International Journal of Innovative Research in Engineering
Volume 7, Issue 4 (July - August 2026), PP: 13-20.
https://www.doi.org/10.59256/ijire.20260704003
www.theijire.com
ISSN No: 2582-8746

Geometry or density: metric comparison for geographical Tor node distribution

Ruo Ando[1], Suyam Ghale[2], Yuto Arai[3], Ko Kumada[4], Yoshiyasu Takefuji[5]
[1]*National Institute of Informatics, Tokyo, Japan.*
[2,3,4,5]*Musashino University, Tokyo, Japan.*

How to cite this paper:
Ruo Ando[1], Suyam Ghale[2], Yuto Arai[3], Ko Kumada[4], Yoshiyasu Takefuji[5], "Geometry or density: metric comparison for geographical Tor node distribution", IJIRE-V7I4-13-20.

Copyright © 2026 by author(s) and Fifth Dimension Research Publication. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Abstract: *The geographic distribution of Tor nodes, particularly Exit nodes, plays an important role in cybersecurity research because attackers frequently exploit the Tor network to conceal their identities, evade tracking, and obscure the origin of malicious traffic. Unfortunately, the previous efforts mainly focuses on the frequency of IP address of Tor nodes, which makes it difficult to conduct detecing structural change-points detection.This study investigates metric comparison for the distributional change-point detection in Tor Exit node data collected daily from October 24, 2025 to March 12, 2026. IP addresses were converted into geographic coordinates using the GeoLite2-City database, and temporal changes in the spatial distribution were analyzed using six distance measures: Wasserstein (Earth Mover's) distance, Sinkhorn distance, Energy distance, Jensen–Shannon distance, Hellinger distance, and Total Variation distance. It was found that using the geometry-based Wasserstein distance yielded clear results in the detection of day-to-day changes. Specifically, significant points of change were detected in early December and mid-February. On the other hand, it was found that density-based distances, such as the JS and Hellinger distances, were effective for detecting changes in trends, such as drift shifts. The timing of the drift shifts largely coincided with the timing detected by the day-to-day change detection method. Conversely, it became apparent that the use of geometry-based methods resulted in the false detection of drift shifts around January.*

Key Words: *Tor nodes; Distance Metric; Change point detection; Wasserstein, JS, Hellinger*

I.INTRODUCTION

The geographic distribution of Tor nodes, particularly Exit nodes, plays an important role in cybersecurity research because attackers frequently exploit the Tor network to conceal their identities, evade tracking, and obscure the origin of malicious traffic. While previous studies often focused on the number of nodes or traffic volume, simple counting-based analysis fails to capture structural and geometric changes in node distributions over time. This study investigates distributional change-point detection in Tor Exit node data collected daily from October 24, 2025 to March 12, 2026. IP addresses were converted into geographic coordinates using the GeoLite2-City database, and temporal changes in the spatial distribution were analyzed using six distance measures: Wasserstein (Earth Mover's) distance, Sinkhorn distance, Energy distance, Jensen–Shannon distance, Hellinger distance, and Total Variation distance. The results show that count-based time-series analysis alone is insufficient for detecting meaningful structural changes in the Tor network. In contrast, several distribution-based metrics successfully highlighted significant temporal anomalies. Among them, the Wasserstein distance produced the clearest and most distinguishable change-point peaks, while Jensen–Shannon, Hellinger, and Total Variation distances exhibited similar but more noise-sensitive behavior. Robust z-score analysis further clarified the temporal magnitude and progression of these changes.

The study demonstrates that preserving geometric and distributional information is essential for effective change-point detection in cybersecurity applications. In particular, Optimal Transport–based methods provide a powerful framework for detecting structural shifts in anonymization networks such as Tor.

II.TOR NODE DISTRIBUTION

Core Objectives of an Attacker

The reasons attackers use Tor are relatively straightforward, but highly effective in practice. Their primary goal is to conceal the original source IP address, making it difficult to identify the true origin of malicious traffic [1]. In addition, Tor helps them avoid tracking by reducing traceability across networks, preventing investigators from easily linking activities back to a single entity. Attackers also use Tor to spoof their geographic location, allowing them to appear as if they are operating from different countries or regions, which can help bypass regional restrictions or raise false assumptions during analysis. In essence, Tor functions as an anonymization infrastructure that shifts the burden of identification onto the defender, increasing the cost and complexity of attribution.

