How to build socio-organizational information from remote IP addresses to enrich security analysis?

Camille Moriot, Francois Lesueur, Nicolas Stouls, Fabrice Valois

How to build socio-organizational information from remote IP addresses to enrich security analysis?

Camille Moriot[1], François Lesueur[2], Nicolas Stouls[1], Fabrice Valois[1]

[1] Univ Lyon, INSA Lyon, Inria, CITI, EA3720, 69621 Villeurbanne, France
[2]IRISA, Université Bretagne Sud, 56017 Vannes, France

September, 2022

Abstract

There is a constant threat of having our computing systems under attack. Information regarding the origins of the traffic we receive can be valuable. Currently, the AS-number and the localization are the most commonly used IP-related information to characterize an attack.

In this paper, we propose expanding knowledge about a remote IP's owner to improve defensive reaction effectiveness and obtain in-depth analyzes of attacker profiles. We introduce the enrichment with socio-organizational information (such as organization type, activity field, etc.) about the entities owning the IP in addition to infrastructural information. This integration is driven by combining RDAP and Wikidata. We demonstrate that this proposal is promising.

Keywords— IP qualification, Social Labels, Wikidata, Security analysis

1 Introduction

IT systems are regularly rocked by attacks. In recent years, it has become an important means of pressure on competing companies, governmental organizations, and individuals. But unfortunately, attacks can bring down a system and all the social structures behind it.

Consequences can be, at the enterprise level, financial costs and failures of operational services. Attacks on public services also have serious consequences: numerous attacks took place against hospitals during the 2020 health crisis[1] and led to slow operations in the establishments. Given the consequences, and frequency of attacks[2,3] developing efficient solutions appears as a necessity.

[1]https://www.wsj.com/articles/hospitals-suffer-new-wave-of-hacking-attempts-11612261802?mod=tech_lead_pos13

[2]`https://www.forbes.com/sites/chuckbrooks/2022/01/21/cybersecurity-in-2022--a-fresh-look-at-some-very-alarming-stats/?sh=1d4aea3b6b61`

[3]`https://www.orange.com/en/newsroom/press-releases/2021/number-cyberattacks-against-organizations-increases-13-noticeable-rise`

In-depth analysis of attacks, including the attacker's infrastructure, improves counterattack mechanisms as they provide precious description and awareness. Unfortunately, attacks keep evolving, and attackers keep circumventing protection mechanisms, resulting in a need to develop or readjust defense mechanisms constantly. Thus, automated tools are needed to keep up with the evolution. We claim that new approaches are mandatory to deal with the complexity of attack strategies and the increased frequency of attacks. We propose to characterize misbehaving host sending attack traffic and accordingly characterize the attackers in terms of social organization and technical capabilities.

In this paper, we present a new and original methodology that increases the knowledge about IP addresses used by attackers by assigning organizational labels to these addresses. We introduce an algorithm collecting the following labels: the type of structure (*e.g.*, IT companies, universities), the field of work (*e.g.*, web service hosting) and the human size (*e.g.*, number of employees). Our approach aims to know the socio-organizational characteristics of the attacking host rather than the attack to establish the best security policies. We don't focus on the source of the attack but on the last host of an attack.

The following section provides a state of the art highlighting the context of our work, as well as some useful existing tools to analyze threats. Then, in section 3, we present a new and original methodology that increases the knowledge about the hosts' IP addresses by assigning socio-organizational labels. Finally, we provide promising results of our characterization algorithm in section 4.

2 State of the art

Defense mechanisms against attacks can be described by the following steps: detect, analyze and react. The analysis phase is crucial as it helps to develop countermeasures. Different aspects of an attack can be inspected, such as the attack mechanism or the origin of the attack. This paper focuses on analyzing the last host relaying the attack.

Why being focused on the source side, and how to track the last host? An in-depth study of the sources can be beneficial in two cases. First, it helps to identify the attack's origin and, if possible, to deploy solutions directly at the level of the last known host. Second, adding knowledge could differentiate legitimate traffic's sources from illegitimate ones during an attack. For instance, the attacker's infrastructure characterization is part of the widely-used Mitre ATT&CK matrix[10].

