We have to go back: A Historic IP Attribution Service for Network Measurement

Florian Streibelt[1], Martina Lindorfer[2], Seda Gürses[3], Carlos H. Gañán[3], and Tobias Fiebig[1]

[1] Max Planck Institute for Informatics {fstreibelt,tfiebig}@mpi-inf.mpg.de
[2] TU Wien martina.lindorfer@tuwien.ac.at
[3] TU Delft {f.s.gurses,c.hernandezganan}@tudelft.nl

Abstract. Researchers and practitioners often face the issue of having to attribute an IP address to an organization. For *current* data this is comparably easy, using services like whois or other databases. Similarly, for historic data, several entities like the RIPE NCC provide websites that provide access to historic records. For large-scale network measurement work, though, researchers often have to attribute millions of addresses. For *current* data, Team Cymru provides a bulk whois service which allows bulk address attribution. However, at the time of writing, there is no service available that allows *historic* bulk attribution of IP addresses. Hence, in this paper, we introduce and evaluate our 'Back-to-the-Future whois' service, allowing historic bulk attribution of IP addresses on a daily granularity based on CAIDA Routeviews aggregates. We provide this service to the community for free, and also share our implementation so researchers can run instances themselves.

1 Introduction

A common issue in the network measurement domain–but also in industry fields from Threat Intelligence to traffic engineering–is attributing an IPv4 or IPv6 address to an organization. While, technically, Regional-Internet-Registries (RIRs) allocate IP addresses to organizations [15], and provide a whois [7] infrastructure to make this information accessible, common whois interfaces are impractical for bulk requests. This is mostly due to whois providing unstructured text data, which has to be appropriately parsed [33]. Furthermore, organizations may have multiple organizational objects with overlapping and semantically equivalent data, which is not bit-equivalent or hides relationships due to subsidiaries from, e.g., different countries [4]. To address the needs of, especially, the threat hunting community, Team Cymru operates a bulk whois service, which allows users to bulk-request AS attribution for thousands of requests.

However, when working with *historic* data-sets, sometimes ranging back decades, *current* whois information may be ill suited to correctly attribute IP addresses, especially in the wake of IPv4 exhaustion [26] and the accelerating IPv4 market [21, 10, 24, 20]. Hence, in this paper, we introduce our historic whois

service–Back-to-the-Future whois–which we implemented to address these challenges, leveraging the public CAIDA Routeviews aggregates [28, 3]. Our service is publicly available to the community at bttf-whois.as59645.net port tcp/10000. The service provides a historic address attribution service starting in May 2005 and for IPv4 and in January 2007 for IPv6. It can be queried using a simple syntax, and provides structured JSON output.

In summary, we make the following contributions in this paper:

- We introduce 'Back-to-the-Future whois' (BTTF whois) as a public service for the networking research community providing a simple way to obtain historic IP address attribution. Our service can handle around 1,000 request per second per connection, with 30 instances running, enabling around 30,000 requests per second.
- We document the methodology we used for this service, so researchers can independently utilize it to distil historic IP attribution from Routeviews or the CAIDA aggregates.
- We evaluate the efficacy of BTTF whois in terms of coverage over time using an example research project, and find BTTF to perform comparably to Team Cymru's bulk whois service on recent data, while outperforming it in accuracy for historic data.

The remainder of this paper is structured as follows: First, we introduce the datasets we use and our methodology for BTTF whois in Section 2. Next, we evaluate BTTF whois against Team Cumry's bulk whois in a sample case. Finally, we first discuss our results and limitations in Section 4, before concluding in Section 5.

2 Dataset and Methodology

2.1 Utilized Data

CAIDA Data for BTTF Whois The historic whois service leverages the aggregates of the RouteViews project compiled daily by CAIDA [3]. The dataset spans the time from May 2005 for IPv4 until today, and the time from January 2007 until today for IPv6, both with a daily resolution. We decided to utilize the aggregates computed by CAIDA instead of aggregating the routing tables provided by the RouteViews project [28] ourselves, as the RouteViews dataset is large (tenth of TB as compressed files), and aggregation of this data is already a significant task in itself.

This prefix data alone is, however, insufficient to estimate a whois service based on routing data. Routing data in itself only maps IP addresses to ASes that announced the prefix at a specific time. Especially when looking at historic data, ASes may change the organization they are allocated to. Furthermore, we may find ASes that announce prefixes which are not registered to the announcing AS' organization, see for example Cogent announcing various customer prefixes,[4] see also Section 4.2.

[4] `https://bgp.tools/as/174#prefixes`

We address the issue of tying ASes to organizations by leveraging the AS2ORG dataset, also published by CAIDA [11, 4]. The AS2ORG dataset covers the period from April 2004 up until today, with a quarterly resolution. However, this reduced resolution will lead to a reduced reliability of the AS2ORG mappings, meaning that changes of ownership/authority over an AS may be reflected up to three months too late, while temporary changes of a duration less than three months may remain completely unnoticed, see Section 4.2.

