

EtherBee: A Global Dataset of Ethereum Node Performance Measurements Coupled with Honeypot Interactions and Full Network Sessions

Scott Seidenberger and Anindya Maiti
University of Oklahoma
seidenberger@ou.edu, am@ou.edu

Abstract—We introduce *EtherBee*, a global dataset integrating detailed Ethereum node metrics, network traffic metadata, and honeypot interaction logs collected from ten geographically diverse vantage points over three months. By correlating node data with granular network sessions and security events, *EtherBee* provides unique insights into benign and malicious activity, node stability, and network-level threats in the Ethereum peer-to-peer network. A case study shows how client-based optimizations can unintentionally concentrate the network geographically, impacting resilience and censorship resistance. We publicly release *EtherBee* to promote further investigations into performance, reliability, and security in decentralized networks.

I. INTRODUCTION AND RELATED WORK

Web3 networks are becoming critical public infrastructure, and Ethereum is among the most widely used and robust public smart contract platforms. As the network grows in size and complexity, understanding the performance, behavior, and risk characteristics of Ethereum nodes is a key component to ensuring the blockchain's security, reliability, and scalability. Nodes are the core components that validate transactions, execute smart contracts, and maintain global consensus; however, the decentralized nature of blockchain networks complicates holistic data collection and analysis of these nodes. Existing datasets often fail to capture the full breadth of node activity when striving for global coverage, granular metrics, and long observation periods [1], [2], [3].

Recent work has examined Ethereum node performance and decentralization, showing that the effectiveness of the network can shift over time due to major events like client software bugs, and that a small number of large entities often control most validators [4], [5]. Furthermore, economic incentives may unintentionally promote centralization [6]. Other studies focus on measuring the Ethereum networking stack through active monitoring [7], [8], [9], yet most publicly available datasets prioritize on-chain activity rather than node or security-focused data. Meanwhile, multimodal cyber threat intelligence has proven essential to uncovering sophisticated attacks, especially when analyzing logs, network telemetry, and threat feeds in tandem [10], [11].

To address these gaps, we present *EtherBee*, a first-of-its-kind global dataset that integrates detailed Ethereum node metrics, network traffic metadata, and honeypot interaction logs. By capturing both benign and malicious activity across diverse

geographic vantage points, *EtherBee* facilitates a deeper understanding of peer-to-peer (P2P) communication, node stability, and network-level threats. To the best of our knowledge, no existing resource provides this level of multimodal detail that covers both operational node data and honeypot collections. In summary, our contributions are: (1) *EtherBee*, a dataset with fine-grained metrics from 10 vantage points over three months; (2) a scalable methodology for collecting, storing, and transforming the data; and (3) analyses that generate new research questions around performance, reliability, and security for decentralized networks.

II. DATA COLLECTION

We deployed ten AWS servers across five global regions (North America, Europe, Middle East, Asia, and South America). In each region, one server ran Lighthouse [12] (Consensus Layer), Nethermind [13] (Execution Layer), plus honeypot and packet capture software; a second server ran only honeypot and packet capture. All three data streams—network sessions, node metrics, and honeypot logs—were indexed into a six-node Elasticsearch cluster (227 TB storage, 132 GB heap), yielding 30.1 TB of primary data across 1,419 shards and 33.7 billion documents. We used a WireGuard-based mesh to encrypt inter-server traffic, and afterwards exported raw Elasticsearch data to tabular formats for public release. This final, time-synchronized dataset enables correlating network-level events, node-level performance, and cyberattacks at global scale.

A. Network Traffic Capture

Using Arkime [14], we indexed start/end times, IP addresses (with geolocation), ports, and byte counts for every connection. Over three months (Aug–Nov 2024), 94,659 unique IPs accessed our P2P ports, yielding 20.97 TB of metadata. Since Ethereum mainnet typically has 6–10k persistent beacon nodes [15], our dataset encompasses both stable peers and transient endpoints.

B. Node Metrics and Honeypot Logs

We instrumented Lighthouse (including its attestation simulator¹) and Nethermind to capture chain health (slots since best block, attestors, aggregator participation), node performance (latencies, errors, reorganizations), and peer/network stability

¹<https://blog.sigmaprime.io/attestation-sim.html>

(disconnects, discovery events). Simultaneously, a diverse honeypot suite [16] logged 133.8 million inbound attacks, from port scans to CVE-specific exploits. Of these, 58% were DDoS-related, closely matching Cloudflare Radar’s 57% [17]. Collecting these security events alongside blockchain metrics offers a unique multimodal lens on both benign and hostile interactions targeting Ethereum nodes.

III. CASE STUDY

A. Motivation & Setup

Peer Pruning and Geographic Centralization. In latency-sensitive networks like Ethereum, client software scores peers based on metrics such as RTT [18], [19], actively pruning those with poorer scores. While this can enhance propagation efficiency, existing research suggests it may inadvertently favor geographically closer peers, reducing overall decentralization [20], [9]. To evaluate this, we analyzed three months of P2P sessions from *EtherBee*, mapping peer IPs to geographic coordinates (via MaxMind GeoLite2) and computing a data-weighted “focal point.” Each peer’s weight corresponded to its total bytes exchanged, reflecting its relative influence on the node’s connectivity.

