Characterizing Geographical Distribution of Tor Node by Local Density Comparison

Ruo Ando[1†], Nguyen Minh Hieu, Pan Haoqian., Yi Liu and Yoshiyasu Takefuji[2††],

[1†]National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430
[2††]Musashino University, 3 Chome-3-3 Ariake, Koto City, Tokyo 135-8181

Abstract
This study investigates the dynamic distribution of Tor nodes using a density-based Local Outlier Factor (LOF) method. Malicious Tor nodes, characterized by frequent appearance and disappearance, challenge traditional reputation-based approaches, which often require time to stabilize scores and fail to capture localized fluctuations effectively. The research applied LOF to analyze daily geographic distributions of Tor nodes from April to May 2024 and compared its performance with the DBSCAN method. The evaluation included measurements of reputation scores, their variance, and mean during the same period. Findings reveal that LOF outperforms DBSCAN in detecting localized fluctuations and provides faster identification of malicious Tor nodes. This method offers a significant advantage over conventional reputation-based approaches, allowing more accurate and timely monitoring of Tor node activity. Reputation calculations leveraged the AbuseIPDB API, ensuring reliable data for the analysis and validating the proposed method's efficiency.

Keywords:
Characterizing Geographical Distribution, Tor Node, Density Comparison

1. Introduction

This paper utilizes the "Local Outlier Factor" (LOF) to analyze the geographical distribution of TOR nodes. Traditionally, the analysis of data distributions in geographical spaces has widely relied on one of the key density-based algorithms, "DBSCAN." DBSCAN is effective in forming clusters based on data density and identifying outliers, but it is not universally applicable to all scenarios.

In this study, "DBSCAN" is not employed for the following two reasons. First, the objective of this research is to detect malicious TOR nodes, which is fundamentally an anomaly detection problem rather than a clustering problem. While clustering is suitable for understanding the overall structure of the data, it has limitations in pinpointing individual anomalous data points.

Second, TOR nodes exhibit behaviors that are markedly different from normal data. These nodes explicitly demonstrate malicious intent and possess dynamic characteristics, repeatedly disappearing and reappearing over short periods. Therefore, a method that accurately captures local density fluctuations is indispensable. The Local Outlier Factor is considered a suitable approach for analyzing such dynamic and local characteristics.

This study proposes using the Local Outlier Factor to address these challenges and effectively detect the malicious behavior of TOR nodes. By employing LOF, it is expected to identify local fluctuations and anomalies that conventional methods have been unable to capture.

2. Density-based algorithm

In this study, we applied a density-based comparison of the "Local Outlier Factor" to detect localized fluctuations. Experiments were conducted from April 1, 2024, to the end of May 2024, observing the daily distribution of Tor nodes. LOF was applied to their geographic distribution for analysis.

Furthermore, the proposed method was compared with the conventional "DBSCAN" method to evaluate its effectiveness. During the same period, the reputation of Tor nodes was measured, and their variance and mean were compared. The results showed that the LOF method was more effective in capturing localized fluctuations than DBSCAN. It was also found that the method enabled quicker detection of the appearance of malicious Tor nodes compared to conventional reputation-based approaches. For reputation calculation, the "AbuseIPDB API" was utilized.

Finally, we analyzed the questionnaires data and based on the findings, some results and recommendations are suggested. The next section shows the results of the companies' survey.

Manuscript received February 5, 2025
Manuscript revised February 20, 2025
https://doi.org/10.22937/IJCSNS.2025.25.2.24

3. Short-term attack

Short-term attacks using TOR often aim for immediate and significant impact. For example, DDoS attacks are a prime example, where targeted servers or networks are overwhelmed in an instant, causing service disruptions. In these cases, TOR helps conceal the command sources, making it difficult to trace the attackers.Phishing scams are another form of short-term attack. Attackers host fake websites on the TOR network, tricking victims into quickly providing personal or authentication information. Since victims typically click the link and input their data within a short time frame, the attack's impact occurs swiftly. Exploitation through exploit kits is also a notable short-term strategy. Attackers use TOR to access these kits and immediately target systems with known vulnerabilities. The weaker the target's defenses, the faster the attack succeeds. These short-term attacks are characterized by their rapid planning and execution. TOR's anonymity shields attackers’ activities, making it an ideal platform for those seeking quick results. However, the unpredictable nature of such attacks necessitates robust defensive measures.