IP Address Research Data To evaluate BTTF whois, we have to compare its efficacy against a 'current' whois extract on a historic IP address dataset, where we can also investigate the impact of BTTF whois on the analysis results. For this purpose, we selected the University IP Address dataset collected from the Farsight SIE DNS Dataset [8] by Fiebig et al. in their study of cloudification in universities [9]. This dataset consists of A, AAAA, and CNAME records attributed to universities' domains between January 2015 and May 2022. It spans a total of 133M records, ranging between 600k and 6M individual IP addresses per month, with records within each month being unique, while several months may contain the same addresses. Fiebig et al. used this dataset to identify which universities utilize services located in one of the three major cloud providers–Amazon, Google, and Microsoft–by identifying which universities have DNS entries pointing to IP addresses belonging to these cloud providers. In this process, CNAMEs are resolved to the IP addresses they ultimately point to.

Team Cymru Whois Data As a base-line, we requested bulk whois data from Team Cymru's bulk whois service for all unique addresses in June 2022. We used the Team Cymru whois to resolve all 14M unique IP addresses in the university dataset. For each IP address the bulk whois service of Team Cumry returns the currently associated AS number, the requested address, and the AS Name and location of the corresponding AS.

2.2 Methodology

In this section, we describe how we organized the CAIDA AS2ORG and AS2Prefix datasets in our service daemon to enable quick queries for individual addresses against the dataset. The major challenge–preventing a traditional RDBMS from being used–is that these datasets contain whole prefixes, instead of individual IP addresses, and relations between objects are complex. This would lead to, for example in SQL, a nested JOIN structure which limits performance of an RDBMS. To prevent this bottleneck, our implementation uses a completely in-memory prefix trie, i.e., pytricia [1].

AS2ORG Data-Structure. To use the supplied dataset to identify the AS and organization announcing a specific IP address, we first create a data-structure mapping time-frames, organizations, and ASes to each other. The challenge here is that the resolution of the supplied data is relatively low. Furthermore, we find that the supplied data regularly contains parsing errors, as it has been sourced

from RIR supplied whois data, which is known to be often unstructured and to have volatile formats [19].

To handle the sparseness of the supplied data, we do have to make decisions on the margin of error that is acceptable for a whois service when making an educated guess for the organizational affiliation of an AS in between two quarterly files. There, we have to handle four cases:

- **AS2ORG unchanged:** If, in both files, the AS is mapped to the same organization, we assume that it was continuously mapped to the same organization between the two dates for which we have data.
- **AS missing from newer file (AS removed):** If an AS has been removed, we consider it to be removed from the day directly following the last quarterly file's date in which the AS could be found.
- **AS missing from older file (AS added):** If an AS has been added, we consider this AS mapping to be valid from the date of the file in which the AS first occurs (again).
- **AS2ORG changed:** If the AS2ORG mapping changes between two adjacent files, we consider this change to have come into effect on the day after the older files' collection date.

Following this approach, we can then construct a continuous mapping of ASes to organizations in our data-structure.

Prefix Tree (Trie). Next, we iterate through the list of available files by date, and add the prefixes we find to an IP trie [1]. In that trie, each added prefix holds a list at date ranges when it was observed. For each prefix in our input files, we check if the prefix exists in the trie. Here, we have to handle four cases:

- **Prefix is not in the trie:** We add the prefix to the trie, setting the first 'first seen' field to the date of the collection date of the currently processing file.
- **Prefix is in the trie:**
  - **No gap to last-seen date:** If the last-seen date of the prefix is the date of the day before the collection time of the currently processing file, we update the last-seen date of the most recent date-range to the date of the currently processing file.
  - **Gap to last-seen date:** If the last-seen date of the prefix is not the date of the day before the collection time of the currently processing file, we add a new date-range to the list of date-ranges, and set the first seen date to the date of the currently processing file.
  - **Originating AS changed:** If the originating AS(es; see below) changed from the last seen state, we treat the prefix as a new prefix, i.e., start a new date range associated with the new ASes.

In all cases, the prefix is attributed to the ASes we observe as announcing the prefix. There, we also have to handle several special cases:

- **Prefix originated by exactly one AS:** If a prefix is originated by exactly one AS, we add this AS as the authoritative AS.

– **MOAS prefix:** If a prefix is announced by multiple ASes at the same time, commonly known as a MOAS (Multi Origin AS) prefix, we add all these ASes to the announcement state, see the section on handling requests for details on the presentation.
– **ASSET aggregate:** Under certain conditions ASes may aggregate prefixes received from downstream ASes. Fore example, if AS65536 announces 198.51.100.0/25 to AS65538, and AS65537 announces 198.51.100.128/25 to AS65538, AS65538 can aggregate these announcements to 198.51.100.0/24, only announcing that to its peers, while also aggregating AS65536 and AS65537 to { AS65536, AS65537 } in the AS path of that announcement. The information whether 198.51.100.0/25 was originated by AS65536 or AS65537 is lost in this process. As this is suggested to occur only on provider aggregatable IP space [6], we attribute the whole /24 to the aggregating AS, i.e., AS65538 in this case.

