BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **18**, No 4
Sofia • 2018 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2018-0050

A Security-Oriented Analysis of Web Inclusions in the Italian Public Administration

A. Bartoli, A. De Lorenzo, E. Medvet, M. Faraguna, F. Tarlao

Dipartimento di Ingegneria e Architettura, University of Trieste, Italy
E-mails: bartoli.alberto@units.it mimmuz2k5@gmail.com emedvet@units.it
faragunamarco@gmail.com ftarlao@gmail.com

***Abstract*:** *Modern web sites serve content that browsers fetch automatically from a number of different web servers that may be placed anywhere in the world. Such content is essential for defining the appearance and behavior of a web site and is thus a potential target for attacks. Many public administrations offer services on the web, thus we have entered a world in which web sites of public interest are continuously and systematically depending on web servers that may be located anywhere in the world and are potentially under control of other governments. In this work we focus on these issues by investigating the content included by almost 10000 web sites of the Italian Public Administration. We analyse the nature of such content, its quantity, its geographical location, the amount of dynamic variations over time. Our analyses demonstrate that the perimeter of trust of the Italian Public Administration collectively includes countries that are well beyond the control of the Italian government and provides several insights useful for implementing a centralized monitoring service aimed at detecting anomalies.*

***Keywords*:** *Web security, e-Government, Javascript.*

1. Introduction

Modern web sites serve content that browsers fetch automatically from a number of different web servers, usually associated with organizations different from the one associated with the web site. Such web servers may be placed anywhere in the world and end users do not see any explicit notification of those *third-party* interactions. Most of third-party interactions are made for purposes of user tracking and advertising [1], but these interactions are also often used for fetching code to be executed within the browser in the form of JavaScript libraries. These libraries provide common functionalities and simplify the development of complex web applications.

Web servers that host third-party content are an attractive target for attacks, in particular, for attacks aimed at modifying JavaScript libraries [1-3]: First, a single successful attack may impact thousands of web sites; second, a successful attack provides almost complete control on the browsers that visit the impacted web sites;

third, timely detection of such attacks is very difficult. From a different point of view, a web site that includes external libraries may lose control of the content actually delivered to *its* clients as a result of intrusions to *other* sites.

Attacks of this sort are not a theoretical possibility and have already started appearing. In particular, in February 2018 a cryptojacking malware was injected into a script used by more than 4000 web sites, including the UK's Information Commissioner Office, US courts, the UK's financial ombudsman, several Australian government websites [4, 5]. Similar incidents have been observed on other government web sites of several countries [6-8]. The behavior of government web sites was thus altered in a way that was hidden to both their users and their administrators. The motivations behind that attack were financial and the attack probably was carried out by a criminal organization. The underlying issue is much deeper and has much broader ramifications, though: what if the attack had been executed by a state actor? What if the objective of the attack had been introducing subtle alterations in the site behavior in order to undermine trust in the government organization behind the site? What if a coordinated and large scale attack occurred on many such sites?

In other words, the security incidents mentioned above have made it manifest that the integrity of content and behavior of government web sites may depend on the integrity of *many other sites* belonging to administration domains fully *disjoint* from those of the government. The fact that third-party content may be located *anywhere* in the world makes these issues especially critical: the government of the territory in which third-party content happens to be located could force specific changes to that content, perhaps *tailored* to specific web sites or countries that use that content. We have thus entered a world in which government web sites are *continuously* and *systematically* depending on web sites that are beyond control of the government and that may even be under control of other governments. Indeed, a recent report by the NATO Cooperative Cyber-Defence Center of Excellence points out that "*one of the least explored areas of cyber vulnerabilities concerns cross-border dependencies of critical information infrastructure*" and that "*cross-border dependencies create additional vulnerabilities and a potential source of instability even for countries that have addressed these issues domestically*" [9]. On the other hand, the legal frameworks necessary for providing concrete end operational guidelines suitable for these novel forms of dependence are often still excessively vague [10].

It is crucial to emphasize that the tight dependence induced by the inclusions of web resources is intrinsically different from the one associated with the supply chain of hardware and software components [11], because changes in third-party content included by web sites can be executed much more quickly and with much less planning. Furthermore, subtle modifications in third-party content can be *targeted*, applied and *undone* at will, unlike modifications in hardware and software components.

In this work we focus on these issues and investigate the potential risks associated with government web sites, based on a large scale analysis of almost 10000 web sites of the Italian Public Administration. Specifically, we investigate the *nature of the content* obtained from other sites, the *number of those sites* and their

geographical location, i.e., whether they are abroad and in which countries. The nature of the content obtained from other sites determines the kind of attacks that can be executed. Scripts is the potentially more dangerous content category: Depending on how the script is used by the including web page, a script may change the appearance of the site using that script completely. Images can be misused, essentially, only for attempting to inject malware in the browser, by exploiting possible browser vulnerabilities related to image rendering. The number of sites that are automatically contacted by the browser for assembling the content of a given site indicates the number of different targets potentially available for attacking that site. Finally, the countries where those targets are located highlight the implicit dependencies for ensuring integrity of content and behavior of the main site. For example, the fact that a site of the Italian Public Administration serves third-party content from web servers located in Russia, and Israel is certainly a relevant information from a strategic point of view.