Current research and practice

In practical security research, these aspects play a critical role in understanding and mitigating threats. Researchers and practitioners continuously monitor changes in node distribution to identify shifts in the structure of the network over time. They also focus on detecting anomalies, such as sudden increases in specific regions or unusual concentration (bias) of nodes, which may indicate manipulation or coordinated activity. Furthermore, analyzing these patterns helps uncover early signs of potential attacks, enabling more proactive defense strategies and improving overall situational awareness.

III.METRICS AND MEASURES

IP address: data source and API

Tor node IP addresses, including relays and exit nodes, are publicly available and can be obtained through several official sources provided by the Tor Project. One of the simplest methods is to use the official exit node list at tor project[2] , which provides up-to-date information on exit node IP addresses along with fingerprints and related metadata. This approach is useful for quick reference and straightforward applications. For more structured and programmatic access, the Onionoo API [3] offers detailed information in JSON format. It includes IP addresses, node types such as Guard, Middle, and Exit, as well as bandwidth and uptime data. This makes it particularly suitable for data analysis and automated processing. A more comprehensive but complex option is to use Tor's consensus data, available at [4]. This dataset contains information about all relays in the network, including IP addresses, ports, and role flags. While it provides the most complete view of the Tor network, it requires more effort to parse and interpret.

Wasserstein Distance (EM Distance)

IP Optimal Transport theory has been widely studied in recent years [5]. The Wasserstein distance measures how much "mass" must be moved to transform one distribution into another. Because it preserves geometric information, it is effective for detecting spatial shifts in distributions.

$$W_p(P,Q) = \left(\inf_{\gamma \in \Pi(P,Q)} \int d(x,y)^p \, d\gamma(x,y) \right)^{1/p}$$

- P, Q : probability distributions
- $\Pi(P,Q)$: set of joint distributions with marginals P and Q
- $\gamma(x,y)$: transport plan (mass moved from x to y)
- $d(x,y)$: distance function (e.g., Euclidean)
- p : order of the distance
- inf : infimum (minimum transport cost)

Sinkhorn Distance

Entropy-regularized Optimal Transport enables faster approximation [6]. The Sinkhorn distance is an approximation of the Wasserstein distance that introduces entropy regularization. It is computationally faster and more stable for large-scale data.

$$W_\varepsilon(P,Q) = \min_{\gamma \in \Pi(P,Q)} \sum_{i,j} \gamma_{ij} c_{ij} + \varepsilon \sum_{i,j} \gamma_{ij} \log \gamma_{ij}$$

- γ_{ij} : transport from x_i to y_j
- c_{ij} : cost (distance between x_i and y_j)
- ε : entropy regularization parameter

Energy Distance

Energy statistics provide a non-parametric framework for distribution comparison [7]. Energy distance compares distributions using average pairwise distances between samples. It does not require explicit density estimation and works well for non-parametric analysis.

$$D_E(P,Q) = 2\mathbb{E}\|X - Y\| - \mathbb{E}\|X - X'\| - \mathbb{E}\|Y - Y'\|$$

- $X, X' \sim P$: independent samples from P
- $Y, Y' \sim Q$: independent samples from Q
- $\|X - Y\|$: distance between samples
- $\mathbb{E}$: expectation

Jensen-Shannon Distance

The Jensen--Shannon distance is a symmetric divergence measure between probability distributions [8]. Unlike KL divergence, the JS distance is symmetric and numerically stable. It is widely used in information theory and machine learning for comparing probability distributions.

$$JS(P \parallel Q) = \frac{1}{2}KL(P \parallel M) + \frac{1}{2}KL(Q \parallel M)$$

$$M = \frac{1}{2}(P + Q)$$

- $KL(P \parallel Q)$: Kullback-Leibler divergence
- M : mixture distribution

Hellinger Distance

The Hellinger distance is a classical metric between probability distributions [9]. It measures the similarity between two probability distributions and is bounded between 0 and 1, making interpretation straightforward.

$$H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_i (\sqrt{p_i} - \sqrt{q_i})^2}$$

- p_i, q_i : probabilities of category i
- $\sum_i$: summation over all categories

Total Variation Distance

The Total Variation distance measures the maximum difference between two probability distributions. It is simple, intuitive, and widely used in statistics and probability theory.