After having determined the ASes to which we attribute a prefix, we look up the associated AS2ORG mapping from our first datastructure and add that information to the date range. Please note that the trie data structure handles the occurrence of more specific prefixes by a branching approach, i.e., we can add 198.51.100.128/25 to the trie, even if 198.51.100.0/24 is already present. When looking up addresses, the more specific will match, and we will have to traverse the tree upward, see also below under 'Lookups'. Loading the full data set into the implementation takes around 24 hours.

Filtering. Prefix announcements on the Internet are noisy. Specifically, we may regularly observe organizations announcing prefixes they are not supposed to announce [31], announce prefixes that are more specific than the maximum agreed prefix size in the global routing table (/24 for IPv4 and /48 for IPv6) [30], announce prefixes that are unreasonably short, e.g., when leaking default routes, or announce prefixes and AS numbers from reserved ranges [25] (see also IANA's registires[5,6]). Reserved prefixes are statically added to our lookup daemon, and reported as such upon lookup. Hence, when importing prefixes we are filtering all announcements less specific than a /8 for IPv6 and /18 for IPv6, and more specific than a /24 for IPv4 and /48 for IPv6. Similarly, we exclude all prefixes originated by private and reserved AS numbers, i.e., 0 [17, 18], 23456 [32], 64496-64511 [16], 64512-65534 [13, 23], 65535 [12], 65536-65551 [16, 32], 65552-131071 (IANA Reserved), 4200000000-4294967294 [23], and 4294967295 [12]. Finally, we exclude broken data, as for example, AS numbers that include a dot.

Lookups. The implementation of the historic whois service allows lookups with daily granularity. When an IP address or prefix is looked up, we first identify the most specific match. Next, we check if the prefix has been announced at the given date, i.e., if it has a date-range covering the requested date. If it does not

[5] `https://www.iana.org/assignments/iana-ipv4-special-registry/iana-ipv4-special-registry.xhtml`,

[6] `https://www.iana.org/assignments/iana-ipv6-special-registry/iana-ipv6-special-registry.xhtml`

have a corresponding date range, we traverse the tree until we either find a less specific prefix with a covering date-range or arrive at the root of the address tree. If we arrive at the root, we return that the prefix could not be found at the given date.

For the most specific prefix with a covering date range, we return the requested IP address or prefix, the requested date, and the result set. The result set contains the dates when the prefix was first and last observed for the date-range covering the requested date, with the last-seen date being null if the prefix was still being observed in the newest file imported into the daemon. Additionally we return the identified prefix and the list of ASes associated with the prefix. For each AS we also return an AS2ORG mapping, listing the ASN, the ASNAME and RIR where the ASN has been registered. Furthermore, we return all organizations associated with the AS at the time of the request, which includes the country code registered for the organization, the RIR the organization object has been obtained from, and the name of the organization.

Implementation, Infrastructure, and Performance. We implemented the historic whois system in a team using roughly three person months between May and August 2022 in Python. To handle our request load, we deployed forty instances behind a load-balancing frontend on a cluster of four hardware machines. Each instance consumes roughly 16GB of memory (including caches) and has access to two dedicated CPU threads, leading to a total resource consumption of 80 CPU cores and 640GB of memory, without Kernel Same-Page Merging (KSM) applied. An instance can process around 1.2K lookups a second, allowing us to perform the address resolution for the 133M addresses over 7 years in a bit more than 1.5 hours given noise in actual lookup rates and a maximum parallelization factor of 40.

3 Results

In this section, we describe how we evaluate the efficacy of BTTF whois. First, we analyze its coverage in comparison to Team Cymru's historic whois. Next, we analyze how the use of BTTF whois would have improved the analysis of Fiebig et al., i.e., which additional insights would have been possible, had they used our BTTF whois implementation instead of relying on Team Cymru's bulk whois.

3.1 Experimental Setup and Assumptions

To evaluate the dataset, we work on the assumption that using Team Cymru whois data is accurate when requesting information on IP addresses *'as of now.'* Furthermore, accuracy of the Team Cymru dataset should decline, the further back we are looking, with prefixes being transferred between organizations.

Hence, we should be able to test the accuracy of our historic whois service with the following experiment: For a time period, in our case from January 2015

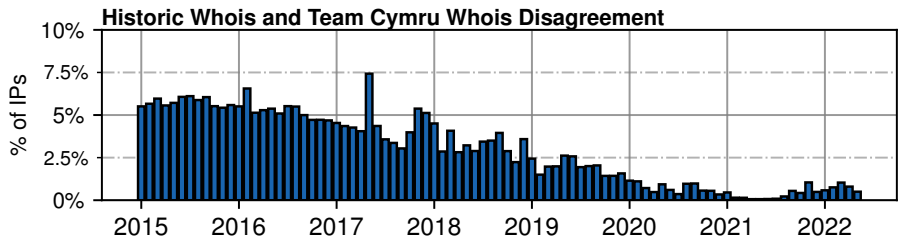

Fig. 1: Percentage of IP addresses in the dataset on which the historic whois service we implemented and the data from Team Cymru's bulk whois service disagree.

to May 2022, we gather a varying set of IP addresses. We then attribute these IP addresses to organizations, once using current Team Cymru whois data, and once requesting the historic whois state from our service. We then compare both datasets. If the implemented historic whois service is accurate, we should observe the following:

1. For recent lookups, there should be a high agreement between data from the historic whois service and the Team Cymru provided data.
2. The farther back we go in time, the higher the discrepancy becomes.

