WHOactuallyIS? Finding the Companies Behind the Networks

Joshua Levett, Vassilios Vassilakis, Poonam Yadav
Department of Computer Science, University of York

Abstract

Extracting accurate data from WHOIS and Registration Data Access Protocol (RDAP) services currently depends on the level of transparency and willingness of organisations to update their information within Internet registries, including after mergers and acquisitions. What this never shares, however, is what resources are used for, and whether their users are the same organisations as their owners. With the rise of IP leasing and long-seen prevalence of hosting platforms or PaaS providers, the information returned is not fully representative. In this paper, we discuss the challenges of identifying the companies behind networks, and our approach to building a WHOactuallyIS? service, which uses a variety of different data services to provide contextually collected information about the owners and users of Internet resources.

I. Introduction

The Internet is – by design – a connected mesh of independent entities where little is known (or needed to be known) about other networks to connect. This makes the Internet highly scalable but also presents challenges in acquiring even high-level information about other networks, and particularly, their operators. From the perspective of researchers and other operators, it then requires a manual task using technical-user-centric tooling to determine whether any two networks have the same operator, where they may be registered, and who may be a relevant point of contact. For law enforcement agencies, this is an even greater challenge, as it restricts the acquisition of such information to people with particular technical expertise.

For decades, the standardised approach to querying Internet resource registries has been with WHOIS, a tool originally introduced to provide contact information within the ARPANET [1]. Its most recent derivative, RFC 3912 [2], stores and retrieves records using the Routing Policy Specification Language (RPSL) [3], [4], a text-based format with attributes separated from their keys with a colon. More recently, an effectively derivative protocol, the Registration Data Access Protocol (RDAP) specification [5] has been introduced by a number of registries to help standardise query and output structures, in addition to improving exchange security and broadening support for different local or language requirements, with the intention of ultimately replacing WHOIS. Based on HTTP, RDAP allows for REST-based queries that can be returned in the machine-readable JSON format. However, despite its introduction over 10 years ago, the implementation of the protocol at domain registry level is only a relatively recent development, with generic Top-Level Domains (TLDs) switching over in January 2025 as a contractual obligation, and many country-code TLDs still not providing RDAP services. This in some ways introduces a new challenge – wherein some registries have already retired WHOIS tooling, and others have not yet implemented RDAP, meaning there is now no single approach to querying registry information.

The assignment of IP addresses and Autonomous System Numbers (ASNs) is undertaken through a structure of delegated bodies, centred around the Internet Assigned Numbers Authority (IANA) who are responsible for the overall management and allocation of IP addresses. The IANA delegates blocks of IP addresses to the five Regional Internet Registries (RIRs): AFRINIC, APNIC, ARIN, LACNIC and the RIPE NCC, each of whom is responsible for further delegation (to a Local Internet Registry, or LIR) or assignment within their respective service regions. A similar system exists for domain names. At the top of the hierarchy is the IANA who manage the DNS root zone, and as part of this function, are responsible for the delegation of each TLD to domain name registries. Information about a given domain name's registrant is obtained through the delegated registry for that TLD – but unlike in the case of IP addresses, the number of these registries is far more substantial, with hundreds of different generic and country-code based TLD variations. These TLD registries are also able to delegate the management of further levels (such as the commonly used second-level domain `ac.` for education institutions). This scheme of delegation means that information about the registered owner of any given IP address, ASN, or domain name is not held centrally.

A further challenge is that these registries effectively operate in silos, and therefore an entity may have inconsistent details recorded in each registry. For researchers or operators querying an entity's resources, this means knowing both *where* an organisation may have registered a resource and *by which name*. For instance, an entity may use the name "`Example Operator, Inc`" with one registry, "`Example Operator Ltd`" with another and "`Example Subsidiary`" with another.