We also investigate the *temporal changes* in third-party content, with the aim of assessing the feasibility of a monitoring service capable of detecting any *anomaly* in the content served by a web site automatically. Defensive services of this kind can be implemented in the cloud in the form of Software as a Service [12] and have been proposed for providing a systematic, continuous monitoring of the visible contents of a web site, i.e., as a defense toward *defacements* [13-17]. The framework consists in observing the visible content of a web site, first for constructing a profile capturing the dynamic variability of the site, and then for notifying the site administrator of any observed anomaly with respect to the site-specific profile. Such services are potentially very useful because they can monitor a large set of web sites automatically, without any need of installing dedicated software in the sites to be monitored or of providing any site-specific service configuration. Our investigation allows gaining insights into the possibility of implementing similar frameworks for monitoring the *integrity of third-party resources*, i.e., of components which determine the site *behavior* beyond its visible appearance. A key problem along this path consists in assessing the number and the frequency of changes for third-party resources. If most web sites change most of their third-party resources every few days, then an automatic and site-agnostic monitoring service may be very hard to implement in practice. On the other hand, if changes in third-party resources do not occur frequently, then notifying site administrators in order to verify the genuinity of the change could be feasible. Extensions to the HTTP protocol have been defined which allow the administrator of a web site to specify the approved sources and content for resources to be included in that site, in the form of *content security policies* contained in HTTP responses [18]. When connecting to a web site, a browser will apply any content security policy previously received from that site and will thus use only resources explicitly whitelisted (approved) by the administrator. Recent studies have demonstrated that very few sites actually use content security policies [19-21], though, which makes the need for an external and zero-configuration monitoring service even more evident.

As an aside, our interest in this research was triggered by the fact that a malicious script located in Singapore was found on the site of the Italian Agency for

the Evaluation of Universities. That script remained unnoticed for an unknown interval until it was spotted by a media outlet [22]. A monitoring service like the one we are devising would have been able to detect such a script extremely quickly, even without the need of solving the difficult problem of classifying the script as malicious – more precisely, site administrators would have been promptly notified of the presence in the site of an anomaly, in the form of a never seen before script located in Singapore.

Although we are not aware of any work with a focus similar to ours, the web security community has long recognized the risks intrinsic in the inclusion of resources from remote servers. A significant work in this area is [1], which analyzed a large-scale crawl of the top 10000 sites in the Alexa ranking worldwide and identified the trust relationships of these sites with their library providers. The feasibility of attempting to uncover malicious content by comparing two versions of the same site has been proven in [23], which also discovered previously unknown infection campaigns. The attack vectors made possible by JavaScript libraries included by many web sites have been analyzed in [24]. Techniques for whitelisting safe scripts automatically in web sites with dynamically changing content have been proposed in [25].

2. Data collection and methodology

We used a dataset consisting of 9846 web sites obtained from the Italian National Index of the Public Administration (**https://www.indicepa.gov.it**). Many of these sites satisfy the definition of "information infrastructures of national interest" provided by the Italian Government [9]. We have not quantified this figure exactly because the corresponding requirements are rather high level and cannot be extracted automatically.

We performed a static analysis based on one observation of the home page of each site and a dynamic analysis based on four observations of each home page taken approximately 10 days from each other. For each observation, we stored information for describing all the web resources that the browser automatically fetches for rendering that observation. For each resource, we stored what follows (we actually stored more information and list here only the pieces relevant to the presented analysis).

- **Resource type:** Obtained from the HTML element that provoked the automatic download of the resource, one of: `<audio>`, `<embed>`, `<iframe>`, `<img>`, `<link>`, `<object>`, `<script>`, `<track>`, `<video>`.
- **Final URL**: Web address from which the resource was downloaded (an HTTP request for a resource at a given URL may receive a *redirection* HTTP response, i.e., a response indicating that the resource should be requested at a different URL, specified in the response itself; in this case the browser sends an HTTP request for the new URL automatically; the new request may in its turn receive a further indirection response).
- **Resource hash:** A hash of the resource content. This information allows to efficiently detect whether the content of a given resource has changed in different

observations of the same home page, as well as whether the same resource is part of the content of different home pages.

- **IP server**: IP address of the server from which the resource was downloaded.

When analyzing resource types, we avoided to make any hypothesis on the attack vector, attack objective, nature of the attackers. We considered all resources as a potential risk and only emphasize that scripts constitute the potentially more dangerous content category. We provide some background on this fact below, for completeness of presentation.

Depending on how a script is used by including a web page, the script may alter the behavior of the web page completely (in the case of the cryptojacking incident previously mentioned, the script augmented the behavior of the web site with a cryptomining routine). Such alteration may include changing the appearance of the site using that script completely and altering the data exchanged between the browser and the site. In other scenarios a script may steal data obtained from the site, i.e., send those data at an attacker-controlled location. In all scenarios, a script may force the browser to exchange data with an attacker-controlled server: such *drive-by* attacks are aimed at injecting malware in the browser by exploiting browser vulnerabilities [26]. Drive-by attacks are very common and are the basis, for example, of the so-called *malvertising*, i.e., injection of malware in browsers by means of malicious advertisements [27-30]. Resources of type `iframe` are, essentially, equivalent to scripts used for drive-by attacks [31, 32]. The other resource types can be misused only for attempting to inject malware in the browser, by exploiting vulnerabilities in the browser code for handling the corresponding resources [33-36].