3.2 Coverage Comparison Results

To test the reliability to the historic whois service, we compared the lookup results for all IP addresses in the dataset provided by Fiebig et al. (see Fig. 1). To this end, we strictly compared the returned sets of ASes, and only considered an exact match to be agreement, i.e., if one service would return a subset of ASes of the other, we would consider this a disagreement. As depicted in Fig. 1, we indeed observe the expected pattern. While disagreement started out at around 5.5% in 2015, it continuously decreases over time, reaching a low point of 0.06% disagreement in mid 2021. Note that, thereafter, we see a slight increase in disagreement with 0.5-1.0% disagreement in early 2022. Overall, this result aligns with our predictions in terms of reliability for the historic whois service. Hence, as it is comparably reliable on data where the Team Cymru whois is reliable, we assume our historic whois to be reliable for historic data as well.

3.3 Impact of BTTF Whois on Case-Study Analysis

For demonstrating the benefits of our historic whois service, we analyze how its different perspective on IP address ownership would have influenced the results Fiebig et al. presented [9]. To this end, we compared the final cloud hosting verdict for several countries between an analysis where our historic whois service has been used and one where Team Cymru's whois has been used (see Fig. 2).

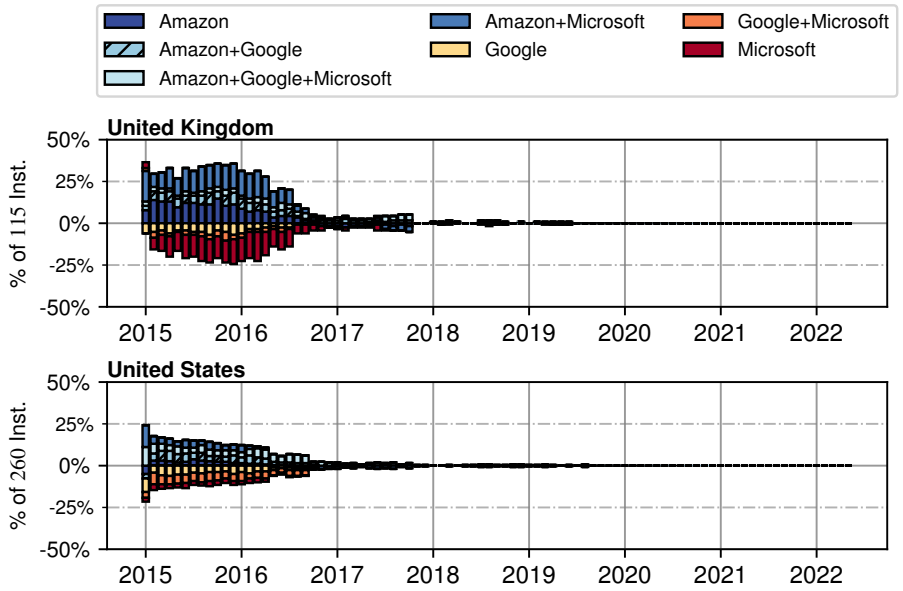

Fig. 2: Difference in cloud use attribution for universities in the U.K. and the U.S. (January 2015–May 2022) between Team Cymru bulk-whois data and historic bulk-whois data as absolute percent values as relative change considering Team Cymru as the base-line, i.e., positive values mean *more based on Team Cymru whois data*, while negative values mean *more based on historic bulk whois data*.

Over all countries in our analysis, we only observe a significant impact in the U.K. and the U.S.. Other nations show a picture similar to that of Germany, listed for comparison. For the U.K. and the U.S., we find that, overall, the number of universities attributed to Amazon (i.e., Amazon, Amazon+Google, Amazon+Microsoft, Amazon+Google+Microsoft) are estimated higher by data from Team Cymru's whois service. Using our historic whois service, these numbers drop, with corresponding increases for Google, Microsoft, and Google+Microsoft, i.e., overall we see less addresses attributed to Amazon. This effect slowly declines for the U.S., nearly completely vanishing in late 2016, while in the U.K. we observe a more significant share of this discrepancy, with a more rapid decline in 2016, with the effect being mostly gone in Q4 of 2017. Afterwards, both the U.S. and the U.K., show no measurable impact of using our historic whois implementation over Team Cymru's whois.