In this extended abstract, we outline a tool, *WHOactuallyIS?*, which attempts to mitigate these issues and returns more detailed and comprehensive information about the companies operating Internet networks. WHOactuallyIS? takes as input an IP address, ASN or domain name, and through a series of recursive registry queries and response matching techniques, achieves both a list of connected Internet resources owned by the same entity, and in doing so identifies the most representative organisation behind these resources (which may not be obvious, especially in the case of mergers or subsidiaries). The codebase for the tool is available on GitHub[1].

[1] https://github.com/LevettJ/who-actually-is

II. Related Work

There are numerous studies, such as [6] that use the information provided by Internet registries, most especially in collecting name and contact information to associate with Internet resources, but this has the limitation that business relationships (including parent organisations or mergers) are not captured, as such information is not supplied to the registry. This challenge remains a limitation even in recent work aiming to resolve the issue of linking information about Internet resources to the resources themselves. In [7], inferences between resources and organisations depend on the information from multiple sources, including the PeeringDB [8] dataset of self-declared network information, supported by the work of Chen et al. [9] in improving AS-to-organisation mappings. Whilst this work is significant in determining the presence of sibling ASes (Internet networks operated by the same organisation), our work makes further progress – extending beyond AS-level inferencing to determine for any Internet resource (AS, IP address or domain name) its related organisation(s).

III. Methodology & Implementation

Fundamentally, the concept of WHOactuallyIS? is to explore the pool of resources associated with a given input in place of solely the target resource to be able to identify a more complete perception of who the resource owner and operator is. For larger organisations, the owner and operator entities may be the same, such as in Fig 1. In other cases, such as for websites hosted with smaller or independent hosting companies, the resource owners and operators (for each of the domain, IP address and ASN) may all be different.

System Overview. If the initial resource is an IP address, we perform a Reverse DNS (rDNS) lookup, potentially returning a short list of associated DNS records. For an input domain name, a forward DNS (fDNS) lookup returns associated IP address records for which a rDNS lookup can then be performed. This results in the collection of three datasets (which can be zero-length, if, for instance, an rDNS `PTR` record has not been configured):

- **IP address(es).** The input IP address; or a list of the IP addresses associated with `A` or `AAAA` records of the input domain name.
- **fDNS.** The `A` or `AAAA` records associated with the IP address list. Some commercial solutions can provide a more complete list of associated `A` or `AAAA` records by retaining DNS lookup results or brute-force DNS enumeration, but this can miss results or be quickly outdated. We do not undertake this step meaning our fDNS results are more restricted but simultaneously guaranteed to be more current.
- **rDNS.** `PTR` records associated with the IP address list.

IP Ownership and Usage. For each address in the IP address list, we query the applicable RIR's RDAP service. The returned RDAP objects are neither consistent within, or between, registries, and so require a degree of parsing to be able to obtain usable information. Of particular interest are the *entities* objects related to a given resource. These have partially standardised types, including *registrants* (who the resource is registered to), *registrar* (information about who registered the resource – usually an RIR or LIR) and an *administrative* contact (usually where the `mntner` handle is present). This latter contact can