Weighted Geographic Focal Point Calculation. For N peers with latitude ϕ_i , longitude λ_i , and weight w_i , we convert each location to 3D Cartesian coordinates on the unit sphere:

$$x_i = \cos(\phi_i) \cos(\lambda_i), \quad y_i = \cos(\phi_i) \sin(\lambda_i), \quad z_i = \sin(\phi_i).$$

We then compute:

$$X = \frac{\sum w_i x_i}{\sum w_i}, \quad Y = \frac{\sum w_i y_i}{\sum w_i}, \quad Z = \frac{\sum w_i z_i}{\sum w_i},$$

and finally:

$$\phi_c = \arctan 2(Z, \sqrt{X^2 + Y^2}), \quad \lambda_c = \arctan 2(Y, X).$$

The resulting (ϕ_c, λ_c) indicates the daily spatial centroid of peer connectivity.

B. Key Findings

Four of the five vantage points showed a significant shift toward lower-latency peers. For example, the North American node’s average peer distance shrank by 6.17 km/day ($p = 0.002$), indicating intensified geographic clustering (Figure 1). The South American node did not exhibit local centralization ($p = 0.797$), but re-centering it as if it were in North America produced a strong clustering effect (-4.16 km/day, $p < 0.001$), reflecting sparse local peers and dependence peer connections in NA, afforded by undersea cables connecting the regions. Daily weighted focal points converged along NA–EU submarine routes (Figure 2), underscoring reliance on limited physical links. Meanwhile, nodes in remote regions were pruned more often, reinforcing existing hubs and reducing regional connectivity.

These results support the hypothesis that **pruning based on latency unintentionally drives geographic centralization**, raising resilience and censorship concerns. Reliance on a handful of undersea cable routes concentrates risk: any

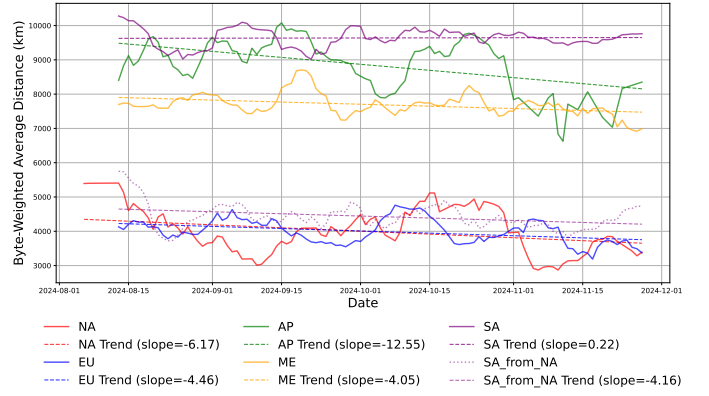


Fig. 1: Daily byte-weighted average distance to peers for each vantage point over 107 days, with linear regression and 7-day smoothing.

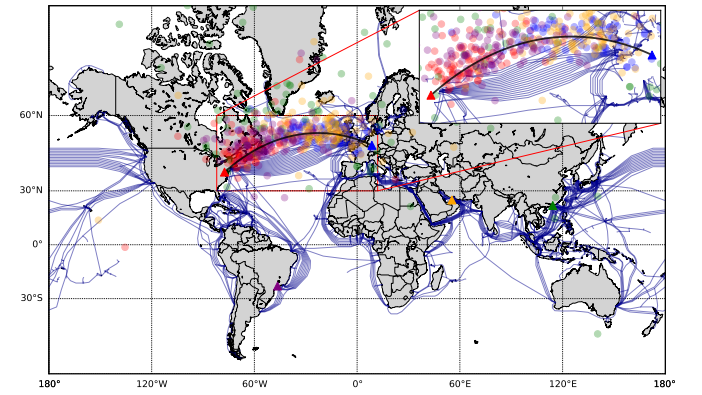


Fig. 2: Data-weighted geographic focal points of five Ethereum nodes over 107 days. The black line is the geodesic path between N. Virginia and Frankfurt; blue lines show key undersea cables, highlighting convergence along critical routes.

disruption along these paths can disproportionately affect large swaths of the network. Balancing performance with broader peer diversity will require refined scoring mechanisms that look beyond latency alone, ultimately safeguarding Ethereum’s decentralization at the P2P layer.

IV. CONCLUSION AND DATA RELEASE

We introduced *EtherBee*, a global dataset combining Ethereum node metrics, honeypot data, and full network traffic sessions from ten vantage points over three months, enabling holistic analyses of node performance, P2P topology, and security threats. Our findings show how client-based networking decisions can inadvertently drive geographic centralization, affecting resilience and censorship resistance. The complete ~ 16 TB compressed dataset is available upon request and other access options are available on OSF at <https://doi.org/10.17605/OSF.IO/C5UPF>.