A closer investigation of the observed effect revealed that it is related to 18.0.0.0/8, the IPv4 address block allocated to the Massachusetts Institute of Technology (MIT). In 2017, MIT announced its intent to sell large parts (87.5%) of this address block to Amazon [29]. The transfer of addresses was finalized in 2019, with the creation of associated route objects [2], but the networks to be

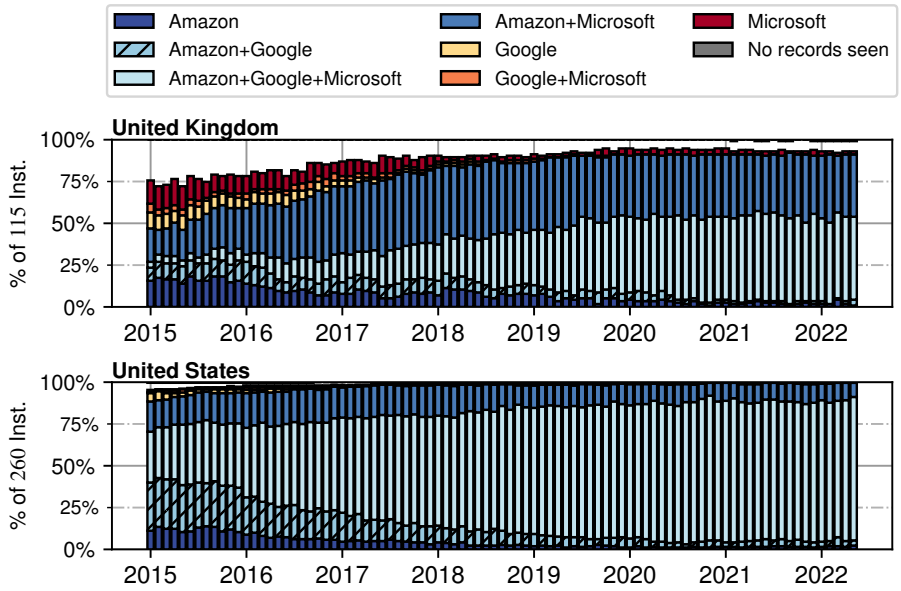

Fig. 3: Cloud use attribution for universities in the U.K. and the U.S. (January 2015–May 2022) based on Team Cymru bulk-whois data.

sold were cleared ahead of time. What our observations in the dataset mean is that several U.S. and U.K. universities that were not the MIT had DNS records pointing to hosts and services hosted at MIT.

With recent whois data attributing these netblocks to Amazon instead of MIT, we of course mis-identified Amazon cloud usage for several U.S. and U.K. universities in 2015 and 2016. To better understand the significance of this attribution error, we also compare the cloud usage graphs generated when using whois data sourced via the Team Cymru whois service (see Fig. 3) with the updated version relying on our historic whois now used in the paper (see Fig. 4). We find that for both countries, the U.S. and the U.K., using the historic whois service reveals a richer pattern in the data, even though the effect is less elaborate for the U.S.. There, we find that the initial share of universities also using Amazon hosted services now hovers around 70% instead of the around 85% initially observed. Still, this effect quickly reduces with the still continuous market growth of major cloud providers.

In the U.K. the effect has been more pronounced. Instead of the gradual increase initially assumed based on Team Cymru's whois data, we now find that the UK had a considerably low share of Amazon service usage in 2015. This quickly increased over the course of, especially, 2016. Hence, by not using our

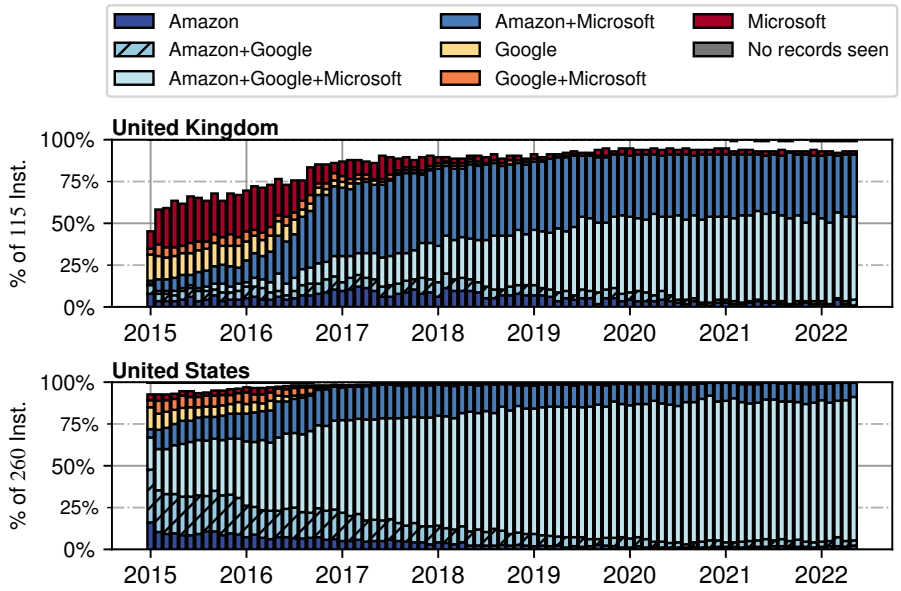

Fig. 4: Cloud use attribution for universities in the U.K. and the U.S. (January 2015–May 2022) based on historic bulk-whois data.

historic whois service, Fiebig et al. missed an important growth effect in their data.

4 Discussion

In this section, we discuss lessons learned for research on historic datasets, discuss the limitations of our approach, and outline further work.