What intelligence can be gathered about the misbehaving hosts? How do we know if a source of traffic is legitimate? Threat intelligence (TI) can be broken down into several categories [11] such as strategic TI, operational TI, tactical TI and technical TI. Operational intelligence designates the information that can be gathered by devices and used to analyze attacks. Nowadays, information such as localization tags is used to increase knowledge about traffic's senders automatically. Such information can identify abnormal behaviors. Multiple localization databases such as Maxmind's Geolite[4] database or IP2Location database[5] can be used to obtain such information. Different geo-tracking databases have been compared in [5] showing that such geo-tracking

[4]https://dev.maxmind.com/geoip/geolite2-free-geolocation-data
[5]https://lite.ip2location.com/

tools tend to have the same result regarding the country of origin of an IP address. Crawlers can also be used to get information about hosts. Net crawlers such as Censys [2], Shodan[6] or Onyphe[7] yield knowledge regarding open ports and services behind IP addresses. Net crawlers use a pro-active approach, meaning that the knowledge is gained from active measurements. They also provide information about the accessibility of IP addresses from the Internet and can be used to detect vulnerabilities [12, 1]. Unfortunately, crawlers focus only on technical and factual information without proposing any interpretation of the entity behind.

As a result of threat intelligence, it is possible to obtain information about how trustworthy sources are and help to prevent attacks. Available blocklists[8] are a first step to knowing if a source of traffic can be trusted. In the case of an IP address belonging to such databases, it can be dropped or filtered. This information is enriched only after attempted attacks have taken place. Moreover, due to the fact that IP addresses can be shared, side effects can impact legitimate users [8]. To avoid these drawbacks, it is better to try to obtain more information about the sources before blocking its traffic. However, new blocklists using localization and political actualities [9] tend to have better results as they integrate context into the analysis.

Why is there a need to integrate social components in attack analysis and mitigation? All the previous tools mainly focus on the infrastructural part of the network, whereas they have social, political, or economics dimensions[4]. In some cases, attackers can abuse specific infrastructure, and there is a need to evaluate phenomena based on who owns the IP addresses. In [6], the authors observed a new phenomenon that consists of using residential IP addresses through proxies to bypass security rules that can be encountered in other types of networks. With this work, we can see the need to explore social and organizational network features. To identify residential traffic, they used Whois answers and DNS features for the corresponding network.

The information regarding the ownership of an address or a net block are held by the Regional Internet Registry (known as RIR) and can be retrieved from those RIR. The RDAP [7] protocol, progressively replacing Whois, allows to easily retrieve that information in JSON format. Observing the nature of the traffic as well as its organization also makes it possible to identify characteristics specific to attacks. For example, this allows in the context of certain attacks to determine whether certain structures can be abused to generate attacks. In [3], the authors conducted an evaluation of the presence of cloud IP addresses within different malicious activities. To do so, they studied IP addresses belonging to four of the biggest cloud service provider. They studied the apparition of those IP addresses in 39 blocklists. In this case, they concluded that cloud IPs could be blocklisted for another 30 days after participating in malicious activities. This result again proves the need for a tool characterizing IP addresses without blocking them.

All the previous work and tools demonstrate the need for efficient characterization of the traffic sources. This characterization needs to be not only technical but also focusing on social and organizational aspects as it was demonstrated that in many cases, attack phenomena could not be dissociated from a social

[6] `https://www.shodan.io/`
[7] https://www.onyphe.io/
[8] `https://www.iblocklist.com/`

context. Our work presents an original methodology characterizing misbehaving hosts in attacks based on organizational features.

