44] A. Chancusi, P. Diestra, and D. Nicolalde, "Vulnerability analysis of the exposed public IPs in a higher education institution," in *2020 the 10th International Conference on Communication and Network Security*, ICCNS 2020, (New York, NY, USA), pp. 83–90, Association for Computing Machinery, Nov. 2020.

[45] "Advanced persistent threat compromise of government agencies, critical infrastructure, and private sector organizations." `https://us-cert.cisa.gov/ncas/alerts/aa20-352a`. Accessed: 2021-4-2.

[46] "Mitigate microsoft exchange server vulnerabilities." `https://us-cert.cisa.gov/ncas/alerts/aa21-062a`. Accessed: 2021-4-2.

[47] A. Bartoli, E. Medvet, and F. Onesti, "Evil twins and WPA2 enterprise: A coming security disaster?," *Comput. Secur.*, vol. 74, pp. 1–11, May 2018.

[48] W. H. Sanders, "Quantitative security metrics: Unattainable holy grail or a vital breakthrough within our reach?," *IEEE Security Privacy*, vol. 12, pp. 67–69, Mar. 2014.

[49] N. Nikiforakis, L. Invernizzi, A. Kapravelos, S. Van Acker, W. Joosen, C. Kruegel, F. Piessens, and G. Vigna, "You are what you include: large-scale evaluation of remote javascript inclusions," in *Proceedings of the 2012 ACM conference on Computer and communications security - CCS '12*, (New York, New York, USA), p. 736, ACM Press, 2012.

[50] A. Bartoli, A. De Lorenzo, E. Medvet, M. Faraguna, and F. Tarlao, "A security-oriented analysis of web inclusions in the italian public administration," *CYBERNETICS AND INFORMATION TECHNOLOGIES*, vol. 18, no. 4, 2018.

[51] M. Zimmermann, C.-A. Staicu, C. Tenny, and M. Pradel, "Small world with high risks: A study of security threats in the npm ecosystem," in *28th USENIX Security Symposium (USENIX Security 19)*, pp. 995–1010, 2019.

[52] E. S. Alashwali, P. Szalachowski, and A. Martin, "Exploring https security inconsistencies: A cross-regional perspective," *Computers & Security*, vol. 97, p. 101975, 2020.

[53] P. V. Mockapetris, *RFC1034: Domain names - concepts and facilities*. USA: RFC Editor, 1987.

[54] R. Elz, R. Bush, S. Bradner, and M. Patton, *RFC2182: Selection and Operation of Secondary DNS Servers*. USA: RFC Editor, 1997.

[55] S. Singanamalla, E. H. B. Jang, R. Anderson, T. Kohno, and K. Heimerl, "Accept the risk and continue: Measuring the long tail of government https adoption," in *Proceedings of the ACM Internet Measurement Conference*, IMC '20, (New York, NY, USA), pp. 577–597, Association for Computing Machinery, Oct. 2020.

[56] A. Bhardwaj, V. Mangat, R. Vig, S. Halder, and M. Conti, "Distributed denial of service attacks in cloud: State-of-the-art of scientific and commercial solutions," *Computer Science Review*, vol. 39, p. 100332, 2021.

[57] S. Behal, K. Kumar, and M. Sachdeva, "Characterizing ddos attacks and flash events: Review, research gaps and future directions," *Computer Science Review*, vol. 25, pp. 101–114, 2017.

[58] S. Behal and K. Kumar, "Detection of ddos attacks and flash events using information theory metrics–an empirical investigation," *Computer Communications*, vol. 103, pp. 18–28, 2017.

[59] G. Huston, "Measuring ROAs and ROV." `https://labs.apnic.net/?p=1420`. Accessed: 2021-3-26.

[60] M. Müller, T. Chung, A. Mislove, and R. van Rijswijk-Deij, "Rolling with confidence: Managing the complexity of dnssec operations," *IEEE Transactions on Network and Service Management*, vol. 16, pp. 1199–1211, 2019.

[61] M. Müller, M. Thomas, D. Wessels, W. Hardaker, T. Chung, W. Toorop, and R. V. Rijswijk-Deij, "Roll, roll, roll your root: A comprehensive analysis of the first ever dnssec root ksk rollover," *Proceedings of the Internet Measurement Conference*, 2019.

[62] Y.-D. Song, A. Mahanti, and S. C. Ravichandran, "Understanding evolution and adoption of top level domains and dnssec," *2019 IEEE International Symposium on Measurements & Networking (M&N)*, pp. 1–6, 2019.

[63] G. Huston, "DNSSEC validation revisited." `https://blog.apnic.net/2020/03/02/dnssec-validation-revisited/`, Mar. 2020. Accessed: 2021-6-22.

[64] G. Huston, "DNS trends," *The Internet Protocol Journal*, vol. 24, pp. 2–17, Mar. 2021.

[65] "Making .gov more secure by default." `https://home.dotgov.gov/management/preloading/dotgovhttps/`. Accessed: 2021-3-31.

[66] Google, "A safer default for navigation: HTTPS." `https://blog.chromium.org/2021/03/a-safer-default-for-navigation-https.html`. Accessed: 2021-4-2.