4.1 Lessons Learned for Research on Historic Datasets

In Section 3, we have seen the major impact incorrect address attribution can have on research results when working with historic records. Especially research that investigates research questions in which IP address ownership and control is instrumental–as the case-study work by Fiebig et al.–becomes more robust by selecting a more accurate IP address attribution methodology. Given the growing availability of historic datasets containing IP addresses, for example, the Farsight SIE dataset [8], the OpenINTEL dataset [14, 27], but also historic trace-route datasets [22], or IXP datasets [5], we expect more future research to deal with historic address datasets. At the same time, the exhaustion of IPv4 [26], and the associated growth of the IP address and leasing market [21, 10, 20] will make

real-time whois information increasingly unreliable for such historic datasets. As such, our service fills an important gap for the research community.

4.2 Limitations

Despite our successful validation of the historic whois service by demonstrating it performs comparable to established bulk whois services on recent data, there are several limitations which should be discussed. First, the utilized CAIDA data exhibits several inconsistencies in data, e.g., AS numbers having a dot in the middle. The CAIDA prefix data is an aggregate of RouteViews data. The aggregation process may occlude specific announcements, e.g., if a prefix is not yet visible at the single route collector used by CAIDA. Similarly, prefix hijacks [31] may inject routes into the aggregate table, which are then wrongly attributed to the hijacking organization. This issue could only be addressed by a more elaborate data structure, that includes Internet Registry Routing object data as well as RPKI [17] data – which is difficult to obtain in historic form – and heuristics to identify and exclude route hijacks. Given the current accuracy of the historic whois service, we consider this approach as out-of-scope.

Furthermore, there are several limitations in the AS2ORG mapping data. AS family calculation [4], i.e., grouping of ASes to a common organization if the organizational objects of theses ASes are, e.g., subsidiaries of a common corporation, are unreliable over time. Fields from the whois data provided by RIRs is not consistently parsed, and fields contain faulty data if the base format on the RIRs side changes without the parsers that generate the AS2ORG extract we rely on being adjusted. In addition, the AS2ORG maps have a quarterly granularity, which makes AS2ORG attribution unreliable when changes occur, as discussed in our methodology.

Finally, as noted before, a prefix being announce by an AS does not necessarily mean that this prefix is allocated to, or owned by said AS, see the example of announcements of AS174. Hence, our implemented historic whois service will mis-attribute prefixes that are registered to an organization that is not the organization to which the announcing AS is associated.

Nevertheless, again given the observed reliability in comparison with Team Cymru's whois service, we consider the current implementation of our historic whois service as sufficiently robust to provide historic whois data. Effectively, it is comparably accurate to the commonly used Team Cymru whois service on recent data, while providing higher accuracy in historic data, as highlighted by the case of MIT's /8 network.

4.3 Future Work

As discussed in our limitations section, our reliance on the CAIDA aggregates of the Routeviews BGP announcement collections still limits the accuracy of our data. To improve our service, it would hence be advisable to not only provide routing information based IP attribution, but also access other sources for historic whois information, and attach it to returned records if it is available. For

example, RIPE NCC provides a historic non-bulk whois service. We are in conversations with relevant RIRs and registrars to obtain access to these datasets, so that our service can–along with routing based attribution information, i.e., the announcing AS–also return information from RIR databases. If these datasets become available, it would also be prudent to compare RIR information with actual routing information over the historic timeframe covered by our service.

Similarly, Routeviews data is available for a longer timeframe than the CAIDA aggregates. Hence, we also plan to aggregate Routeviews information from before the first CAIDA aggregates became available–as early as 2000–to include in our BTTF whois service.

5 Conclusion

In this paper, we introduce and evaluate BTTF whois as a public community service. This historic whois service allows more accurate estimations of IP address ownership, especially when the concerned IP address has been observed in the past. Based on a case-study, we demonstrate how the use of an accurate historic whois service allows deeper insights into datasets, and reveal developments that would remain shrouded when only relying on recently available whois information.

Nevertheless, several challenges exist, which should be resolved in further iterations of the development of our service. This includes aggregating the RouteViews dataset ourselves – especially as older data-sets are available than aggregated by CAIDA – and continuously collecting RIR provided data for generating AS2ORG maps ourselves, including addressing the issue of organizational families more reliably. Furthermore, future implementations should include IRR and RPKI data to make the implementation more robust against data noise due to prefix hijacks and the announcement of prefixes by ASes not belonging to the prefix-holder's organization.

Test our service: You can test the BTTF whois service at bttf-whois.as59645.net port tcp/10000. See Appendix A for a brief usage documentation.

References

[1] J. S. et al. *pytricia: an IP address lookup module for Python.* Aug. 30, 2022. URL: `https://github.com/jsommers/pytricia` (visited on 08/30/2022).

[2] ARIN. *WHOIS for NET-18-32-0-0-1.* Oct. 7, 2019. URL: `https://whois.arin.net/rest/net/NET-18-32-0-0-1` (visited on 09/01/2022).

[3] CAIDA. *Routeviews Prefix to AS mappings Dataset for IPv4 and IPv6.* Aug. 30, 2022. URL: `https://www.caida.org/catalog/datasets/routeviews-prefix2as/` (visited on 08/30/2022).