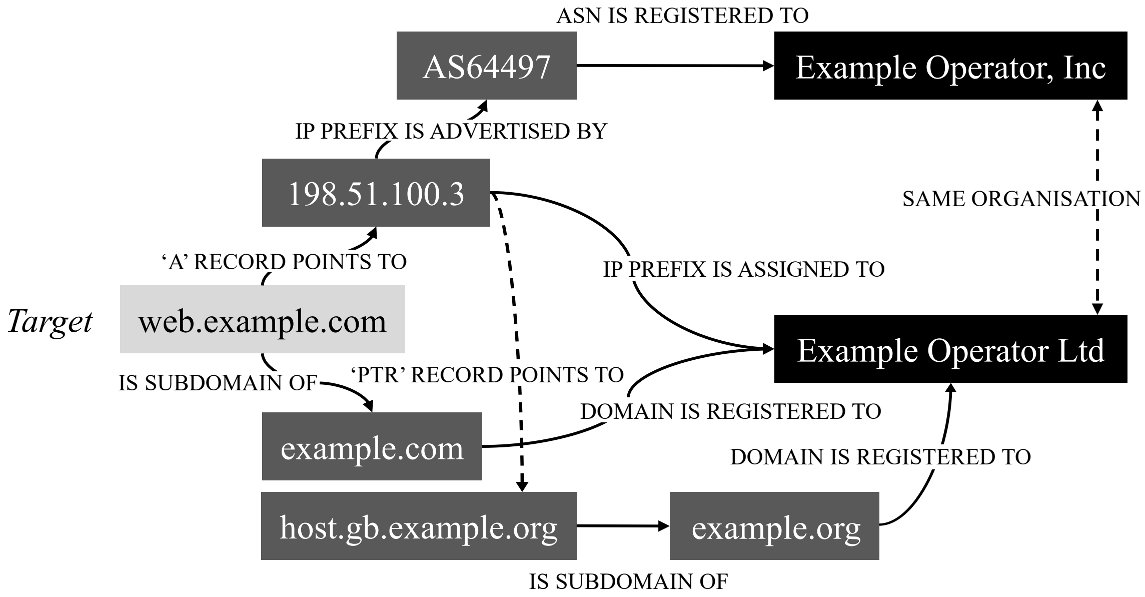

Fig. 1. The concept of WHOactuallyIS? – presented with an input target, to use the related network information (such as the IP addresses and ASNs associated with a domain) to determine both the owner and operator of the resource (which may or may not be the same).

also assist in identifying leased IP addresses as the administrative contact is often set to an `mntner` entity associated with the lessor [10].

Domain Name Owners. For each domain in the fDNS or rDNS list, the parent DNS zone is identified (which may itself be a common domain of note). The domain and its parent zone are then subject to RDAP queries to find associated registry records, and responses are similarly parsed to extract *entities* objects.

Registered and Advertised ASN. For each IP address, we compare the registered entity with the entity associated with the ASN advertising the smallest prefix within which the applicable IP address is contained as seen by the RIPE Routing Information Service (RIS) [11]. This captures the network to which traffic is routed, and hence we can determine whether the *owning* and *operating* networks are the same. Where the AS numbers match, it is likely that the owner is also the operator at a network level, but where there is an AS number disparity there is the potential for further resources from the same organisation to be identified (which can be validated using attributes returned through RDAP). We supplement information gathered through BGP capture and RDAP queries with PeeringDB [8].

Matching Entities. For each entity we extract key attributes, including the type of entity (for instance, *administrative*), the RIR-assigned `handle` (if present), a name, an address, and contact information (such as telephone or email). We then begin the process of using string matching techniques to identify common entities. Attempting first to detect name-based matches (such as the previous "Example Operator, Inc"), we use Hamming distance to check the similarity between strings over a lower length bound of 6. We then also measure the Levenshtein distance between the strings (which we anticipate to more regularly return higher levels of similarity). If a Hamming match is over 80%, or the Levenshtein match is above 80% with a Hamming match over 40%, we assume the name matches. We employ a similar technique with addresses. Compiling address information into a singular string, we identify both the Hamming and Levenshtein distances and assume two addresses are the same where the Hamming match is over 70%, or the Levenshtein match is over 70% with a Hamming match over 30%. The lower thresholds for address-based matches account for more varieties in address records, which may skip fields or present them in different orders depending on the registry.

A similar approach cannot be applied to contact information. If a telephone number is present, we require a complete match (including country code) to be valid. For email addresses we prefer complete address matches. It is also possible to match email addresses using simply the domain name portion, however some entities use common email service providers or inbox anonymising services, leading to false positive matches.