3. Results and discussion

3.1. Static analysis

In this section we report on data obtained by analysing only one observation of the content of the web sites. We defined the following *regions* whose names are self-explanatory: Italy, OutsideItaly, OutsideEU, OutsideED-US. Of course, OutsideItaly is a superset of OutsideEU and the latter is a superset of OutsideEU-US. We geolocated servers based on their IP address and associated each web site with two regions, as follows. The *real* region is the one where the server hosting the main page of the web site is located; the *virtual* region is determined by the location of the servers hosting the resources that are included by that web site:

- **Italy:** All the content is obtained from servers located in Italy.
- **OutsideItaly**: At least a resource is obtained from a server located outside of Italy.
- **OutsideEU**: At least a resource is obtained from a server located outside of the European Union.
- **OutsideEU-US**: At least a resource is obtained from a server located outside of the European Union and of the United States.

The composition of our dataset with respect to the geolocation of web sites is in Table 1 (columns real and virtual). It is interesting to note that 1687 web sites (17.1%) are located outside of Italy, with 319 web sites (3.9%) outside of the EU. It is even

more interesting to note that the virtual location is OutsideItaly for 8007 web sites (81.3%): the vast majority of sites of the Italian Public Administration indeed rely on resources that are located outside of Italy. Furthermore, 78.8% of the sites have a virtual location in the OutsideEU region and 10.2% in the OutsideEU-US region. Interestingly, by looking at the raw data in more detail, we found that 1318 web sites (13.4%) do not include *any* resource from Italy.

For each web site, we counted the resources included by the site that are in the same virtual region as the site itself (e.g., for a web site in virtual region OutsideEU, we counted resources hosted in servers located outside the EU). The basic indexes of the corresponding distributions (average, median, 90th percentile, maximum) are given in Table 1; the All row shows the indexes for all the sites considered in our study. The numerical values of these indexes confirm that there is indeed a substantial dependence of sites in our dataset from content located outside of Italy.

Table 1. Dataset description. Statistic indexes are for resources in the same virtual region as the site

Region	Real	Virtual	Average	Median	90th percentile	Maximum
Italy	8159	1839	39	36	70	1607
OutsideItaly	1687	8007	15	4	52	266
OutsideEU	319	7759	6	3	11	263
OutsideEU-US	29	1007	2	0	1	61
All	9846	9846	58	53	96	1607

We looked at the first 10 sites for number of resources located outside of Italy and found that all those sites correspond to public schools, with only one exception corresponding to a local professional association (10th position in this list, with 172 resources located outside of Italy). We believe such large values (more than twice the 90th percentile) should be considered as a form of anomaly to be notified to site administrators in order to verify the legitimacy of so many inclusions.

The previous data determined the virtual region of web sites by considering all resources are being equivalent. Next, we considered the *type* of the resource because this property determines the potential risk associated with the resource. The results are in Table 1. For example, the first row indicates that 66% of the web sites requires downloading a script from outside of Italy. The most important finding is that reliance of scripts located outside of Italy is indeed a crucial phenomenon in our dataset.

By looking at the figures for the three categories, it can be seen that most of the scripts located outside of Italy are located in the US. This fact may be explained by the wide diffusion of scripts for tracking visitors to web sites that are generally provided by US-based companies. It is important to remark that only 0.42% of web sites rely on scripts located outside of the European Union and of the United States. From a different point of view, scripts located outside of the EU and the US may be seen as anomalies and a monitoring service could signal anomalies of this kind to the site administrators. Similar remarks can be made for resources of the other types, with the observation that site administrators should prioritize the analysis of scripts as they are the most dangerous resource type. Furthermore, even the mere presence of either audio or video content may be seen as an anomaly irrespective of the location of the

content. Notifying site administrators of the presence of such content in order to verify whether it is indeed supposed to be present may be important: since certain video engines used by browsers are frequently affected by newly discovered vulnerabilities, a fraudulent insertion of a video component in a web site may be an effective way for distributing malware on visitors of that web site.

Table 2. Regions of resource types

Resource type	OutsideItaly	OutsideEU	OutsideEU-US
Script	66%	60%	0.42%
Link	52%	48%	1%
Object	0.48%	0.23%	0%
Img	36%	24%	0.46%
Iframe	21%	11%	0.081%
Form	13%	2.4%	0.21%
Embed	0.34%	0.21%	0.40%
Audio	0.081%	0.081%	0%
Video	0.35%	0.34%	0%

We then considered the countries from which the resources are obtained. Table 2 lists all those countries, along with the number of sites that use content located in each country. We believe that the size and composition of the set of those countries is rather surprising, as one would have hardly expected that sites in the Italian Public Administration depend on such a large and broad set of countries.

Furthermore, the number of sites that obtain at least one resource from Italy is significantly smaller than our dataset (8528 vs. 9846): It follows than more than 1300 sites do not obtain *any* resource from Italy. We inspected a sample of those sites and observed that they mostly obtain resources from France, Ireland and the US.

It is also interesting to observe the presence of countries which do not belong to the NATO alliance and that one would not expect to be routinely serving content for sites of the Italian Public Administration (e.g., Japan, Israel, Russia, China, India, Ukraine, Seychelles Islands). Indeed, since those countries are not used by many sites, one could consider those countries as a form of anomaly to be notified to site administrators in order to verify whether content from those countries is indeed supposed to be present.

