IP geolocation through the lens of an academic: Where do we stand?

Kevin Vermeulen
LIX, CNRS, Ecole Polytechnique

ABSTRACT

IP geolocation and academic research is a long-told story. And yet, after decades of research, it is still an open problem, and there exists no public, accurate, large-scale, IP geolocation dataset coming from the academic world. However, IP geolocation is one of the most used metadata in the community, and researchers would love to have access to such a dataset. As both recent users of IP geolocation data and contributors to design of new IP geolocation techniques, we share in this paper our experience on how the research community apprehends geolocation, and open a discussion on research on IP geolocation and its usage and how we could improve it.

1 INTRODUCTION AND MOTIVATION

For academics, IP geolocation has all the properties of an interesting research problem. It is very easy to understand, can be tackled with multiple approaches, and has a multitude of use cases. Indeed, after all, it is just mapping an IP address to its physical location. When looking a bit at the work that has been done in the field, one realizes that they could use different types of network measurements, such as latency [6, 9, 10, 20, 22], topology [4, 12], DNS [3, 11, 14, 18], statistical techniques [6], machine learning techniques [7], and even recently, generative AI [19]. With all this prior work, one could wonder where to find the public datasets associated with the different techniques, and take the union of them to have a large-scale, accurate IP geolocation dataset that has been validated by the research community, or the code to replicate the different techniques and obtain their own dataset. However, it does not exist. It does not exist because a lot of academic research in IP geolocation (and still today!) do not provide a usable dataset for other researchers, or a code to use the techniques.

The consequences are two-fold: First, as there is no consensus of what “good” geolocation is among the community, researchers perform their own IP geolocation in various ways, without using the state-of-the-art techniques. Some common techniques include latency measurements from vantage points with known geolocation [8, 17, 21], retrieving geolocation information from rDNS names [2, ; 17; 13], or use private geolocation databases [15], or use a combination of these different methods [17]. Given the lack of consensus, our experience, on both the author and the reviewer side, is that there is often, if not always, a discussion around the geolocation methodology. We argue that the community needs to define its expectations of what a “good” geolocation methodology is, so IP geolocation is no longer a subject of discussion when this methodology is respected. In this paper, we propose to start the discussion to reach this goal (§2).

Second, for designers, the absence of code and datasets of state-of-the-art techniques can be frustrating. For instance, some prior work claimed “street level geolocation” [20], whereas their technique was evaluated on a few dozens of IP addresses. As designers of new IP geolocation techniques, we wanted to replicate the street level technique to compare its performance with ours. Carefully replicating this paper took us a tremendous amount of work, and can certainly discourage students from working on this topic. Fortunately, the IMC conference recently encouraged replicability and started a replicability track [1], and our work naturally found its place there [5]. In this paper, we want to share our experience and difficulties that we faced during our replication work, and want to open a discussion to define strong recommendations to help the research in IP geolocation (§3).

2 TOWARDS A CONSENSUS ABOUT ACCEPTABLE IP GEOLOCATION METHODOLOGY

As users of IP geolocation in our different networked systems, we often faced common questions about the geolocation methodology. We split the questions according to the different methods that we describe in more detail: Latency measurements, rDNS names, and private geolocation databases.

2.1 Latency measurements

Shortest ping, Constrained Based Geolocation [9], Topology Based Geolocation (TBG) [12], and all the followup systems using more sophisticated versions of these techniques have the same fundamental idea to use latency measurements from vantage points with known geolocation, to derive the geolocation of a target IP address. There are common questions when one uses these techniques: (1) What is the latency threshold to use to demonstrate a certain degree of accuracy (e.g., country or city level), (2) When using vantage points with known geolocation, but that contain errors (e.g., RIPE

Atlas), how do you deal with the vantage points having an erroneous geolocation?

2.2 Reverse DNS names

DRoP [11], HLOC [18], Hoiho [14], Aleph [19] and other prior work [3] all use the idea of decoding geolocation hints from rDNS names. When using these techniques, there is always the question of how to deal with incorrectness of the DNS hint, e.g., DNS information can be stale, or even voluntarily erroneous.