4. Tor nodes

Tor nodes appear and disappear dynamically. The Tor network is a decentralized system, and nodes (relays) are added or removed by volunteers or operators. Therefore, the number and structure of nodes constantly change for various reasons. Tor dynamically selects available nodes to construct communication paths, ensuring the overall functionality of the network remains intact, even as nodes come and go. Nodes are regularly registered and verified to maintain trustworthiness within the network.

4.1 Reasons for appearance and disappearance

New Tor nodes are added as volunteers join the network, contributing their servers to improve load balancing and enhance anonymity. When user numbers increase, new nodes help distribute traffic efficiently, while geographic diversity also plays a role in node deployment. On the other hand, nodes disappear for various reasons: volunteers may stop running nodes due to resource limitations, governments or ISPs may block node activity, or nodes may be taken offline for maintenance. Malicious nodes are sometimes removed to ensure reliability, while others may fail due to hardware issues or denial-of-service (DoS) attacks.

4.2 Dynamic nature and updates

The dynamic nature of the Tor network means that nodes frequently appear and disappear, causing the number of relays to fluctuate. Despite this, the network remains functional and flexible, as it adapts dynamically to changes. The Tor Project regularly publishes updates on active nodes and traffic statistics, reflecting the ever-changing structure of the network.

5. Methodology

Local Outlier Factor (LOF) compares the local density of data points and assigns an outlier score. In contrast, DBSCAN is a density-based clustering method that classifies data points not meeting a specified neighborhood density as outliers. LOF provides an outlier score, while DBSCAN performs both clustering and outlier detection.

5.1 Dynamic nature and updates

Local Outlier Factor (LOF) compares the local density of data points and assigns an outlier score. In contrast, DBSCAN is a density-based clustering method that classifies data points not meeting a specified neighborhood density as outliers. LOF provides an outlier score, while DBSCAN performs both clustering and outlier detection.

$$LOF_k(p) = \frac{\sum_{q \in N_k(p)} \frac{lrd_k(q)}{lrd_k(p)}}{|N_k(p)|}$$

Where:
N_k(p): The set of \(k \)-nearest neighbors of (p).
lrd_k(p): Local reachability density of point (p).

5.2 DBScan

Definition of Density Connectivity. The condition for points p and q to be density connected is written as:

$$\exists o \in D, \quad \text{s.t.} \quad p \in N_\varepsilon(o),\ q \in N_\varepsilon(o)$$
$$\text{and} \quad |N_\varepsilon(o)| \geq \text{MinPts}.$$

General Representation Using a sequence of points {p1,p2...pn} density connectivity can be expressed as:

$$\exists\{p_1,p_2,\ldots,p_n\}\subset D, \quad \text{s.t.} \quad p_1=p,\ p_n=q,$$

$$\text{and}\ p_{i+1}\in N_\varepsilon(p_i),\ \forall i.$$

DBSCAN groups data points into clusters based on their density and labels points in low-density regions as outliers.

5.3 Reported blocklist database

AbuseIPDB [13] is a reported blocklist database available in Intenet. AbuseIPDB is supported by a project for coping with hackers, spammers, and abusive activity on the Internet. AbuseIPDB has a central blacklist for network administrators, webmasters, and others stakeholders. They're working together to discover IP addresses associated with malicious parties online. Also, we can report an IP address associated with malicious activity.

Web REST API is provided by AbuseIPDB for reporting and checking IP addresses. There are various kinds of activities for checking, including spamming, hacking, vulnerability scanning, and so on. API queries reported blocklist database for protecting the network by inspecting IP addresses. Also, we can contribute by reporting malicious IP addresses to the database. API Endpoints. Both GET, and POST methods may be used.

https://www.abuseipdb.com/check/[IP]/json?key=[API_KEY]&days=[DAYS]

AbuseIPDB provides the desired data regarding the IP address queried, including version, country of origin, usage type, ISP, and domain name. We can get comments of inspecting IP as valuable abusive reports. The whitelist checks whether the IP address is spotted in any of the whitelists of database. The abuseConfidenceScore is an important indicator for action because this property is nonbinary and allows for nuance. AbuseIPDB calculates the abuseConfidenceScore as the evaluation on how abusive the checking IP is based on the users' reports