3 Building organizational labels from IP addresses

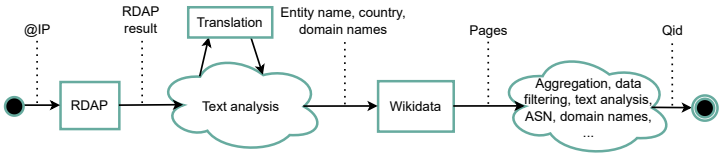

Figure 1: RDAP to Wikidata algorithm

As we introduced in the previous sections, countering attacks requires knowing the attackers in depth. Existing tools have been used and developed to produce technical knowledge regarding devices sending malicious traffic. We claim that social characterization must enrich technical analysis to integrate context information into our analysis.

In order to conduct in-depth analyses, we propose to qualify the sources by the following labels: **type of entity**, **domain of activity** and **human size**. We argue that studying the type of organization owning the IP address is relevant to identify if the threats only come from a certain kind of entity (*e.g.* enterprise, ISP, university network) and to characterize different groups of attackers. The entity's domain of activity is also relevant to trace back some threats linked to the activity, such as, for example, hosting devices.

Regarding the attribution of labels to an IP address, we introduce, in this section, an algorithm retrieving organizational information from Wikidata starting from partial data given by RDAP, illustrated in figure 1.

3.1 Using Wikidata to retrieve social information

Wikidata[9] is a collaborative platform owned by Wikimedia that serves as a support to Wikipedia. Wikidata is a database that contains items identified by Qids retrievable through an API using a SPARQL query. Each item is described by a list of labels. The database not only contains information about entities, but as an encyclopedia, general knowledge about various things. Especially, it contains the following labels relevant to our work: instance of the item (e.g. enterprise, business, telephone company, internet service provider, university), industry, employees, number of subscribers, autonomous system number and IPv4 routing prefix. Wikidata items may have an IPv4 routing prefix field. This can, when we try to qualify an IP address, allow us to easily identify the corresponding item. We notice that almost 25% of the IPv4 space is covered in Wikidata. In the rest of the cases, we need to find the name of the owning entity to perform a text search query to obtain the page corresponding to the entity sought.

[9] `https://www.wikidata.org/`

3.2 Obtaining the entity name from RDAP

In order to perform the Wikidata query, we need to know some information about the entity that owns the IP address: the entity name, country, their domain name and the AS-Number. To do this, we use RDAP, which contains the information provided by the owners to the RIRs. RDAP returns a JSON with multiple fields, and obviously, the field called *entity*, which aims to designate the owner of the IP address, is a very interesting one. Unfortunately, a lot of information contained in this field has been made anonymous in Europe according to the GDPR[10] or are identified by IDs. But, in most cases, there are many information regarding the entity hidden in different fields such as the *remarks* one for example. Information can be an email address with a domain named after a company or a postal address with the name of the entity. Thus, to extract and identify the correct information, we implement a **text analysis** algorithm identifying the most relevant sub-strings of a text based on their frequency. In this step, we also use a translator to normalize all RDAP entries in English. Note that our text analysis does not use ontology or deep learning. We consider the most frequent sub-strings containing the most frequent words in the RDAP entries. We do our comparisons by using the Jaro distance between words (composed of letters) or between fields (composed of words). At the end of this step, we also pinpoint potential discriminatory factors such as the country code, the domain name or an AS number. All this information also comes from RDAP.

3.3 Discriminating Wikidata's results

As represented on figure 1, after identifying key information representing the entity of the owner, we make a text research of the name via the Wikidata API to identify the corresponding resource. Note that this research returns all pages with the entity name within their title, therefore we need to analyze these results to identify the correct result associated with the entity. Many entities have names having several meanings or have changed their name during their existence. To this end, we inspect the following fields: *item label*, *also known as* and *subsidiaries*. The previous fields are used to refine our analysis as we noticed that RIR's database are not always up to date, especially for rebranded names. We **aggregate** results from the different queries and then define multiple criteria to **filter** the results. First of all, we identify Wikidata item with the maximum amount of words in common with the target text string. We also filter our results based on the expected fields we are searching in. Indeed, we noticed that entities have specific fields used to describe them, and if these fields are not present, we are less likely to identify the correct Qid. This particularly helps us to differentiate homographs: for instance, Orange is a french ISP, but it is also a color and the name of a city in the south of France. Finally, we use ASN, domain name, and country to discriminate the results. In Wikidata, there is an ASN field that may be specified for some entities. We also compare the domain address from the website field in Wikidata to the domains found within RDAP. The country may also help us discriminate results. Each of these factors is associated with a score, and the best ranked Wikidata item is then considered the one related to the entity. However, if the global score or the score in some