2.3 Private geolocation databases

Geolocation databases suffer from a lack of explainability. Although they offer a far superior coverage compared to latency measurements which do not work on IP addresses not responding to pings, and to rDNS names that often do not contain any geolocation hint, they lack explainability, i.e., how the geolocation was obtained. And this is largely problematic as they do not have perfect accuracy, because the geolocation of any IP address could be considered suspicious. When using these databases, we often face criticism about them containing errors and that one wants to perform additional measurements to confirm their geolocation.

2.4 Opening a discussion

Given these three different types of geolocation techniques, their limitations and the questions that they raise, can we come up with a methodology that one could use and would be acceptable by the community?

3 GUIDELINES FOR RESEARCH IN IP GEOLOCATION

As designers of new IP geolocation techniques [16], we call the community to share their efforts in the replication of prior work, in particular to share their code, and build open datasets for evaluating IP geolocation techniques.

3.1 Towards open code and datasets for evaluating IP geolocation techniques

When evaluating geolocation techniques, one usually wants to compare state-of-the-art techniques with a new one, and evaluate on a ground truth dataset. For comparison with prior work, sharing the code is essential to help replicability, fairer comparisons, and faster progress. A lot of other fields require code to be published, and if there is no strong reason to not do it (e.g., intellectual property), this should be the norm, and not the exception.

For the evaluation on ground truth, only a few IP addresses with their geolocation are publicly available. For instance, the RIPE Atlas anchors represent only a few hundreds IP addresses. We argue that building an open dataset for evaluating geolocation techniques would be a mine of gold. This dataset could not only contain IP addresses with their geolocation, but also contain their type, such as client, router, or server IP addresses. In addition, the same thing could be done for vantage points, if one designs a new geolocation technique based on latency measurements.

3.2 Opening a discussion

We want to open the discussion around the form that code sharing and open geolocation datasets could take. For instance, there are questions such as whether we could build an evaluation framework, or how to maintain and update the open geolocation dataset, which format it would take, who can update it, etc…

4 CONCLUSION

We shared our experience of using and designing IP geolocation techniques, which is one the most used and important metadata on IP addresses. We hope that this paper will allow us to start a discussion to improve how research uses and design IP geolocation.

REFERENCES

[1] ACM IMC. 2023. Replicability Track. https://conferences.sigcomm.org/imc/2023/cfp/.

[2] Scott Anderson, Loqman Salamatian, Zachary S Bischof, Alberto Dainotti, and Paul Barford. 2022. iGDB: connecting the physical and logical layers of the internet. In *Proceedings of the 22nd ACM Internet Measurement Conference*. 433–448.

[3] Ovidiu Dan, Vaibhav Parikh, and Brian D Davison. 2021. IP geolocation through reverse DNS. *ACM Transactions on Internet Technology (TOIT)* 22, 1 (2021), 1–29.

[4] Ovidiu Dan, Vaibhav Parikh, and Brian D Davison. 2021. IP geolocation using traceroute location propagation and IP range location interpolation. In *Companion Proceedings of the Web Conference 2021*. 332–338.

[5] Omar Darwich, Hugo Rimlinger, Milo Dreyfus, Matthieu Gouel, and Kevin Vermeulen. 2023. Replication: Towards a publicly available internet scale ip geolocation dataset. In *Proceedings of the 2023 ACM on Internet Measurement Conference*. 1–15.

[6] Brian Eriksson, Paul Barford, Bruce Maggs, and Robert Nowak. 2012. Posit: a lightweight approach for IP geolocation. *ACM SIGMETRICS Performance Evaluation Review* 40, 2 (2012), 2–11.

[7] Brian Eriksson, Paul Barford, Joel Sommers, and Robert Nowak. 2010. A learning-based approach for IP geolocation. In *International Conference on Passive and Active Network Measurement*. Springer, 171–180.