IV. Results & Evaluation

We prepare a dataset of Internet resources for which to compare the findings of our WHOactuallyIS? tool against manual analysis. We devise a dataset of 200 Internet resources containing 100 domain names sourced from the ranked list of domain names provided by Cloudflare [12] based on user queries to the '1.1.1.1' DNS resolver service; 50 IP addresses (both IPv4 and IPv6) sourced at random from RIPE Atlas [13] traceroute queries; and 50 highest-ranked ASNs by direct user access from the APNIC Eyeballs dataset [14]. To supplement the collected data, we manually investigate target resources. For domain names, we attempt to establish who is operating the resource by visiting associated website(s) and by using WHOIS to lookup both the domain name and resolved IP address(es). Similarly, for IP addresses we use WHOIS to resolve a likely owner, and we investigate any domain names returned from rDNS queries alongside bgp.tools to establish who is likely to be using the IP address. For ASNs, we use a combination of PeeringDB and bgp.tools services to identify the likely owner of the ASN.

Initial Findings. We manually investigated each resource in the dataset and compared this with the WHOactuallyIS? response. There were a small number of target resources for which WHOactuallyIS? failed to return a result: IP addresses with no associated DNS resources returned no result, but we found these usually related to end-user Internet service provider addresses; Bytedance CDN-related domains were particularly complex, which may relate to recent changes in the way TikTok content is delivered [15] wherein little DNS data or registrant information was returned; domains (almost exclusively `.com`) for which registry information has been redacted and no `A` or `AAAA` records exist, or such records point only to a CDN provider. For each of these cases, it was not possible to obtain much information about any resource owner from information available through network-level queries.

For 93.1% of resources, the manually identified parent owner or operator matched that of the WHOactuallyIS? result or one of its listed 'also known as' names, with 83.0% of the primary names matching their manually sourced equivalents.

Comparison with WHOIS. We compared WHOactuallyIS?-returned results with simple WHOIS queries. Data returned in this way will be of lesser value in future following the retirement of WHOIS services, and indeed a number of WHOIS responses returned warning information about the impending discontinuation of the registrar WHOIS endpoints queried. In total, we found that only 81.1% of the WHOIS responses matched that of the manual discovery process to find the resource owner (11.9% less than that achieved by WHOactuallyIS?). Furthermore, the nature of WHOIS means that only the resource owner, rather than any operator(s), can be determined.

Related Company Data. In this current iteration of WHOactuallyIS?, we use information exclusively sourced from the UK-based Companies House, which limits the findings of the *associated legal entity* aspect of the output to companies based or (in the case of some overseas companies with UK presence) registered in the UK. Nonetheless, where in the validation dataset a UK entity was manually established by finding entities with similar company names, registered offices or company directors with an address associated with an entity, WHOactuallyIS? was able to make the same link in a marginal majority of cases. Extending the WHOactuallyIS? tool with use of a global company data API would likely achieve a higher level of resulting accuracy.

V. Future Work & Conclusion

We present and demonstrate WHOactuallyIS?, a tool for retrieving the identities of the companies owning and operating critical Internet resources, substantially extending the utility of existing registry querying services by connecting returned resources and records, providing more comprehensive information to network operators and more representative information for law enforcement. We found that WHOactuallyIS? performed well against our validation targets, providing accurate ownership and operator information for 93.1% of the targets, a significant improvement on existing WHOIS-based services. In future work, we intend to: evaluate further string matching techniques and thresholds to increase both the quantity and accuracy of entity matches; integrate additional company datasets beyond Companies House to incorporate globally registered legal entities; and explore the potential for WHOactuallyIS? to produce a static snapshot dataset, which could for example be used to improve the detection of ‘sibling’ resources which share an owning entity.

Acknowledgments

This work is supported, in part, by EPSRC and DSIT TMF-uplift: CHEDDAR: Communications Hub For Empowering Distributed ClouD Computing Applications And Research (EP/X040518/1), (EP/Y037421/1) and EPSRC REMOTE (EP/Y019229/1).