[4] CAIDA. *The CAIDA AS Organizations Dataset, all dates.* Aug. 30, 2022. URL: `https://www.caida.org/data/as-organizations` (visited on 08/30/2022).

[5] N. Chatzis, G. Smaragdakis, J. Böttger, T. Krenc, and A. Feldmann. "On the benefits of using a large IXP as an Internet vantage point". In: *Proceedings of the 2013 conference on Internet measurement conference.* 2013.

[6] E. Chen and J. Stewart. *A Framework for Inter-Domain Route Aggregation*. RFC 2519. IETF, Feb. 1999. URL: http://tools.ietf.org/rfc/rfc2519.txt.

[7] L. Daigle. *WHOIS Protocol Specification*. RFC 3912. IETF, Sept. 2004. URL: http://tools.ietf.org/rfc/rfc3912.txt.

[8] Farsight Inc. *Farsight - Security Information Exchange (SIE)*. URL: https://www.farsightsecurity.com/solutions/security-information-exchange/.

[9] T. Fiebig, S. Gürses, C. H. Gañán, E. Kotkamp, F. Kuipers, M. Lindorfer, M. Prisse, and T. Sari. "Heads in the clouds: measuring the implications of universities migrating to public clouds". In: *arXiv preprint arXiv:2104.09462* (2021).

[10] V. Giotsas, I. Livadariu, and P. Gigis. "A first look at the misuse and abuse of the IPv4 Transfer Market". In: *International Conference on Passive and Active Network Measurement*. Springer. 2020.

[11] V. Giotsas, M. Luckie, B. Huffaker, and K. Claffy. "Inferring complex AS relationships". In: *Proceedings of the 2014 Internet Measurement Conference*. 2014.

[12] J. Haas and J. Mitchell. *Reservation of Last Autonomous System (AS) Numbers*. RFC 7300. IETF, July 2014. URL: http://tools.ietf.org/rfc/rfc7300.txt.

[13] J. Hawkinson and T. Bates. *Guidelines for creation, selection, and registration of an Autonomous System (AS)*. RFC 1930. IETF, Mar. 1996. URL: http://tools.ietf.org/rfc/rfc1930.txt.

[14] O. Hohlfeld. "Poster: Operating a DNS-based Active Internet Observatory". In: *Proc. of the 2018 ACM SIGCOMM Conference (SIGCOMM)*. 2018.

[15] R. Housley, J. Curran, G. Huston, and D. Conrad. *The Internet Numbers Registry System*. RFC 7020. IETF, Aug. 2013. URL: http://tools.ietf.org/rfc/rfc7020.txt.

[16] G. Huston. *Autonomous System (AS) Number Reservation for Documentation Use*. RFC 5398. IETF, Dec. 2008. URL: http://tools.ietf.org/rfc/rfc5398.txt.

[17] G. Huston and G. Michaelson. *Validation of Route Origination Using the Resource Certificate Public Key Infrastructure (PKI) and Route Origin Authorizations (ROAs)*. RFC 6483. IETF, Feb. 2012. URL: http://tools.ietf.org/rfc/rfc6483.txt.

[18] W. Kumari, R. Bush, H. Schiller, and K. Patel. *Codification of AS 0 Processing*. RFC 7607. IETF, Aug. 2015. URL: http://tools.ietf.org/rfc/rfc7607.txt.

[19] S. Liu, I. Foster, S. Savage, G. M. Voelker, and L. K. Saul. "Who is. com? Learning to parse WHOIS records". In: *Proceedings of the 2015 Internet Measurement Conference*. 2015.

[20] I. Livadariu, A. Elmokashfi, and A. Dhamdhere. "On IPv4 transfer markets: Analyzing reported transfers and inferring transfers in the wild". In: *Computer Communications* 111 (2017).

[21] I. Livadariu, A. Elmokashfi, A. Dhamdhere, and K. Claffy. "A first look at IPv4 transfer markets". In: *Proceedings of the ninth ACM conference on Emerging networking experiments and technologies*. 2013.

[22] M. Luckie, Y. Hyun, and B. Huffaker. "Traceroute probe method and forward IP path inference". In: *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*. 2008.

[23] J. Mitchell. *Autonomous System (AS) Reservation for Private Use*. RFC 6996. IETF, July 2013. URL: http://tools.ietf.org/rfc/rfc6996.txt.

[24] L. Prehn, F. Lichtblau, and A. Feldmann. "When Wells Run Dry: The 2020 IPv4 Address Market". In: *Proc. of the ACM Conference on emerging Networking EXperiments and Technologies (CoNEXT)*. 2020.

[25] Y. Rekhter, B. Moskowitz, D. Karrenberg, G. J. d. Groot, and E. Lear. *Address Allocation for Private Internets*. RFC 1918. IETF, Feb. 1996. URL: http://tools.ietf.org/rfc/rfc1918.txt.

[26] P. Richter, M. Allman, R. Bush, and V. Paxson. "A primer on IPv4 scarcity". In: *ACM SIGCOMM Computer Communication Review* 45.2 (2015).