[8] Vasileios Giotsas, Thomas Koch, Elverton Fazzion, Ítalo Cunha, Matt Calder, Harsha V Madhyastha, and Ethan Katz-Bassett. 2020. Reduce, reuse, recycle: Repurposing existing measurements to identify stale traceroutes. In *Proceedings of the ACM Internet Measurement Conference*. 247–265.

[9] Bamba Gueye, Artur Ziviani, Mark Crovella, and Serge Fdida. 2004. Constraint-based geolocation of internet hosts. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*. 288–293.

[10] Zi Hu, John Heidemann, and Yuri Pradkin. 2012. Towards geolocation of millions of IP addresses. In *Proceedings of the 2012 Internet Measurement Conference*. 123–130.

[11] Bradley Huffaker, Marina Fomenkov, and KC Claffy. 2014. DRoP: DNS-based router positioning. *ACM SIGCOMM Computer Communication Review* 44, 3 (2014), 5–13.

[12] Ethan Katz-Bassett, John P John, Arvind Krishnamurthy, David Wetherall, Thomas Anderson, and Yatin Chawathe. 2006. Towards IP geolocation using delay and topology measurements. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*. 71–84.

[13] Ioana Livadariu, Ahmed Elmokashfi, and Georgios Smaragdakis. 2024. Tracking submarine cables in the wild. *Computer Networks* 242 (2024), 110234.

[14] Matthew Luckie, Bradley Huffaker, Alexander Marder, Zachary Bischof, Marianne Fletcher, and KC Claffy. 2021. Learning to extract geographic information from internet router hostnames. In *Proceedings of the 17th International Conference on emerging Networking EXperiments and Technologies*. 440–453.

[15] Alagappan Ramanathan and Sangeetha Abdu Jyothi. 2023. Nautilus: A framework for cross-layer cartography of submarine cables and ip links. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 7, 3 (2023), 1–34.

[16] Hugo Rimlinger, Olivier Fourmaux, Timur Friedman, and Kevin Vermeulen. 2025. GeoResolver: An Accurate, Scalable, and Explainable Geolocation Technique Using DNS Redirection. *Proceedings of the ACM on Networking* 3, CoNEXT3 (2025), 1–21.

[17] Loqman Salamatian, Kevin Vermeulen, Italo Cunha, Vasilis Giotsas, and Ethan Katz-Bassett. 2024. metAScritic: Reframing AS-Level Topology Discovery as a Recommendation System. In *Proceedings of the 2024 ACM on Internet Measurement Conference*. 337–364.

[18] Quirin Scheitle, Oliver Gasser, Patrick Sattler, and Georg Carle. 2017. HLOC: Hints-based geolocation leveraging multiple measurement frameworks. In *2017 Network Traffic Measurement and Analysis Conference (TMA)*. IEEE, 1–9.

[19] Kedar Thiagarajan, Esteban Carisimo, and Fabián E Bustamante. 2025. The Aleph: Decoding Geographic Information from DNS PTR Records Using Large Language Models. *Proceedings of the ACM on Networking* 3, CoNEXT1 (2025), 1–20.

[20] Yong Wang, Daniel Burgener, Marcel Flores, Aleksandar Kuzmanovic, and Cheng Huang. 2011. Towards {Street-Level} {Client-Independent} {IP} Geolocation. In *8th USENIX Symposium on Networked Systems Design and Implementation (NSDI 11)*.

[21] Zachary Weinberg, Shinyoung Cho, Nicolas Christin, Vyas Sekar, and Phillipa Gill. 2018. How to catch when proxies lie: Verifying the physical locations of network proxies with active geolocation. In *Proceedings of the Internet Measurement Conference 2018*. 203–217.

[22] Bernard Wong, Ivan Stoyanov, and Emin Gün Sirer. 2007. Octant: A Comprehensive Framework for the Geolocalization of Internet Hosts.. In *NSDI*, Vol. 7. 23–23.