[10] `https://en.wikipedia.org/wiki/General_Data_Protection_Regulation`

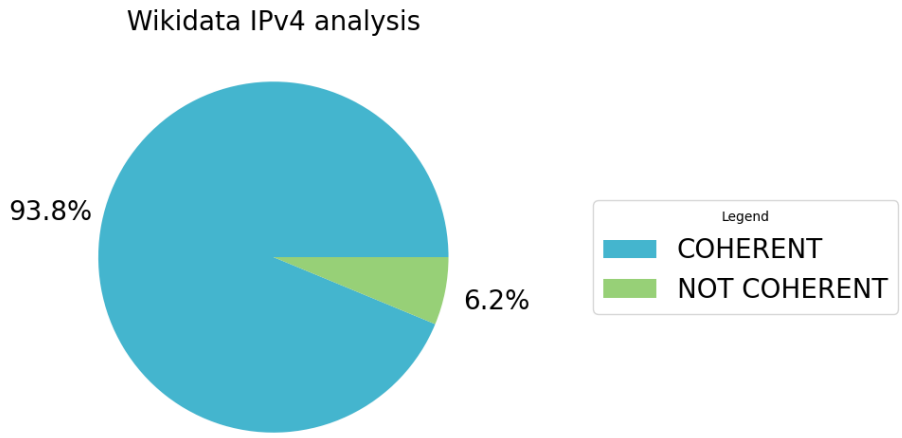

Figure 2: Wikidata comparison with RDAP results

criterion is too low, we consider that there is no result. These thresholds are defined experimentally.

4 Evaluation and discussion

In this section we evaluate the correctness of the results by evaluating the quality of the data contained in Wikidata and the results of our algorithm.

4.1 Evaluation of the data contained in Wikidata

As Wikidata is a collaboratively edited database, we want to ensure the quality of the information in it. Therefore, to verify the veracity of the IP information, we compare them to what we obtain with RDAP as a reference.

To this end, we extract all items in Wikidata having the *IPv4 routing prefix*. We remove all ranges assigned to RIR's items from this dataset. After this, we pick the first IP address of each range, and for each IP, we use RDAP to get the information stocked in the different RIR databases. From RDAP, we obtain a JSON with multiple fields. We inspect all the fields and try to match the label of the item (e.g., the name of the organization), as well as the alternative names within the fields. We obtained a match for 1208 out of 2212 ranges 2. Regarding the rest of the data, we evaluate by trying to find a partial match. After refining our criteria, we found another 697 matches. For the remaining data, we evaluated the results manually. As a result, we obtain a total 93,8% of coherent associations between RDAP and Wikidata. Our evaluation script is available online[11].

To conclude, the evaluation of the data contained in Wikidata validates the shortcut introduced in section 3.1 as we have excellent confidence in Wikidata.

[11] `https://gitlab.inria.fr/cmoriot/lcn_2022`

4.2 Evaluation of our characterization

In this subsection, we evaluate the capacity of our algorithm to find the correct Wikidata Qid from an RDAP result. To this end, we create a dataset containing the association of an IP range with the corresponding Wikidata Qid. This available[11] dataset is an extract of the previous dataset and contains 1055 entries belonging to part of 1208 for which we had a perfect match with RDAP. With this dataset, we evaluate the algorithm described in section 3.2 and 3.3. We use the first IP address of the range given by Wikidata as input for RDAP and the expected correct Qid from the Wikidata block (see fig. 1).