$$TV(P,Q) = \frac{1}{2}\sum_i |p_i - q_i|$$

- $|p_i - q_i|$: absolute difference
- $\frac{1}{2}$: normalization factor

Conversion

This algorithm performs IP geocoding for multiple CSV files by converting IP addresses into geographic coordinates (latitude and longitude) using the GeoLite2-City database. First, the algorithm scans the current directory and collects all CSV files. If no files are found, the process terminates immediately. Otherwise, it opens the GeoLite2-City database, which is used to resolve IP addresses into location information. For each CSV file, the algorithm assumes a two-column structure consisting of a timestamp and an IP address. It then initializes a cache (dictionary) to store previously computed results. This cache plays a crucial role in improving efficiency by avoiding repeated lookups for the same IP address. The algorithm iterates over each IP address in the file. Each IP is normalized as a string and checked against the cache. If the IP address has not been processed before, it is converted into an IP object. The algorithm then filters out non-routable addresses, including private, loopback, multicast, reserved, and unspecified IP addresses. These are assigned null values (None,None) because they do not correspond to meaningful geographic locations. For valid public IP addresses, the algorithm queries the GeoLite2 database to obtain the corresponding latitude and longitude. The result is stored in the cache. If any exception occurs during this process (for example, due to an invalid IP format or missing database entry), the algorithm safely assigns null values instead of stopping execution. After processing all IP addresses, the algorithm appends two new columns, lat and lon, to the dataset. Finally, it writes the augmented data to a new CSV file, whose name is derived from the original file with the suffix geocoded".Overall, the algorithm is designed to be robust and efficient. Its use of caching significantly reduces redundant database queries, and its error handling ensures that invalid or unsupported IP addresses do not interrupt the processing pipeline.

Algorithm 1 IP Geocoding for Multiple CSV Files

```
Obtain the list of CSV files F in the current directory
if F = ∅ then
    Terminate the process
end if
Open the GeoLite2-City database
for each file f ∈ F do
    Read f and assign columns as timestamp and ip
    Initialize an empty dictionary cache
    for each IP address ip do
        ip ← normalize as string
        if ip ∉ cache then
            Convert ip into an IP address object
            if ip is private ∨ loopback ∨ multicast ∨ reserved ∨ unspecified then
                cache[ip] ← (None, None)
            else
                Query the GeoLite2 database
                Obtain latitude lat and longitude lon
                cache[ip] ← (lat, lon)
            end if
            if an exception occurs then
                cache[ip] ← (None, None)
            end if
        end if
        Record cache[ip] as the result
    end for
    Add columns lat and lon to the dataset
    Set output filename as basename(f) + _geocoded.csv
    Save the result as a CSV file
end for
Display completion message
```

IV.NUMERICAL RESULTS

A Tor Exit node is the final relay in the Tor network that connects Tor traffic to the normal Internet. When a user accesses a website through Tor, the website sees the IP address of the Exit node instead of the user's real IP address. Exit nodes are essential for anonymity, but they are also one of the most sensitive parts of the Tor network because they directly communicate with external websites. Currently, there are approximately 2300 to 2500 Exit nodes worldwide, although the number changes over time.

The geographic distribution of Exit nodes is highly concentrated. Many are located in countries such as Germany, the United States, the Netherlands, and France. Researchers often analyze not only countries, but also hosting providers and Autonomous Systems (AS) to understand infrastructure concentration. From a research perspective, Exit nodes are important for studying anonymity, network centralization, geographic entropy, and abnormal changes in network structure over time. The data period spans from October 24, 2025 to March 12, 2026, and the data was collected in the following format: timestamp, ip, latitude, longitude as shown in Table 1.

timestamp	ip	lat	lon
2025-10-24T21:36:26	102.130.113.9	-26.3811	27.8376
2025-10-24T21:36:26	109.70.100.1	48.1936	16.3726
2025-10-24T21:36:26	124.198.132.172	-42.0	174.0

TABLE I
SAMPLE TOR NODE DATA

A Tor Exit node is the final relay in the Tor network that connects Tor traffic to the normal Internet. When a user accesses a website through Tor, the website sees the IP address of the Exit node instead of the user's real IP address. Exit nodes are essential for anonymity, but they are also one of the most sensitive parts of the Tor network because they directly communicate with external websites. Currently, there are approximately 2300 to 2500 Exit nodes worldwide, although the number changes over time.