[27] R. van Rijswijk-Deij, M. Jonker, A. Sperotto, and A. Pras. "A High-Performance, Scalable Infrastructure for Large-Scale Active DNS Measurements". In: *IEEE Journal on Selected Areas in Communications* 34.6 (2016).

[28] RouteViews. *RouteViews Project*. Aug. 30, 2022. URL: http://www.routeviews.org (visited on 08/30/2022).

[29] M. A. Schmidt and I. R. Executive. *Letter to: To the members of the MIT community*. Apr. 20, 2017. URL: https://gist.github.com/simonster/e22e50cd52b7dffcf5a4db2b8ea4cce0 (visited on 09/01/2022).

[30] K. Z. Sediqi, L. Prehn, and O. Gasser. "Hyper-specific prefixes: gotta enjoy the little things in interdomain routing". In: *ACM SIGCOMM Computer Communication Review* 52.2 (2022).

[31] P. Sermpezis, V. Kotronis, A. Dainotti, and X. Dimitropoulos. "A survey among network operators on BGP prefix hijacking". In: *ACM SIGCOMM Computer Communication Review* 48.1 (2018).

[32] Q. Vohra and E. Chen. *BGP Support for Four-Octet Autonomous System (AS) Number Space*. RFC 6793. IETF, Dec. 2012. URL: http://tools.ietf.org/rfc/rfc6793.txt.

[33] L. Zhou, N. Kong, S. Shen, S. Sheng, and A. Servin. *Inventory and Analysis of WHOIS Registration Objects*. RFC 7485. IETF, Mar. 2015. URL: http://tools.ietf.org/rfc/rfc7485.txt.

A BTTF Whois Short Documentation

Here, we document a) how you can use BTTF whois manually with netcat, and b) how to obtain bulk results. Furthermore, we provide an overview over the returned JSON's structure.

A.1 Using BTTF Whois Manually

To manually use BTTF whois, you have to use either `netcat`/`nc` or `telnet`. The date format is YYYYMMDD.

```
% nc 65.21.106.239 10000
# This is the historic IP to AS mapping service
# Contact: <REDACTED_FOR_PEER_REVIEW>
# Trie Status: READY - loaded 2169698 IPv4 and 313354 IPv6 prefixes
# AS2Org Status: 119005 AS and 199948 organisations loaded
# Enter HELP to get basic usage information
# NOTICE: OUTPUT FORMAT: JSON-SHORT
# READY
begin
1.1.1.1 20210101
```

```
{"IP": "1.1.1.1", "QDATE": "20210101", "results": {"DATA_FIRST": [...]
end
# goodbye
```

A.2 Using BTTF Whois for Bulk Requests

To use BTTF whois to handle bulk requests, you can create a file with 'begin' on the first line, and 'end' on the last line, containing IP addresses and dates in between:

```
% cat ./file
begin
1.1.1.1 20210101
1.1.1.1 20120101
8.8.8.8 20210201
end
```

You can then use `netcat`/`nc` to send this file to the bulk whois service and receive the results directly or redirect them to a file:

```
% cat ./file | nc 65.21.106.239 10000
# This is the historic IP to AS mapping service
# Contact: <REDACTED_FOR_PEER_REVIEW>
# Trie Status: READY - loaded 2169926 IPv4 and 319033 IPv6 prefixes
# AS2Org Status: 119005 AS and 199948 organisations loaded
# Enter HELP to get basic usage information
# NOTICE: OUTPUT FORMAT: JSON-SHORT
# READY
{"IP": "1.1.1.1", "QDATE": "20210101", "results": {"DATA_FIRST": [...]
{"IP": "1.1.1.1", "QDATE": "20120101", "results": []}
{"IP": "8.8.8.8", "QDATE": "20210201", "results": {"DATA_FIRST": [...]
# goodbye
```

A.3 BTTF Whois JSON Data Structure

Below, you can find an overview of the response fields returned by BTTF whois.

```
{
  # Requested IPv4 or IPv6 address
  "IP": "1.1.1.1",
  # Date for which data was requested
  "QDATE": "20210101",
  "results": {
    # First time the most specific prefix for address has been seen
    # first with this specific set of announcing ASes
    "DATA_FIRST": 20180320,
    # Last time this entry was seen, i.e., valid until. If it is
```

```
    # null, the most specific is still visible in the most recent
    # dataset (valid NOW).
    "DATA_LAST": null,
    # List of ASNs that announced the most specific prefix for the
    # requested address.
    "asns": [
      13335
    ],
    # The most specific matching prefix from the dataset.
    "prefix": "1.1.1.0/24",
    # AS2ORG mappings for all announcing ASN.
    "as2org": [
      {
        # AS number
        "ASN": 13335,
        # AS name
        "ASNAME": "CLOUDFLARENET-AS",
        # RIR that is the data source in the AS2ORG mappings
        "RIR": "RIPE",
        # Org objects associated with the ASN
        "orgs": [
          {
            # Country code attributed to an organization
            "CC": "US",
            # RIRs that hold an instance of this ORG object
            "RIR": "ARIN,RIPE",
            # Organization name from the ORG object
            "ASORG": "Cloudflare Inc"
          }
        ]
      }
    ]
  }
}
```