In 53% of the cases, we directly obtain a correct result. In another 7% of the cases, the correct result is in a shortlist of three entries which could be shown to an analyst to help him, and in 2% of the cases, the resulting page was the parent company of the one we were looking for (which is still quite good). Finally, in 20% of cases, no results are returned as the confidence score was too low. More than 80% of our results are thus sound (either good, quite good, or explicitly unfound). There are currently a bit less than 20% of wrong characterizations. A large part of these errors come from university networks, which are often managed by a third party that we could identify from the data contained in RDAP but not the university itself.

5 Conclusions

In this paper, we claim that a new approach is needed to cope with the increase of attacks and their complexity. We highlight the necessity to investigate in-depth all the information related to the source or the relay of an attack. We introduce a new and original methodology to enrich security investigations, by characterizing infrastructures and owners of IP addresses thanks to organizational labels. We develop an algorithm assigning social labels to IP addresses. In particular, we want to obtain labels describing an organization owning an IP address by the type of organization, its domain of activity and its human size. Our results are promising and from our evaluation, we underline the following key points. First, the information contained in RDAP is not necessarily up-to-date because of the dynamic of companies (*e.g.*, name change or merge/acquisition). This leads to difficulties to identify correctly an entity. However, thanks to the data available in different fields of the databases we use, it is often possible to identify and to characterize the correct entity. Second, the European anonymization of entity information (GDPR), which is also applied to RDAP, increases the complexity to correctly identify the owner of the IP range.

Since our algorithm affecting social label to an entity owning IP addresses is based on public data, RDAP entry and Wikidata, while our code is open source[12], our results are transparent and reproducible.

To extend our approach, we identify two complementary directions for our next work. First, we will enhance our contribution using a semantic approach or a deep reinforcement learning one. Second, because of the limitations of RDAP, we will propose to collect organizational labels using web crawlers or network topology.

[12] `https://gitlab.inria.fr/cmoriot/ipseen`

References

[1] Roland Bodenheim et al. “Evaluation of the ability of the Shodan search engine to identify Internet-facing industrial control devices”. In: *Int. J. Crit. Infrastructure Prot.* 7 (2014), pp. 114–123.

[2] Zakir Durumeric et al. “A Search Engine Backed by Internet-Wide Scanning”. In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015*, pp. 542–553.

[3] Naoki Fukushi et al. “A Large-scale Analysis of Cloud Service Abuse”. In: *8th IEEE Conference on Communications and Network Security, CNS 2020*. 2020, pp. 1–9.

[4] Robin A. Gandhi et al. “Dimensions of Cyber-Attacks: Cultural, Social, Economic, and Political”. In: *IEEE Technol. Soc. Mag.* 30 (2011), pp. 28–38.

[5] Ioana Livadariu et al. “On the Accuracy of Country-Level IP Geolocation”. In: *ANRW ’20: Applied Networking Research Workshop, 2020*. 2020, pp. 67–73.

[6] Xianghang Mi et al. “Resident Evil: Understanding Residential IP Proxy as a Dark Service”. In: *2019 IEEE Symposium on Security and Privacy, SP 2019*, pp. 1185–1201.

[7] Bruce J. Nikkel. “Registration Data Access Protocol (RDAP) for digital forensic investigators”. In: *Digit. Investig.* 22 (2017), pp. 133–141.

[8] Sivaramakrishnan Ramanathan et al. “Quantifying the Impact of Blocklisting in the Age of Address Reuse”. In: *IMC ’20: ACM Internet Measurement Conference, 2020*.

[9] Henanksha Sainani et al. “IP Reputation Scoring with Geo-Contextual Feature Augmentation”. In: *ACM Trans. Manag. Inf. Syst.* 11 (2020), 26:1–26:29.

[10] Blake E Strom et al. “MITRE ATT&CK®: Design and Philosophy”. In: (2020).

[11] Wiem Tounsi and Helmi Rais. “A survey on technical threat intelligence in the age of sophisticated cyber attacks”. In: *Comput. Secur.* 72 (2018), pp. 212–233.

[12] Natalija Vlajic and Daiwei Zhou. “IoT as a Land of Opportunity for DDoS Hackers”. In: *Computer* 51.7 (2018), pp. 26–34.