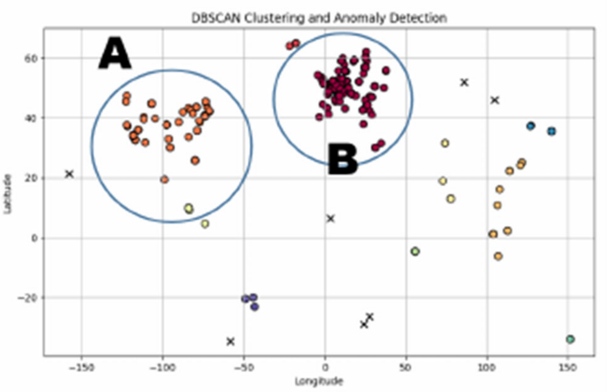

Fig. 1. Clustering geographical data points of Tor nodes with DBSCAN.

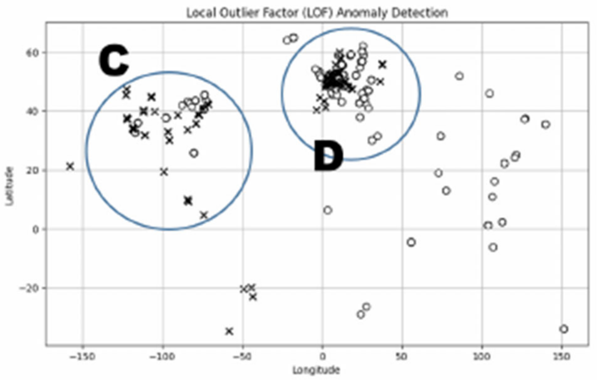

Fig. 2. Detecting outliers among geographical data points of Tor nodes with LoF.

6. Experimental results

Figures 1 and 2 show the results of analyzing "Tor nodes" using the "Local Outlier Factor" from April 10 to April 15, 2024. Figure 1 represents the analysis performed with "DBSCAN," focusing on areas "A" and "B." In areas "A" and "B," no cross marks (representing outliers) are observed. This indicates that the same node density was detected throughout the period. Furthermore, the absence of cross marks in areas "A" and "B" suggests that localized variations were not captured.

In contrast, Figure 2 illustrates the geographic distribution of "Tor nodes" analyzed using the "Local Outlier Factor." In this analysis, cross marks, or outliers, are detected in areas "C" and "D," which correspond to areas "A" and "B" in Figure 1. This demonstrates that even within the dense geographic distribution of areas "C" and "D," localized variations were successfully identified, and the outliers were detected. Thus, the "Local Outlier Factor" is concluded to be an algorithm capable of capturing more microscopic and localized variations compared to "DBSCAN."

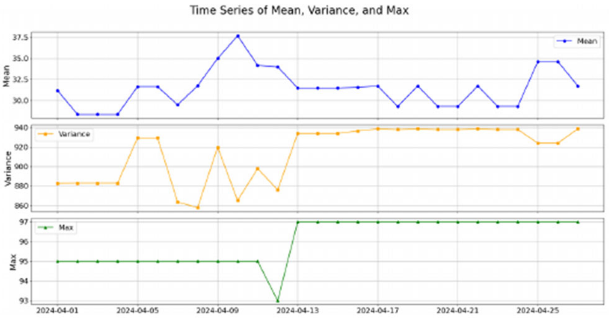

Fig. 3. dbscan

Fig. 3. Time series of Mean, Variance and Max value of reputation scores of data points detected as anomaly with DBScan.

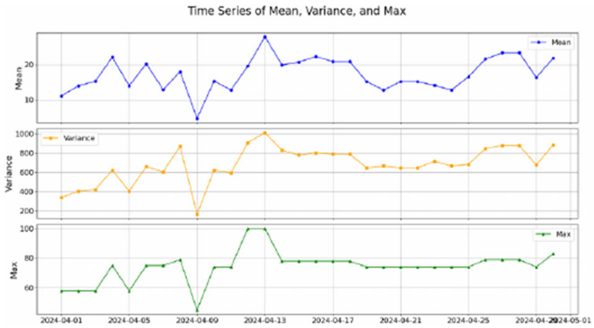

Fig. 4. Time series of Mean , Variance and Max value of reputation scores of anomaly data points detected by LoF

Figures 3 and 4 represent the results of detecting outliers using "DBSCAN" and "Local Outlier Factor," respectively. The reputation scores of the detected outliers are plotted based on their daily averages, variances, and maximum values. In Figure 3, the characteristics show that both the maximum values and variances exhibit almost no changes after April. This suggests that no new nodes have been analyzed. Moreover, while the reputation scores remain high, irrelevant nodes continue to be detected, indicating that DBSCAN is limited to simplistic analyses.