We inspected some of the sites using resources from these countries and we found that several of those sites redirected to online shops selling shoes and similar goods of fashion brands. For example, a certain URL (that we prefer to omit) is redirected to a web site that is apparently selling Hogan shoes at the URL **http://www.hoganscerpeoutlet.com/**. This URL is a form of typosquatting as the Italian word for “shoes” is “scarpe” while the URL contains “scerpe”. We did not analyze whether those sites are legitimate sites and whether they are an authorized reseller of the brand. We interpret this result as a staleness of data in the Italian Index of the Public Administration: the legitimate web site changed URL without reflecting the new URL in the index; the previous URL was somehow acquired by another

organization, either by buying the domain name of the previous URL or by means of a fraudulent intrusion in a server that is no longer maintained but is still active.

Table 3. Sites with content in the corresponding country

Country	Number of sites	Country	Number of sites
Italy	8528	Japan	7
United States	7687	Spain	7
Ireland	1092	Israel	6
Netherlands	1031	Turkey	4
Canada	1002	Russia	4
France	598	Belgium	3
Germany	532	track	3
United Kingdom	317	Luxembourg	3
Austria	79	Denmark	2
Switzerland	15	China	2
Croatia	13	San Marino	2
British Virgin Island	11	Australia	2
Seychelles	10	Estonia	2
Sweden	9	Slovenia	2
Bulgaria	9	India	2
Czechia	9	Ukraine	1

We visually inspected a few tens of sites using resources from countries that do not belong to the NATO alliance. Such sites used 120 resources from those countries, including 69 scripts. Approximately 40% of the inspected sites turned out to be online shops selling goods of differing nature, from shoes to football shirts. Interestingly, we found that the median number of resources in legitimate sites and non-legitimate sites was 4 and 128, respectively. These figures could also be used for tuning forms of anomaly detectors.

We then analyzed which scripts are obtained from different countries. We considered two scripts as being equal if they have the same hash and counted the number of different countries from which the same script is obtained. Table 3 contains the information corresponding to the 10 scripts obtained from the largest number of countries, e.g., the first row corresponds to a script obtained from 14 different countries and used by 1090 different web sites.

Figures in Table 3 illustrate that web sites tend to refer to widely differing locations even when they need to use the very same script content. This fact multiplies the number of potential targets of interest for an attacker, i.e., of locations that could be attacked for modifying a given script. From a different point of view, the fact that different sites in the Italian Public Administration may direct their visitors to as many as 14 different countries for obtaining the very same script is quite odd, at least in principle.

Table 4. Scripts obtained from the largest number of different countries

Number of countries	Number of sites
14	1090
14	1421
14	2259
12	1351
11	1005
11	43
10	143
10	143
9	652
9	34

The fact that a given script be obtained from different countries may be an unavoidable effect of modern web technology that is beyond the control of web site administrators. For example, a large number of web sites could refer a script with the same URL and that URL could be resolved to web servers located in different countries by load balancing policies implemented within the name-to-IP address mapping infrastructure (i.e., the DNS). We ascertained that this is indeed what happens for the script associated with the first row of Table 3, which corresponds to **https://fonts.googleapis.com**. On the other hand, the administrators of different web sites could even refer to different repositories of the same script library.

It is worth pointing out is that the median and 90th percentile of the number of different countries from which a given script is obtained are 2 and 4, respectively. Thus, the phenomenon of identical scripts obtained from different countries is very common, although the number of different locations from which the same script is obtained is generally small. These figures could be used by a centralized service monitoring the integrity of the full set of web sites in order to assess the risk associated with scripts obtained from many different countries (such as those in Table 3) and possibly coordinate the access to the same script from different web sites.

Table 5. Sites with content not legitimate uncovered with script distribution anomalies

Number of sites	Countries	Additional countries
3	Turkey, Seychelles, Estonia	
4	Turkey, Seychelles, Estonia	
5	Turkey, Seychelles, Sweden	US
8	Turkey, Seychelles	Italy, US

Another kind of anomaly analysis may be based on scripts that are obtained from few countries and such that those countries are used by few sites. The rationale for this analysis is that scripts obtained from few countries are unlikely to correspond to widely used libraries, while countries that are used by few sites are unlikely to host widely used computing infrastructures. In other words, such analysis could uncover attacks relying on “unusual” scripts in “unusual” countries. Based on these considerations, we could find relatively quickly 4 scripts used only by 19 sites whose

content is clearly not legitimate (see Table 5, one row for each script). Interestingly, the script in the last row was found in seven sites that have become shopping sites and one site of a major Italian University. The script in the 3rd row was located also in the US while the one in the 4th row was located also in the US and in Italy.

3.2. Dynamic analysis

In this section we report on data obtained by analysing the *variations* between two consecutive observations of the same web site. Let r be a resource of type t, let v_1 be an observation of a web site and let v_2 be the next observation of that web site. We say that there is a variation between v_1 and v_2 when r is in v_1 but not in v_2, or when r is in v_2 but not in v_1, or when r is in both observations with the same URL but different content. We counted the variations across all the consecutive observations available and summarized the results in Table 6, separately for each resource type. In order to place these figures in perspective, the table contains also the statistical indexes computed statically. For example, in 90% of the consecutive observations of the same web sites, there are at most 5 scripts that change; and, in 90% of the web sites, there are at most 38 scripts.