The geographic distribution of Exit nodes is highly concentrated. Many are located in countries such as Germany, the United States, the Netherlands, and France. Researchers often analyze not only countries, but also hosting providers and Autonomous Systems (AS) to understand infrastructure concentration. From a research perspective, Exit nodes are important for studying anonymity, network centralization, geographic entropy, and abnormal changes in network structure over time. The data period spans from October 24, 2025 to March 12, 2026, and the data was collected in the following format: timestamp, ip, latitude, longitude as shown in Table 1. The data was collected by issuing API requests on a daily basis using cron on Linux.

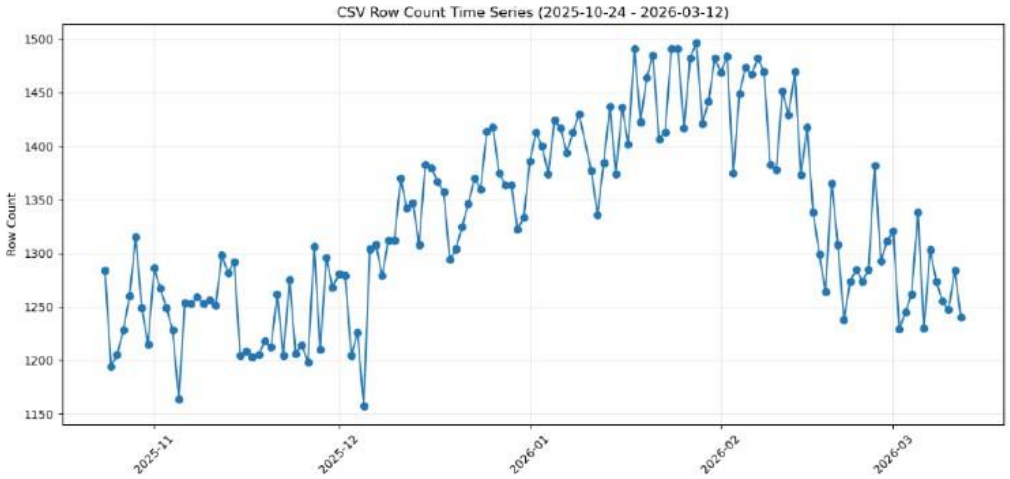


Table no 2: the number of TOR nodes

Table 2 shows the number of TOR nodes from October 2025 to March 2026. The overall shape of the trend is not particularly complex, reaching a peak in February 2026. Additionally, no abrupt increases or decreases are observed in the time series, suggesting that no change points can be detected in the evolution of the TOR node count. As will be explained in Figures 2 and 3, it can be concluded that change points cannot be identified by examining the trend in Figure 1 alone. Therefore, to detect change points in the node count, it is necessary to analyze not only the temporal trend but also the magnitude of distributional shifts and changes in shape.