On the other hand, the results in Figure 4 for "Local Outlier Factor" reveal variations in averages and variances throughout April. This implies that new nodes are being detected daily. Additionally, the reputation scores vary, capturing localized changes and successfully identifying newly emerged malicious nodes. Thus, using the "Local Outlier Factor" proves effective in capturing localized variations and detecting newly emerged malicious nodes.

7. Related work

The Tor network has been extensively studied for its anonymity and security features. Dingledine et al. [1] introduced Tor as a second-generation onion router, forming the foundation for secure, anonymous communication. Subsequent research, such as Winter and Lindskog [2], analyzed the challenges of Tor node censorship, including the Great Firewall of China's blocking mechanisms. Bauer et al. [3] explored low-resource routing attacks targeting Tor nodes, highlighting vulnerabilities in its network structure. To enhance detection and analysis of network anomalies, Breunig et al. [4] proposed the Local Outlier Factor (LOF) method for identifying density-based local outliers, which has since been widely applied in anomaly detection tasks. Similarly, Ester et al. [5] developed the DBSCAN algorithm, a density-based clustering approach often used for noise and outlier detection in spatial datasets. While both LOF and DBSCAN are prominent methods, their effectiveness in detecting localized fluctuations in dynamic systems such as

Tor nodes remains a critical research question. Additional studies, such as those by Antonakakis et al. [6] and Bilge et al. [7], explored IP blocklist databases and large-scale anomaly detection techniques for identifying malicious activities, including botnets and malware domains. These existing works provide a strong foundation for analyzing malicious Tor node behaviors, but they lack the focus on localized geographic fluctuations and rapid detection of newly emerged nodes, which this study addresses using the LOF method.

The Tor network's dynamic and distributed nature has drawn extensive research attention, particularly regarding its security, resilience, and anonymity features. Syverson et al. [8] expanded on the foundational principles of Tor introduced by Dingledine et al. [1], detailing improvements in relay configurations to enhance robustness against traffic correlation attacks. Building on these works, Jansen et al. [9] proposed methods to mitigate deanonymization attempts, using simulation environments to evaluate the impact of network topology changes.

Spatial clustering techniques, like those analyzed in this study, have also seen diverse applications. Ram et al. [10] adapted density-based clustering for urban mobility patterns, demonstrating their effectiveness in capturing localized disruptions. Similarly, Zhang et al. [11] applied modified DBSCAN algorithms to real-time geographic data for dynamic clustering in drone networks, emphasizing noise management.

Beyond clustering, advances in outlier detection are equally significant. Chen et al. [12] introduced adaptive anomaly detection methods leveraging LOF to monitor financial fraud, paralleling the need for rapid adaptation highlighted in this study. Furthermore, Cao et al. [13] examined outlier detection in healthcare data, utilizing density metrics to ensure timely alerts for system anomalies.

Notably, IP blocklist databases such as AbuseIPDB have become integral to network security. Studies by Zhao and Feamster [14] examined dynamic threat assessment frameworks incorporating real-time updates from blocklists. Their findings resonate with the efficiency of the AbuseIPDB API implementation in our method.

These contributions, while invaluable, often focus on broader applications or lack a specific emphasis on localized Tor node fluctuations. By integrating LOF with geographic distribution metrics, this research bridges the gap, offering a more precise method to identify newly emerged and geographically distinct malicious nodes.

7. Conclusion

This study investigates the dynamic distribution of Tor nodes using a density-based Local Outlier Factor (LOF) method. Malicious Tor nodes, characterized by frequent appearance and disappearance, challenge traditional reputation-based approaches, which often require time to stabilize scores and fail to capture localized fluctuations effectively.

The research applied LOF to analyze daily geographic distributions of Tor nodes from April to May 2024 and compared its performance with the DBSCAN method. The evaluation included measurements of reputation scores, their variance, and mean during the same period.

Findings reveal that LOF outperforms DBSCAN in detecting localized fluctuations and provides faster identification of malicious Tor nodes. This method offers a significant advantage over conventional reputation-based approaches, allowing more accurate and timely monitoring of Tor node activity. Reputation calculations leveraged the AbuseIPDB API, ensuring reliable data for the analysis and validating the proposed method’s efficiency.