Table 6. Resources for each web site (static) and variations between consecutive observations (Δ)

Type	Average		Median		90th percentile		Maximum	
	Δ	Static	Δ	Static	Δ	Static	Δ	Static
Any	7	58	7	53	14	96	248	1607
Script	2	21	1	18	5	38	102	1597
Link	2	12	1	9	4	25	109	124
Object	0.03	0.05	0	0	0	0	10	10
Img	2	24	1	20	5	44	130	1041
Iframe	0.50	0.6	0	0	1	2	36	48
Form	0.90	1	1	1	2	2	12	13
Embed	0	0.02	0	0	0	0	1	7
Audio	0	0	0	0	0	0	0	1
Video	0	0	0	0	0	0	5	5

We believe these results are important because they clearly illustrate that changes in the considered resources are quite unlikely events. It follows that, for many sites, a monitoring infrastructure which considers any change in those resources as an anomaly to be investigated by site administrators could be practically feasible and, most importantly, very useful. For other sites, the changes that should be notified and investigated could be filtered based on a profile of the observed site constructed automatically in a preliminary learning phase. An infrastructure that monitors a large set of the Italian Public Administration could even correlate the observed changes across different sites and use the corresponding information appropriately, either in a centralized way or by forwarding the information to the site administrators.

The distribution of variations between consecutive observations exhibits a long tail: The 10 sites with the largest number of variations between consecutive

observations have between 205 and 730 variations. Although such values are not, by themselves, indicative of any security-related problem, we believe they are so large to be considered as anomalies that should be investigated by site administrators.

3.3. Analysis for selected categories

We repeated some of the previous analyses on three significant subsets of our dataset, one containing Municipalities, one containing Ministries, one containing Hospitals. We constructed these subsets with a keyword-based heuristics applied to URL and names of web sites. The size of the resulting subsets was 4282, 7 and 32, respectively. Although these sizes may be too small to derive general conclusions, the corresponding data provide interesting insights. The categorization of web sites based on geolocation of resources (Table 7) confirms that reliance on resources located outside of Italy is an essential feature of the analysed subsets.

Table 7. Statistics of resources included by selected web site categories

	Municipalities	Ministries	Hospitals	All
OutsideItaly				
Number of sites	79%	86%	75%	81%
Median	12	9	20	15
Maximum	215	30	75	266
OutsideEU				
Number of sites	76%	86%	69%	79%
Median	5	9	4	6
Maximum	105	30	14	263
OutsideEU-US				
Number of sites	3%	0%	9%	10%
Median	3	0	1	2
Maximum	49	0	1	61

Interestingly, none of the Ministries web sites in our subset includes resources located outside of the European Union and the United States. We looked at the list of countries in detail and we found that Hospitals, Ministries and Municipalities use resources from increasingly larger sets of countries: Hospitals use resources from Italy, US, The Netherlands, UK; Ministries use resources also from Canada, France, Germany; Municipalities use resources also from Ireland, Austria, Canada, Switzerland, Japan, Croatia, Bulgaria, Czechia, Slovenia, British Virgin Islands, Luxembourg, Spain, Russia. In other words, these data suggest that Hospitals and Ministries tend to depend on a set of countries much smaller and restricted to the EU than Municipalities.

The phenomenon of a same script obtained from many different countries, thereby augmenting the number of potential targets of interest for an attacker, is present also in the three categories Municipalities, Ministries and Hospitals, as illustrated in Table 8.

Variations between consecutive observations are summarized in Table 9, separately for each resource type. These data illustrate the benefits that may be obtained by tailoring the anomaly definition to the profile of each individual site, as it can be seen that the variations do depend on the site category. In all cases, though, the median number of variations is very small for each category, thereby confirming

that even the simple approach of notifying site administrators of each variation may be practical and effective for a large quantity of web sites.

Table 8. Scripts obtained from the largest number of different countries, for different categories of sites

Municipalities		Ministries		Hospitals	
Number of countries	Number of sites	Number of countries	Number of sites	Number of countries	Number of sites
14	525	14	1	14	5
14	208	8	1	14	7
14	368	8	1	14	3
12	188	8	1	12	1
11	124	8	1	9	1
11	1	7	1	8	2
10	33	7	1	8	3
10	70	6	1	8	2
9	38	6	1	8	1
8	15	5	1	8	1

Table 9. Variations between observations, for each web site category

Resource type	Municipalities		Ministries		Hospitals	
	Median	90th	Median	90th	Median	90th
Script	1	6	2	9	1	4
Link	1	3	3	10	1	3
Object	0	0	0	0	0	0
Img	1	5	3	7	0	5
Iframe	0	2	0	2	0	1
Form	0	2	1	2	0	2
Embed	0	0	0	0	0	0
Audio	0	0	0	0	0	0
Video	0	0	0	0	0	0

The distribution of variations between consecutive observations exhibits a long tail: The 10 sites with the largest number of variations between consecutive observations have between 205 and 730 variations. Although such values are not, by themselves, indicative of any security-related problem, we believe they are so large to be considered as anomalies that should be investigated by site administrators.

3.4. Applicability of the findings to other countries

A study of the web sites of the Public Administration of other countries is beyond the scope of this work. However, we repeated some of the previous analyses on a small

sample of web sites of the *Public Administration of the United Kingdom*, selected from **https://www.gov.uk/government/organisations**. Although these additional data are clearly not a substitute for a broader and more complete assessment, they suggest that the issues emerging from our study are indeed of general interest. In this respect, it may be useful to observe that a large-scale analysis of the defensive security mechanisms adopted by more than 22,000 web sites in 28 EU countries, could not find any significative difference between the "security score" of web sites in different countries [37].

