Improving IP Geolocation With Target-Centric IP Graph (Student Abstract)

Kai Yang[1], Jiayang Li[1], Wenxin Tai[1,2*], Zhenhui Li[1], Ting Zhong[1,2], Guangqiang Yin[1,2], Yong Wang[3]

[1]University of Electronic Science and Technology of China
[2]Kashi Institute of Electronics and Information Industry
[3]Hong Kong University of Science and Technology
kaiyang.cs@outlook.com, {jy.li, wxtai, 202012081717}@std.uestc.edu.cn, {zhongting,yingq}@uestc.edu.cn,
wangyongjoy@ust.hk

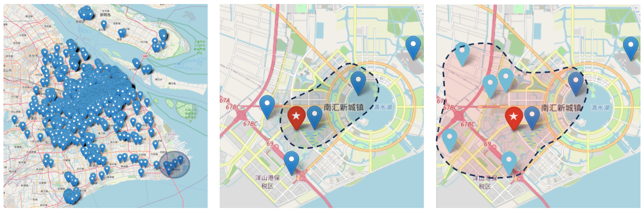
Abstract

Accurate IP geolocation is indispensable for location-aware applications. While recent advances based on router-centric IP graphs are considered cutting-edge, one challenge remain: the prevalence of sparse IP graphs (14.24% with fewer than 10 nodes, 9.73% isolated) limits graph learning. To mitigate this issue, we designate the target host as the central node and aggregate multiple last-hop routers to construct the target-centric IP graph, instead of relying solely on the router with the smallest last-hop latency as in previous works. Experiments on three real-world datasets show that our method significantly improves the geolocation accuracy compared to existing baselines.

Introduction

Numerous client-independent IP geolocation methods have been proposed to estimate geographic locations. These methods do not rely on users willingly sharing their location data but instead employ alternative techniques to determine locations without direct input from clients. Recently, cutting-edge advancements in this field involve the use of graph neural networks (GNNs) with a router-centric IP graph, known for their remarkable ability to harness rich surrounding information (Wang et al. 2022; Tai et al. 2023). Nevertheless, despite the enhanced geolocation accuracy achieved, router-centric IP geolocation faces one challenge: The prevalence of sparse IP graphs limits the efficacy of graph learning. Upon analyzing the datasets, there are 14.24% of the targets possess fewer than 10 neighbors, with 9.73% of them being completely isolated.

In this work, we present a simple but effective graph enrichment method to address the aforementioned challenge. Specifically, we designate the target host as the central node and aggregate multiple last-hop routers to construct the target-centric (TC) IP graph, instead of relying solely on the router with the smallest last-hop latency as in previous works. This straightforward modification substantially reduces the proportion of isolated nodes from 9.73% to 4.46% and lowers the sparsity ratio from 14.24% to 8.64%, which further enhances the geolocation accuracy (cf. Table 1). Experimental results demonstrate that our method can effectively eliminate unnecessary topological neighbors while retaining the most pertinent and critical ones.



(a) Shanghai (b) Router-Centric (c) Target-Centric

Figure 1: From router-centric to target-centric IP graph: (a) router-centric landmark distribution in Shanghai; (b) router-centric IP graph; (c) target-centric IP graph where distinct colors indicate associations with different last-hop routers.

Methodology

Problem Definition. Given a set of landmarks $\{\boldsymbol{l}_i\}_{i=1}^N$ with attribute knowledge $\{\mathbf{x}_i\}_{i=1}^N$ (6-dimensional data extracted from the WHOIS website), network measurements $\{\mathbf{m}_i\}_{i=1}^N$ (24-dimensional ping and traceroute data), and coordinates $\{\mathbf{y}_i\}_{i=1}^N$ (2 dimensions for longitude and latitude), our objective is to predict the geographic location of a target IP:

$$\widehat{\boldsymbol{y}}_T = f(\{\boldsymbol{x}_i\}_{i=1}^N, \{\boldsymbol{m}_i\}_{i=1}^N, \{\boldsymbol{y}_i\}_{i=1}^N, \boldsymbol{x}_T, \boldsymbol{m}_T; \boldsymbol{\Theta}),$$

where $\widehat{\boldsymbol{y}}_T$=$(\widehat{lon}_T, \widehat{lat}_T) \in \mathbb{R}^2$ denotes the estimated locations of the target, and $\boldsymbol{\Theta}$ denotes NN parameters.

