Systematic Detection and Correction of IP Geolocation Anomalies in Network Measurements

ALAGAPPAN RAMANATHAN, University of California Irvine, USA
SANGEETHA ABDU JYOTHI, University of California Irvine, USA

1 Introduction

Traceroute and IP geolocation are essential tools for Internet infrastructure analysis, supporting applications from network topology mapping to censorship detection. However, when these tools are used together in practice, their individual limitations compound to create systematic inaccuracies that significantly impact measurement reliability.

To understand how these limitations compound, we examined each tool individually. IP geolocation aims to map IP addresses to physical locations, but commercial databases often disagree substantially—past studies show accuracy as low as 33% at the country level for some databases. Simultaneously, traceroute measurements face challenges from network infrastructure elements like MPLS tunnels and interface address variability that can obscure true network paths. When researchers combine these imperfect tools to analyze network topology and performance, the resulting errors propagate throughout their analyses.

Our preliminary analysis of a day-long traceroute measurement from RIPE Atlas demonstrates that approximately 5.4% of IP addresses in this dataset exhibit what we term "anomalous geolocations"—locations that are inconsistent with network measurement data. These anomalies affect at least one hop in 55% of traceroutes and 20% of network links, creating widespread contamination of network analysis results.

2 Current Gaps in IP Geolocation Approaches

2.1 Fundamental Database Inconsistencies

To understand the scope of these anomalies, we must first examine the underlying causes. IP geolocation databases suffer from systematic consistency problems that become apparent when subjected to rigorous validation. We analyzed $\approx$ 250,000 unique IPv4 addresses using eight major geolocation databases and found that traditional validation methods achieve consensus for only 30% of IP addresses. This lack of agreement reflects deeper methodological issues in how databases are constructed and maintained.

2.2 Network Infrastructure Effects on Geolocation Accuracy

Beyond database inconsistencies, the network infrastructure itself introduces several systematic complications. We primarily examined two technical factors that compromise geolocation accuracy when combined with traceroute measurements:

MPLS Tunnel Effects: Multiprotocol Label Switching creates uniform RTTs within tunnels, making all hops within a tunnel appear to have the same RTT as the tunnel exit point. This obscures the true geographic path of packets.

Interface Address Variability: Per RFC 1812, routers respond to traceroute probes with the IP address of the interface used to transmit the ICMP reply, not necessarily the interface where the original packet arrived. This can result in off-path interface reporting, particularly problematic near AS boundaries and country borders where precise location is crucial.

2.3 Scalability Limitations of Current Solutions

Given these widespread problems, one might expect existing solutions to address them systematically. However, current approaches face significant practical limitations. Active measurement methods require additional infrastructure and impose measurement overhead that does not scale to large IP datasets. Most critically, current methods address MPLS or interface issues in isolation, failing to provide systematic solutions for the compounding effects we observe.

Authors' Contact Information: Alagappan Ramanathan, University of California Irvine, Irvine, California, USA; Sangeetha Abdu Jyothi, University of California Irvine, Irvine, California, USA.

3 Technical Approach: Systematic Anomaly Detection and Correction

Recognizing these gaps, we developed a prototype: "GeoTrace", a lightweight tool that leverages existing traceroute data to systematically identify, classify, and potentially correct geolocation anomalies without requiring additional measurements.

3.1 Iterative Neighbor-Based Evaluation

GeoTrace employs an iterative approach that begins by aggregating and clustering geolocations from multiple databases. For each IP and its immediate neighbor in the traceroute dataset, GeoTrace evaluates RTT-distance alignment using Haversine distance calculations with accommodations in place for network variability.

GeoTrace computes performance ratios for each geolocation candidate and retains only candidates within 90% of the best-performing option. This iterative process continues until geolocation candidates stabilize, ensuring convergence to stable and accurate location estimates.

3.2 Classification and Targeted Correction

After identifying anomalous IPs (those with consistently low performance ratios), GeoTrace classifies them based on underlying causes:

MPLS-Affected IPs. : Identified when anchor IPs (accurately geolocated neighbors in traceroutes) are dispersed across multiple countries with no single country accounting for more than 95% of anchors. These are flagged but not corrected due to inherent measurement limitations within MPLS tunnels.

Interface-Affected IPs. : Corrected using constraint-based techniques. GeoTrace constructs buffer regions around anchor IP locations using median RTT differences, then employs spatial indexing with city polygons to identify regions with maximum buffer overlap.

4 Preliminary Analysis

4.1 Quantitative Improvements in Geolocation Accuracy

GeoTrace achieved significant improvements over traditional validation methods by nearly doubling the percentage of IPs with consistent geolocation clusters compared to traditional validation methods. Further, GeoTrace effectively identified, classified, and successfully corrected interface-affected IPs to be in line with observed measurement data.

4.2 Systematic Bias Discovery

More significantly, our analysis revealed systematic patterns suggesting methodological biases in existing geolocation databases. We observed consistent over-assignment of anomalous IP addresses to certain Western European countries and under-representation in others, with substantial correction distances averaging over 1,000 km and nearly one-third of corrections involving country-level discrepancies.

5 Implications for Current Use Cases

5.1 Network Analysis Applications

The prevalence of geolocation errors—affecting over half of all traceroutes—suggests that current approaches to network topology mapping, performance optimization, and infrastructure analysis may be based on systematically flawed location assumptions.

5.2 Regulatory and Security Applications

Location inaccuracies have direct policy implications. In censorship analysis, such inaccuracies might obscure the true path of data through restrictive regions. For regulatory compliance, the nearly one-third rate of country-level discrepancies in anomalous IPs (which are typically concentrated at country and AS borders) suggests that jurisdictional determinations based on IP geolocation may be systematically unreliable.

6 Recommendations for the Community

6.1 Systematic Validation Integration

The community should integrate consistency-based validation into existing measurement platforms. Our preliminary analysis demonstrates that neighbor-based validation can significantly improve accuracy without additional measurement overhead. Measurement platforms should implement feedback mechanisms to geolocation database providers based on systematic error detection.

6.2 Enhanced Database Quality Mechanisms

Geolocation database providers should address the systematic geographic biases we identified, particularly the over-assignment patterns in Western European countries for anomalous IPs. Focus should be placed on improving accuracy near AS boundaries and country borders, where precise location is crucial.

6.3 Lightweight Correction Frameworks

The success of constraint-based approaches using virtual anchor points like GeoTrace demonstrates that effective correction can be achieved without additional active measurements. The community should develop standardized frameworks that leverage existing measurement infrastructure.

7 Conclusion

The systematic nature of IP geolocation inaccuracies demands community attention. Our analysis demonstrates that existing measurement infrastructure contains sufficient information to identify and potentially correct systematic errors when properly analyzed through neighbor-based consistency evaluation and constraint-based correction techniques.

The key insight is that lightweight, data-driven approaches can significantly improve accuracy without additional measurement overhead. The prevalence and impact of these issues underscore the need for systematic IP geolocation validation in network measurement studies. Without addressing these fundamental gaps, critical Internet infrastructure decisions will continue to likely be based on systematically flawed location data.