Table 10. Statistics for a sample of UK web sites

Region	Real	Virtual	Average	Median	90th percentile	Maximum
UK	29	2	5	5	6	7
OutsideUK	21	48	19	8	45	102
OutsideEU	13	48	13	6	42	98
OutsideEU-US	0	6	1	1	1	1
All	50	2	37	38	58	105

Table 11. Variations between consecutive observations for resource type (UK)

Resource type	Average	Median	90th percentile	Maximum
Any	22	23	33	54
Script	11	11	16	26
Link	5	4	11	17
Object	0	0	0	0
Img	4	2	8	20
Iframe	0.50	0	2	5
Form	0.80	1	2	3
Embed	0	0	0	0
Audio	0	0	0	0
Video	0	0	0	0

Specifically, Table 10 summarizes the real and virtual regions of web sites and of the included content, with the same schema as in Table 1. These figures confirm the presence of sites of the Public Administration located outside of the country, as well as a substantial dependence from content (i.e., resources) also located outside of the country. Variations between consecutive observations are summarized in Table 11, with the same schema as in Table 6. Also in this case we observe that there is a moderate amount of variations between consecutive observations, albeit higher than for the Italian Public Administration. This fact would have to be taken into account for automatically constructing global and site-specific profiles for a monitoring service.

4. Concluding remarks

The main findings of our analysis may be summarized as follows:

- Web sites of the Italian Public Administration indeed depend implicitly on servers that are beyond their administrative control and, most importantly, that are located in other countries. Such web sites are thus continuously and systematically depending on servers that might be under hidden control of other governments.
- A centralized monitoring service may have several practically feasible strategies for notifying site administrators of anomalies worth investigating. Such strategies may be based on global profiles constructed on the full set of monitored sites, for example the 90th percentile of numerical quantities or the set of countries accessed by 90% of the sites, as well as on profiles automatically tailored to each single web site.
- The practical absence of certain resource types in our dataset, for example audio and video content, may allow more aggressive definitions of anomalies. Similar considerations could be made, for example, on the sets of countries involved in providing a web site content.
- Web sites of the Italian Public Administration tend to refer to widely differing locations even when they need to use the very same script content: visitors of different sites may be directed to as many as 14 different countries for obtaining the very same script. While this fact might be unavoidable, it also multiplies the number of potential targets of interest for an attacker and thus constitutes an issue that could be effectively monitored by a centralized service.

A realistic assessment of the effective risk for the Italian society as a whole cannot be obtained on the sole basis of our data and requires further work, in particular, with the analysis of more pages for each web site, with a more granular view of script usage by web sites, with a scrutiny of the interdependencies between the computing infrastructures of the various organizations [38] and with an assessment of the operational security practices followed by each organization (similarly, e.g., to [39]). We plan to broaden our investigation by attempting to uncover possible correlations between the nature of resource inclusions by web sites and the content of those sites, by looking for similarities that should emerge from unsupervised analysis of large collections of web content [40-43]. We believe that analyses of this kind may provide important insights for the implementation of a centralized monitoring service.

References

1. Nikiforakis, N., L. Invernizzi, A. Kapravelos, S. Van Acker, W. Joosen, C. Kruegel et al. You Are What You Include: Large-Scale Evaluation of Remote Javascript Inclusions. – In: Proc. of 2012 ACM Conference on Computer and Communications Security (CCS'12), New York, NY, USA, ACM Press, 2012, p. 736.
2. Uesugi, S. You Could've Submitted a Pull Request to Inject Arbitrary JS Code into Donald Trump's Site. – In: Medium [Internet]. Medium, 18 August 2016 [Cited 21 August 2018]. **https://medium.com/@chibicode/you-can-submit-a-pull-request-to-inject-arbitrary-js-code-into-donald-trumps-site-here-s-how-782aa6a17a56**