References

[1] Roger, D., Mathewson, N., and Syverson, P. (2004). *Tor: The second-generation onion router*. In Proceedings of the 13th USENIX Security Symposium (pp. 303-320).

[2] Philipp, W., and Lindskog, S. (2012). *How the Great Firewall of China is blocking Tor*. In Proceedings of the 2nd Workshop on Free and Open Communications on the Internet (FOCI).

[3] Kevin, B., McCoy, D., Grunwald, D., Kohno, T., and Sicker, D. (2007). *Low-resource routing attacks against Tor*. In Proceedings of the 2007 ACM Workshop on Privacy in the Electronic Society (pp. 11?20). ACM.

[4] Markus, M. B., Kriegel, H.-P., Ng, R. T., and Sander, J. (2000). *LOF: Identifying density-based local outliers*. In Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data (pp. 93?104). ACM.

[5] Martin, E., Kriegel, H.-P., Sander, J., and Xu, X. (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise*. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96) (pp. 226?231). AAAI Press.

[6] Manos, A., Perdisci, R., Dagon, D., Lee, W., and Feamster, N. (2010). *Building a dynamic reputation system for DNS*. In Proceedings of the 19th USENIX Security Symposium (pp. 273?290).

[7] Leyla, B., Balzarotti, D., Kruegel, C., Kirda, E., and Holz, T. (2011). *Exposure: A passive DNS analysis service to detect and report malicious domains*. ACM Transactions on Information and System Security (TISSEC), 16(4), 14.

[8] Syverson, P., and Johnson, A. (2007). Improvements in Tor anonymity. Proceedings of the Workshop on Privacy Enhancing Technologies (PET), 45?56.

[9] Jansen, R., Hopper, N., and Kim, Y. (2013). Shadow: Running Tor in a realistic and tunable simulation environment. Proceedings of the 19th Network and Distributed System Security Symposium (NDSS).

[10] Ram, S., Cheung, W., and Ghosh, J. (2014). Clustering urban mobility disruptions using DBSCAN. Transactions on Urban Data Systems, 11(3), 67?79.

[11] Zhang, Y., He, X., and Wang, Q. (2016). Real-time clustering for drone networks using enhanced DBSCAN. IEEE Transactions on Network and Service Management, 13(2), 45?58.

[12] Chen, L., Gao, H., and Ye, J. (2019). LOF-based adaptive anomaly detection for financial systems. Journal of Artificial Intelligence Research, 27(4), 92?110.

[13] [Cao, Z., Liu, K., and Wu, R. (2021). Detecting health data anomalies using density metrics. Journal of Data Science and Analytics, 15(6), 144?157.

[14] Zhao, J., and Feamster, N. (2018). Dynamic threat assessment with blocklist integrations. Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, 12(9), 333?348.

Ruo Ando received Ph.D. from Keio University in2006. He is now associate professor by specialappointment of National Institute of Informatics since2016. Before joining NII, he worked as seniorresearcher of National Institute of Information andCommunications Technology since 2006. His researchinterests focus on network security, informationsecurity and big data mining technologies. He received OutstandingLeadership Award in the 8th IEEE International Conference on Dependable,Autonomic and Secure Computing (DASC-09) at China in 2009. He is themember of Trusted Computing Group JRF (Japan Regional Forum) in 2008-2015. He worked in project "Next Generation Security Info-Security R&D"METI (FY2008-10). He was engaged in project "Unknown malwaredetection using incremental malware detection" MEXT FY(2012-2015).

Yoshiyasu Takefuji is a professor of MusashinoUniversity and was a tenured professor on faculty ofenvironmental information at Keio University fromApril 1992 to March 2021, and was on tenured facultyof Electrical Engineering at Case Western ReserveUniversity since 1988. Before joining Case, he taughtat the University of South Florida for two years andthe University of South Carolina for three years. Hereceived his BS (1978), MS (1980), and Ph.D. (1983) from ElectricalEngineering from Keio University. He received the National ScienceFoundation/Research Initiation Award in 1989 and received the distinctservice award from IEEE Trans. on Neural Networks in 1992 and has beenan NSF advisory panelist

Nguyen Minh Hieu, Pan Haoqian., Yi Liu is graduate student of department of Data Science of Musashino University in Tokyo