V. ACKNOWLEDGMENTS

The funding for this dataset was provided by the Ethereum Foundation under grants FY24-1546 and FY24-1547.

REFERENCES

- [1] A. Khan, "Graph analysis of the ethereum blockchain data: A survey of datasets, methods, and future work," in *2022 IEEE International Conference on Blockchain (Blockchain)*, 2022, pp. 250–257.
- [2] Q. Wang, Z. Zhang, Z. Liu, S. Lu, B. Luo, and B. He, "EX-graph: A pioneering dataset bridging ethereum and x," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: <https://openreview.net/forum?id=juE0rWGCJW>
- [3] T. Chen, Z. Li, Y. Zhang, X. Luo, A. Chen, K. Yang, B. Hu, T. Zhu, S. Deng, T. Hu *et al.*, "Dataether: Data exploration framework for ethereum," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, pp. 1369–1380.
- [4] D. Grandjean, L. Heimbach, and R. Wattenhofer, "Ethereum proof-of-stake consensus layer: Participation and decentralization," 2023. [Online]. Available: <https://arxiv.org/abs/2306.10777>
- [5] M. Cortes-Goicoechea and L. Bautista-Gomez, "Discovering the ethereum2 p2p network," in *2021 Third International Conference on Blockchain Computing and Applications (BCCA)*. IEEE, 2021, pp. 81–88.
- [6] M. Cortes-Goicoechea, T. Mohandas-Daryanani, J. L. Muñoz-Tapia, and L. Bautista-Gomez, "Unveiling ethereum's hidden centralization incentives: Does connectivity impact performance?" in *2023 Fifth International Conference on Blockchain Computing and Applications (BCCA)*, 2023, pp. 89–96.
- [7] S. K. Kim, Z. Ma, S. Murali, J. Mason, A. Miller, and M. Bailey, "Measuring ethereum network peers," in *Proceedings of the Internet Measurement Conference 2018*, ser. IMC '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 91–104. [Online]. Available: <https://doi.org/10.1145/3278532.3278542>
- [8] T. Wang, C. Zhao, Q. Yang, S. Zhang, and S. C. Liew, "Ethna: Analyzing the underlying peer-to-peer network of ethereum blockchain," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 3, pp. 2131–2146, 2021.
- [9] A. E. Gencer, S. Basu, I. Eyal, R. Van Renesse, and E. G. Sirer, "Decentralization in bitcoin and ethereum networks," in *Financial Cryptography and Data Security: 22nd International Conference, FC 2018, Nieuwpoort, Curaçao, February 26–March 2, 2018, Revised Selected Papers 22*. Springer, 2018, pp. 439–457.
- [10] H. Almohannadi, I. Awan, J. Al Hamar, A. Cullen, J. P. Disso, and L. Armitage, "Cyber threat intelligence from honeypot data using elasticsearch," in *2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA)*. IEEE, 2018, pp. 900–906.
- [11] A. Mahboubi, K. Luong, H. Aboutorab, H. T. Bui, G. Jarrad, M. Bahutair, S. Camtepe, G. Pogrebna, E. Ahmed, B. Barry *et al.*, "Evolving techniques in cyber threat hunting: A systematic review," *Journal of Network and Computer Applications*, p. 104004, 2024.
- [12] Sigma Prime, "Lighthouse: An ethereum 2.0 client," <https://github.com/sigp/lighthouse>, 2024, version v5.2.1-9e12c21.
- [13] Nethermind Project, "Nethermind: The ethereum execution client," <https://github.com/NethermindEth/nethermind>, 2024, version 1.27.1+d2d5d4ae.
- [14] A. Project, "Arkime: Open source full packet capture and analysis tool," 2024, accessed: 2024-12-22. [Online]. Available: <https://github.com/arkime/arkime>
- [15] MigaLabs, "Ethereum cl layer public dashboard," <https://monitoreth.io/nodes>, 2024, accessed: 16 December 2024. Data retrieved from Armirma by MigaLabs.
- [16] T-Pot Project, "T-pot: The all in one honeypot platform," <https://github.com/telekom-security/tpotce>, 2024, version 24.04.0.
- [17] Cloudflare, "Cloudflare radar," <https://radar.cloudflare.com/>, 2024, accessed: 16 December 2024.
- [18] O. Krajsa and L. Fojtova, "Rtt measurement and its dependence on the real geographical distance," in *2011 34th International Conference on Telecommunications and Signal Processing (TSP)*. IEEE, 2011, pp. 231–234.
- [19] Z. Hu, J. Heidemann, and Y. Pradkin, "Towards geolocation of millions of ip addresses," in *Proceedings of the 2012 Internet Measurement Conference*, 2012, pp. 123–130.
- [20] Y. Mao, S. Deb, S. B. Venkatakrishnan, S. Kannan, and K. Srinivasan, "Perigee: Efficient peer-to-peer network design for blockchains," in *Proceedings of the 39th Symposium on Principles of Distributed Computing*, 2020, pp. 428–437.