3. H u n t, T. The JavaScript Supply Chain Paradox: SRI, CSP and Trust in Third Party Libraries. – In: Troy Hunt Blog [Internet]. Troy Hunt, 12 February 2018 [Cited 21 August 2018]. **https://www.troyhunt.com/the-javascript-supply-chain-paradox-sri-csp-and-trust-in-third-party-libraries/**
4. Z h o u, N. Cryptojacking Attack Hits Australian Government Websites. – The Guardian, 12 February 2018. Accessed 21 August 2018. **http://www.theguardian.com/technology/2018/feb/12/cryptojacking-attack-hits-australian-government-websites**
5. L o m a s, N. Cryptojacking Attack Hits ~4,000 Websites, Including UK's Data Watchdog. TechCrunch. TechCrunch; 12 February 2018. Accessed 21 August 2018. **http://social.techcrunch.com/2018/02/12/ico-snafu/**
6. Russian Government Website Was Affected by a Malicious Cryptocurrency Mining Script. – In: Altcoin Today [Internet]. 12 Jun 2018 [Cited 24 August 2018]. **https://altcointoday.com/russian-government-website-was-affected-by-a-malicious-cryptocurrency-mining-script/**
7. US Government Site Was Hosting Ransomware. – In: Threatpost [Internet]. 1 September 2017 [Cited 24 August 2018]. **https://threatpost.com/us-government-site-removes-link-to-cerber-ransomware-downloader/127767/**
8. B a k e r, P. "Malicious Attack" on Government Site Hijacked Computers to Mine XMR – The Market Mogul. – In: The Market Mogul [Internet]. 16 March 2018 [Cited 24 August 2018]. **https://themarketmogul.com/crypto-jack-malicious-attack/**
9. L K k: T. Regulating Cross-Border Dependencies of Critical Information Infrastructure [Internet]. 2015. **https://ccdcoe.org/sites/default/files/multimedia/pdf/CII_dependencies_2015.pdf**
10. H a r a š t a, J. Legally Critical: Defining Critical Infrastructure in an Interconnected World. – Int. J. Crit. Infrastruct Prot., Vol. **21**, 2018, pp. 47-56.
11. W i n d e l b e r g, M. Objectives for Managing Cyber Supply Chain Risk. – Int. J. Crit Infrastruct Prot. Vol. **12**, 2016, pp. 4-11.
12. K u m a r, R. P., P. H. R a j, P. J e l c i a n a. Exploring Security Issues and Solutions in Cloud Computing Services – A Survey. – Cybernetics and Information Technologies, Vol. **17**, 2017, No 4, pp. 3-31. **http://www.cit.iit.bas.bg/CIT_2017/v-17-4/01_paper.pdf**
13. M a g g i, F., M. B a l d u z z i, R. F l o r e s, L. G u, V. C i a n c a g l i n i. Investigating Web Defacement Campaigns at Large. – In: Proc. of 2018 on Asia Conference on Computer and Communications Security. New York, NY, USA, ACM, 2018, pp. 443-456.
14. B o r g o l t e, K., C. K r u e g e l, G. V i g n a. Meerkat: Detecting Website Defacements through Image-Based Object Recognition. – USENIX Security Symposium. usenix.org, 2015, pp. 595-610.
15. B a r t o l i, A., G. D a v a n z o, E. M e d v e t. A Framework for Large-Scale Detection of Web Site Defacements. – ACM Trans. Internet Technol. New York, NY, USA, ACM, 2010, 10: 10:1–10:37.
16. B a r t o l i, A., G. D a v a n z o, E. M e d v e t. The Reaction Time to Web Site Defacements. – IEEE Internet Comput. ieeexplore.ieee.org, 2009, 13, pp. 52-58.
17. D a v a n z o, G., E. M e d v e t, A. B a r t o l i. Anomaly Detection Techniques for a Web Defacement Monitoring Service. – Expert Syst. Appl. Elsevier, 2011, 38, pp. 12521-12530.
18. Content Security Policy Level 3 [Internet]. [Cited 4 September 2018]. **https://www.w3.org/TR/CSP/**
19. W e i s s b a c h e r, M., T. L a u i n g e r, W. R o b e r t s o n. Why Is CSP Failing? Trends and Challenges in CSP Adoption. Research in Attacks, Intrusions and Defenses. – Springer International Publishing, 2014, pp. 212-233.
20. P a n, X., Y. C a o, S. L i u, Y. Z h o u, Y. C h e n, T. Z h o u. CSPAutoGen: Black-Box Enforcement of Content Security Policy Upon Real-World Websites. – In: Proc. of 2016 ACM SIGSAC Conference on Computer and Communications Security, New York, NY, USA, ACM, 2016, pp. 653-665.