Sparsity Investigation. To investigate the influence of neighbor quantity on geolocation performance, we selectively mask nodes in the adjacency matrix to control neighbor counts. Subsequently, we apply the model proposed in (Tai et al. 2023) and evaluate its performance on IP graphs with varying numbers of neighboring nodes. We observe a rapid decrease in distance error as the number of neighbors increases from 1 to 10 (cf. Figure 3(a)). This observation showcases a potential limitation of GNN-based geolocation methods: a constrained amount of neighbor landmarks will significantly influence the geolocation accuracy.

Learning on Target-Centric IP Graphs. Different from previous works (Wang et al. 2022; Tai et al. 2023) that narrow the region by clustering the IPs at the router level, we designate the target host as the central node and aggregate

[*]Corresponding author

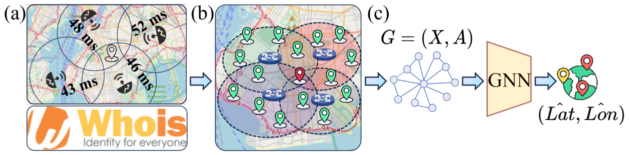


Figure 2: Overview of the TCGeo architecture: (a) Processing data collection and feature engineering; (b) Constructing target-centric IP graph; (c) Using GNN for geolocation.

multiple last-hop routers to construct the target-centric IP graph. Figure 1 is a target-centric example. Note that these router statistics (and their connected landmarks) are acquired from traceroute data gathered by four strategically positioned probing hosts spanning various regions.

For each IP graph $G = (\boldsymbol{X}, \boldsymbol{A})$, we define node features as a composite of attribute knowledge, network measurements, and coordinates, totaling 32 dimensions. As for edge weights, we employ an attention mechanism to dynamically learn interactions between landmarks and target nodes, following (Tai et al. 2023):

$$\boldsymbol{A}_{T,l} = \exp\left(\mathbf{v}^T \sigma\left(\mathbf{W}_1\{\mathbf{x}_T, \mathbf{m}_T\} + \mathbf{W}_2\{\mathbf{x}_l, \mathbf{m}_l\} + \mathbf{b}\right)\right),$$

where $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{(d_x+d_m)\times(d_x+d_m)}$, and $\mathbf{b}, \mathbf{v} \in \mathbb{R}^{d_x+d_m}$ are trainable matrices and vectors. Subsequently, we apply one GNN layer to facilitate the learning of representations for the target IP address. Once the final representation is obtained, we utilize a non-linear layer to estimate the geographic location of the target IP. We treat IP geolocation as a deterministic regression task and optimize the model by minimizing the mean squared error (MSE) between the estimated location $\widehat{\mathbf{y}}_T$ and the ground truth $\mathbf{y}_T$. Figure 2 shows the detail of TCGeo.

Experiments

Datasets and Setup. Following previous works (Wang et al. 2022; Tai et al. 2023), we evaluate our method on three real-world IP geolocation datasets (New York, Los Angeles and Shanghai), which consist of 91,808, 92,804, and 126,258 IP addresses respectively. We take 70% IP as landmarks and 30% as target IPs in training process. During testing, we treat the training set as landmarks and others as target IPs. We set the learning rate as 0.002 for the New York, Los Angeles datasets, and 0.001 for the Shanghai dataset. The hidden size of each layer (except the last layer) is fixed to 32.