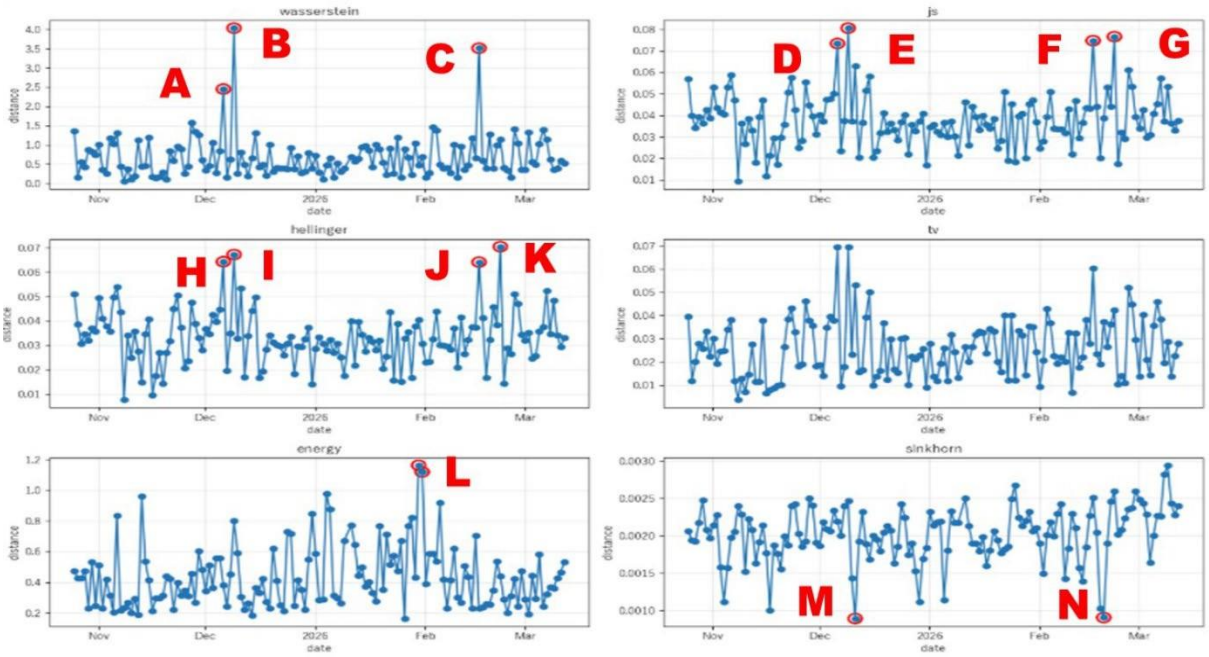


Table no 3: daily change point detection

Table 3 shows changes in the distribution of TOR nodes from October 24, 2025 to March 12, 2026, measured using six different distance metrics. At each time point, the distance was calculated between the current TOR node distribution and the distribution from the previous period (one day earlier).

Among the six methods, four metrics correctly detected anomalies: the EM (Wasserstein) distance, JS distance, Hellinger distance, and TV distance. These metrics produced large distance values in early December and mid-February. The JS distance and Hellinger distance produced almost identical results, each detecting two change points around early December and mid-February. Compared with the JS and Hellinger distances, the TV distance appears to be somewhat more

sensitive to noise. However, the deviations of these three distances from the overall average were smaller than those of the EM distance. From this perspective, the EM distance can be considered the most accurate in detecting change points.

In addition, while the EM distance detected clearly outstanding peaks in both the first and second halves of the observation period, the JS distance, Hellinger distance, and TV distance each detected two spikes in both periods, making it more difficult to determine which spikes actually corresponded to true change points. Considering these observations as a whole, it can be concluded that the EM distance most appropriately detected the change points.

From Figures 1 and 2, it can be seen that simply counting the number of observed nodes is insufficient for effective change-point detection in cybersecurity, because important density and geometric information contained in the distribution is lost. There are two reasons for this. First, in TOR node traffic, the total number of nodes itself is not the primary issue; rather, the important factors are the geographical information and the connectivity structure of the nodes. Second, In general, "TOR nodes that repeatedly appear and disappear," especially Exit nodes or Relay nodes that operate only for a short period of time, tend to attract attention from both operational and security perspectives. In the TOR network, nodes that operate stably over long periods are generally considered more trustworthy, and directory authorities also place importance on stability and uptime. For this reason, nodes that appear only briefly and then disappear may have the following risks or characteristics.