References

[1] K. Harrenstien and V. White, “NICNAME/WHOIS,” RFC 812, Mar. 1982. [Online]. Available: https://www.rfc-editor.org/info/rfc812

[2] L. Daigle, “WHOIS Protocol Specification,” RFC 3912, Sep. 2004. [Online]. Available: https://www.rfc-editor.org/info/rfc3912

[3] D. Kessens, T. J. Bates, C. Alaettinoglu, D. Meyer, C. Villamizar, M. Terpstra, D. Karrenberg, and E. P. Gerich, “Routing Policy Specification Language (RPSL),” RFC 2622, Jun. 1999. [Online]. Available: https://www.rfc-editor.org/info/rfc2622

[4] J. da Silva Damas, A. Robachevski, L. Blunk, and F. Parent, “Routing Policy Specification Language next generation (RPSLng),” RFC 4012, Mar. 2005. [Online]. Available: https://www.rfc-editor.org/info/rfc4012

[5] A. Newton and S. Hollenbeck, “Registration Data Access Protocol (RDAP) Query Format,” RFC 7482, Mar. 2015. [Online]. Available: https://www.rfc-editor.org/info/rfc7482

[6] A. Arturi, E. Carisimo, and F. E. Bustamante, “as2org+: Enriching as-to-organization mappings with PeeringDB,” in *Passive and Active Measurement: 24th International Conference, PAM 2023, Virtual Event, March 21–23, 2023, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2023, p. 400–428. [Online]. Available: https://doi.org/10.1007/978-3-031-28486-1_17

[7] R. Fontugne, M. Tashiro, R. Sommese, M. Jonker, Z. S. Bischof, and E. Aben, “The wisdom of the measurement crowd: building the Internet Yellow Pages a knowledge graph for the internet,” in *Proceedings of the 2024 ACM on Internet Measurement Conference*, ser. IMC ’24. New York, NY, USA: Association for Computing Machinery, 2024, p. 183–198. [Online]. Available: https://doi.org/10.1145/3646547.3688444

[8] PeeringDB, “PeeringDB.” [Online]. Available: https://www.peeringdb.com/

[9] Z. Chen, Z. S. Bischof, C. Testart, and A. Dainotti, “Improving the inference of sibling autonomous systems,” in *Passive and Active Measurement*, A. Brunstrom, M. Flores, and M. Fiore, Eds. Cham: Springer Nature Switzerland, 2023, pp. 345–372.

[10] B. Du, R. Fontugne, C. Testart, A. C. Snoeren, and K. C. Claffy, “Sublet your subnet: Inferring IP leasing in the wild,” in *Proceedings of the 2024 ACM on Internet Measurement Conference*, ser. IMC ’24. New York, NY, USA: Association for Computing Machinery, 2024, p. 328–336. [Online]. Available: https://doi.org/10.1145/3646547.3689010

[11] RIPE NCC, “Routing Information Service (RIS).” [Online]. Available: https://www.ripe.net/analyse/internet-measurements/routing-information-service-ris/routing-information-service-ris

[12] Cloudflare, “Cloudflare Radar.” [Online]. Available: https://radar.cloudflare.com

[13] RIPE NCC, “RIPE Atlas.” [Online]. Available: https://atlas.ripe.net/

[14] G. Huston, “How big is that network?” Oct. 2014. [Online]. Available: https://labs.apnic.net/index.php/2014/10/02/how-big-is-that-network/

[15] D. Madory, “TikTok emerges from shutdown without Bytedance’s US CDN,” Jan. 2025. [Online]. Available: https://www.kentik.com/blog/tiktok-emerges-from-shutdown-without-bytedances-us-cdn/

Appendix A
Reference Acronyms

- **API** Application Programming Interface
- **ASN** Autonomous System Number
- **CDN** Content Delivery Network
- **DNS** Domain Name System
- **fDNS** Forward DNS
- **rDNS** Reverse DNS
- **IANA** Internet Assigned Numbers Authority
- **JSON** JavaScript Object Notation
- **LIR** Local Internet Registry
- **RDAP** Registration Data Access Protocol
- **REST** Representational State Transfer
- **RIR** Regional Internet Registry
- **RPSL** Routing Policy Specification Language
- **TLD** Top-Level Domain