21. W e i c h s e l b a u m, L., M. S p a g n u o l o, S. L e k i e s, A. J a n c. CSP is Dead, Long Live CSP! On the Insecurity of Whitelists and the Future of Content Security Policy. – In: Proc. of 2016 ACM SIGSAC Conference on Computer and Communications Security, ACM, 2016, pp. 1376-1387.
22. Vai sul sito ANVUR? Uno script maligno registra il tuo profilo e lo manda a Singapore. – In: ROARS [Internet]. 18 April 2017 [Cited 22 Aug 2018]. **https://www.roars.it/online/vai-sul-sito-anvur-uno-script-maligno-registra-il-tuo-profilo-e-lo-manda-a-singapore/**
23. B o r g o l t e, K., C. K r u e g e l, G. V i g n a. Delta: Automatic Identification of Unknown Web-Based Infection Campaigns. – In: Proc. of 2013 ACM SIGSAC Conference on Computer & Communications Security (CCS'13), New York, New York, USA, ACM Press, 2013, pp. 109-120.
24. L a u i n g e r, T., A. C h a a b a n e, S. A r s h a d, W. R o b e r t s o n, C. W i l s o n, E. K i r d a. Thou Shalt Not Depend on Me: Analysing the Use of Outdated Javascript Libraries on the Web. – In: Proc. of 24th Annual Network and Distributed System Security Symposium (NDSS'17) The Internet Society. pdfs.semanticscholar.org, 2017. **https://pdfs.semanticscholar.org/50b5/56396ebc887461015b48ce89c572424bcedf.pdf**
25. S o n i, P., E. B u d i a n t o, P. S a x e n a. The SICILIAN Defense: Signature-Based Whitelisting of Web JavaScript. – In: Proc. of 2nd ACM SIGSAC Conference on Computer and Communications Security, New York, NY, USA, ACM, 2015. pp. 1542-1557.
26. C o v a, M., C. K r u e g e l, G. V i g n a. Detection and Analysis of Drive-by-download Attacks and Malicious JavaScript Code. – In: Proc. of 19th International Conference on World Wide Web, New York, NY, USA, ACM, 2010, pp. 281-290.
27. L i, Z., K. Z h a n g, Y. X i e, F. Y u, X. W a n g. Knowing Your Enemy: Understanding and Detecting Malicious Web Advertising. – In: Proc. of 2012 ACM Conference on Computer and Communications Security (CCS'12), New York, NY, USA, ACM Press, 2012, p. 674.
28. V a a s, L. Massive Malvertising Attack Poisons 288 Sites. – In: Naked Security [Internet]. 12 April 2016 [Cited 4 Sep 2018]. **https://nakedsecurity.sophos.com/2016/04/12/massive-malvertising-attack-poisons-288-sites/**
29. G o o d i n, D. Home Routers under Attack in Ongoing Malvertisement Blitz. – In: Ars Technica [Internet]. 16 December 2016 [Cited 4 September 2018]. **https://arstechnica.com/information-technology/2016/12/home-routers-under-attack-in-ongoing-malvertisement-blitz/**
30. Microsoft Patches Zero Day Flaw Used in Two Massive Malvertising Campaigns. – In: Dark Reading [Internet] [Cited 4 September 2018]. **https://www.darkreading.com/attacks-breaches/microsoft-patches-zero-day-flaw-used-in-two-massive-malvertising-campaigns/d/d-id/1326908**
31. ThreatLabz, M. P i e r c y, A. S i n g h. China's NCGA Government Site Infected with Hidden Malicious Iframe | Zscaler Blog. – In: Zscaler [Internet] [Cited 24 August 2018]. **https://www.zscaler.com/blogs/research/chinas-ncga-government-site-infected-hidden-malicious-iframe**
32. Mavrommatis NPP, Monrose MARF. All Your Iframes Point to Us. – In: USENIX Security Symposium USENIX. usenix.org, 2008, pp. 1-16.
33. A r s h a d, S., S. A. M i r h e i d a r i, T. L a u i n g e r, B. C r i s p o, E. K i r d a, W. R o b e r t s o n. Large-Scale Analysis of Style Injection by Relative Path Overwrite. – In: Proc. of 2018 World Wide Web Conference. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2018, pp. 237-246.
34. H e i d e r i c h, M., M. N i e m i e t z, F. S c h u s t e r, T. H o l z, J. S c h w e n k. Scriptless Attacks: Stealing the Pie Without Touching the Sill. – In: Proc. of 2012 ACM Conference on Computer and Communications Security, New York, NY, USA, ACM, 2012, pp. 760-771.
35. H a s h i m, A. Microsoft Edge Vulnerability Could Allow for Email and Facebook Data Scraping. – In: Latest Hacking News [Internet]. 22 Jun 2018 [Cited 23 August 2018]. **https://latesthackingnews.com/2018/06/22/microsoft-edge-vulnerability-could-allow-for-email-and-facebook-data-scraping/**

36. C i m p a n u, C. Chrome Bug Lets Attackers Steal Web Secrets via Audio or Video HTML Tags. – In: BleepingComputer [Internet]. BleepingComputer.com; 15 August 2018 [Cited 23 August 2018].
https://www.bleepingcomputer.com/news/security/chrome-bug-lets-attackers-steal-web-secrets-via-audio-or-video-html-tags/
37. V a n G o e t h e m, T., P. C h e n, N. N i k i f o r a k i s, L. D e s m e t, W. J o o s e n. Large-Scale Security Analysis of the Web: Challenges and Findings. Trust and Trustworthy Computing. – Springer International Publishing, 2014, pp. 110-126.
38. D e N i c o l a, A., M. L. V i l l a n i, M. C. B r u g n o l i, G. D'A g o s t i n o. A Methodology for Modeling and Measuring Interdependencies of Information and Communications Systems Used for Public Administration and e-Government Services. – Int. J. Crit. Infrastruct. Prot., Vol. **14**, 2016, pp. 18-27.
39. K i r i l o v, R. Effectiveness of Information Security in the Banks. – Cybernetics and Information Technologies, Vol. **6**, 2006, No 2, pp. 70-85.
http://www.cit.iit.bas.bg/CIT_06/v6-2/70-85.pdf
40. M e d v e t, E., A. B a r t o l i, G. D a v a n z o, A. D. L o r e n z o. Automatic Face Annotation in News Images by Mining the Web. – In: 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 2011, pp. 47-54.
41. M e d v e t, E., A. B a r t o l i, G. P i c c i n i n. Publication Venue Recommendation Based on Paper Abstract. – In: 2014 IEEE 26th International Conference on Tools with Artificial Intelligence, 2014, pp. 1004-1010.
42. M e g u e b l i, Y., M. K a c i m i, B.-L. D o a n, F. P o p i n e a u. Unsupervised Approach for Identifying Users' Political Orientations. Advances in Information Retrieval. – Springer International Publishing, 2014, pp. 507-512.
43. T r e m b l a y, M. C., C. P a r r a, A. C a s t e l l a n o s. Analyzing Corporate Social Responsibility Reports Using Unsupervised and Supervised Text Data Mining. New Horizons in Design Science: Broadening the Research Agenda. – Springer International Publishing, 2015, pp. 439-446.

Received 02.09.2018; Accepted 29.10.2018