Baselines. We compared our method with the following state-of-the-art baselines, including one *delay-based* measurement method (Wang et al. 2020), one *attribute learning* method (Arik and Pfister 2021), and three *graph learning* methods (Ding et al. 2022; Wang et al. 2022; Tai et al. 2023).

Sparsity Analysis. We explored the influence of different numbers of neighbors on geolocation performance. As Figure 3(a) shown, the TCGeo's performance has no significant improvement when the number of neighbors exceeds 10. Therefore, we only use the target-centric IP graph for targets with fewer than 10 neighbors, and for the rest, we stick with the router-centric IP graphs. This strategy can significantly improve the geolocation performance while maintaining the computational efficiency. In Figure 3(b), we demonstrate the extent to which TCGeo alleviates the issue of graph sparsity across three datasets.

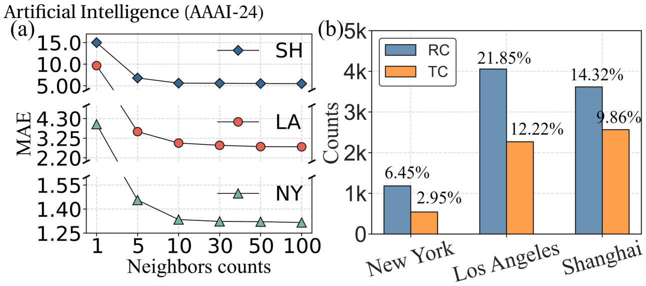


Figure 3: Sparsity investigation. (a) Influence of the number of neighbors; (2) Sparsity reduction from router-centric graph to target-centric graph.

Method	New York		Los Angeles		Shanghai	
	MAE	Median	MAE	Median	MAE	Median
XLBoost-Geo	2.179	1.572	4.577	4.129	6.850	5.242
TabNet	3.272	3.198	6.262	5.189	6.722	5.012
GNN-Geo	2.135	1.618	4.655	4.039	6.026	4.482
GraphGeo	1.614	1.118	3.778	2.269	5.981	3.982
TrustGeo	1.316	0.888	2.793	1.786	5.457	3.619
TCGeo	**1.271**	**0.866**	**2.658**	**1.745**	**5.189**	**3.461**

Table 1: Performance comparisons with recent baselines. All results are measured in kilometers (km).

Overall Performance. The results of our comparative evaluation experiments on three datasets are summarized in Table 1, unveiling two significant findings: (1) Graph-based IP geolocation methods outperform other methods, underscoring the importance of harnessing contextual information. (2) Under the same NN architecture design, our method outperforms TrustGeo, in alignment with our initial motivation that integrating more neighbors can effectively mitigate the issue of graph sparsity.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (Grant No.62176043 and No.62072077).

References

Arik, S. Ö.; and Pfister, T. 2021. TabNet: Attentive Interpretable Tabular Learning. In *AAAI*, 6679–6687.

Ding, S.; Zhang, F.; Luo, X.; and Liu, F. 2022. GNN-Geo: A Graph Neural Network-based Fine-grained IP Geolocation Framework. *arXiv preprint arXiv:2112.10767*.

Tai, W.; Chen, B.; Zhou, F.; Zhong, T.; Trajcevski, G.; Wang, Y.; and Chen, K. 2023. TrustGeo: Uncertainty-Aware Dynamic Graph Learning for Trustworthy IP Geolocation. In *SIGKDD*.

Wang, Y.; Zhu, H.; Wang, J.; Liu, J.; Wang, Y.; and Sun, L. 2020. XLBoost-Geo: An IP Geolocation System Based on Extreme Landmark Boosting. *arXiv preprint arXiv:2010.13396*.

Wang, Z.; Zhou, F.; Zeng, W.; Trajcevski, G.; Chunjing, X.; Yong, W.; and Kai, C. 2022. Connecting the Hosts: Street-Level IP Geolocation with Graph Neural Networks. In *SIGKDD*.