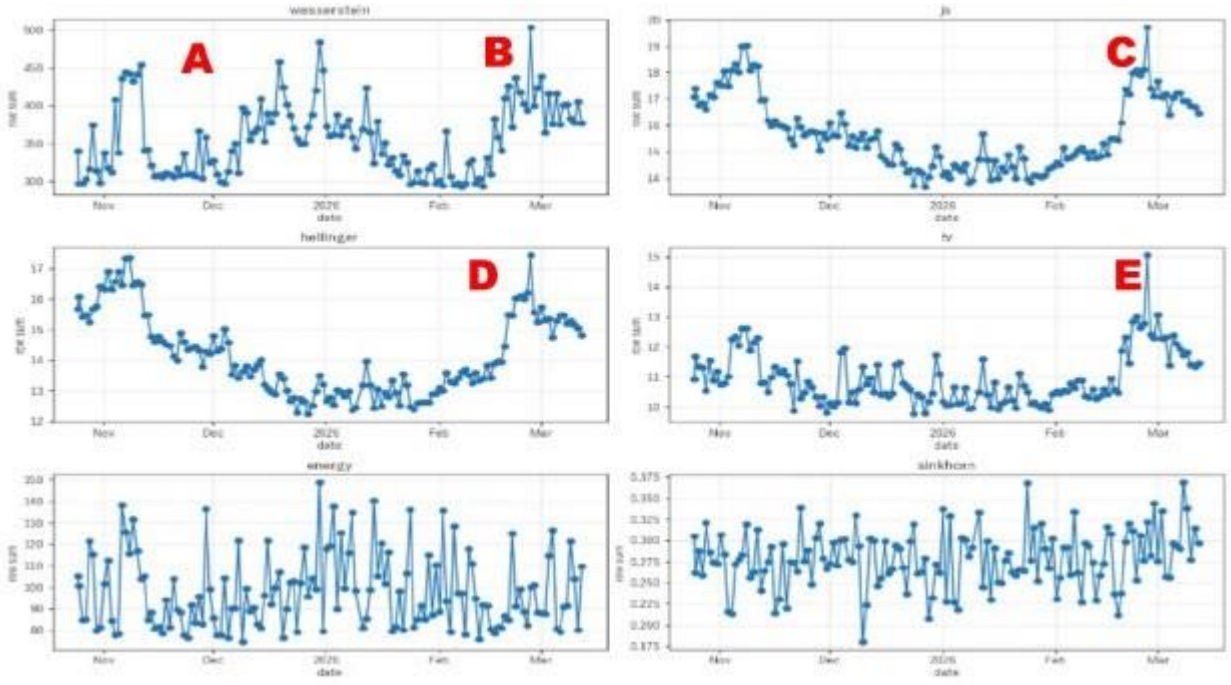


Table no 4: cumulative change point detection

Table 4 shows the robust z-scores computed and plotted for the time series of each distance measure calculated in Figure 2. The robust z-score is a statistical measure used to detect outliers while being less sensitive to extreme values than the standard z-score. Unlike the ordinary z-score, which uses the mean and standard deviation, the robust z-score uses the median and the Median Absolute Deviation (MAD). This makes it more stable when the data contain large outliers or skewed distributions.

This makes it possible to observe the progression and magnitude of changes over time. For the four methods in which change-point detection worked well (EM, JS, Hellinger, and TV), it can be confirmed that each contains regions with large magnitudes of change in the latter half of the observation period (C, D, E).

In addition, the density-based methods — JS, Hellinger, and TV — exhibit very similar trend shapes over time. However, the EM distance is able to capture more fine-grained trend variations, detecting one major peak of change in the first half and another in the latter half of the observation period. It is also interesting that the geometry-based method, EM, and the density-based methods, JS and Hellinger, produce almost symmetrical results during the middle phase of the observation period.

V.DISCUSSION

The six distances can be broadly classified into three categories: Optimal Transport, Integral Probability Metrics (IPM), and f-divergences. First, the EM distance (Wasserstein distance) and the Sinkhorn distance belong to Optimal Transport. These distances are defined from the perspective of how one probability distribution can be transformed into another with minimal cost. They strongly reflect the geometric structure of the underlying space and are particularly effective at capturing shifts in location and changes in distribution shape. However, since they are formulated as optimization problems, they can be computationally expensive. The Sinkhorn distance alleviates this issue by introducing entropy regularization, providing a faster approximation.

Algorithm 2 Wasserstein-Based Analysis (Simplified)

Require: $\{(t_i, X_i)\}_{i=1}^{n},\ X_i = \{(lat, lon)\}$
Ensure: Distance matrix D, adjacent distances V, outliers
1: **if** $n < 2$ **then**
2: Terminate
3: **end if**
4: Compute pairwise distance matrix:

$$D_{ij} = \sqrt{W_1(X_i.\text{lat}, X_j.\text{lat})^2 + W_1(X_i.\text{lon}, X_j.\text{lon})^2}$$

5: Compute adjacent distances:

$$V_i = \sqrt{W_1(X_{i-1}.\text{lat}, X_i.\text{lat})^2 + W_1(X_{i-1}.\text{lon}, X_i.\text{lon})^2}$$

6: Compute robust Z-score:

$$z_i = 0.6745 \cdot \frac{V_i - \text{median}(V)}{\text{MAD}(V)}$$

7: Detect outliers: $|z_i| \geq Z_{\text{thr}}$
8: Compute row sums: $S_i = \sum_j D_{ij}$
9: Output $D, V, \text{outliers}, S$

Next, the Energy distance falls under the category of IPM. In this framework, the difference between two distributions is measured through the difference in expectations over a class of functions. The Energy distance, in particular, is defined based on the average pairwise distances between samples, making it a non-parametric method that does not require explicit density estimation. While it does not explicitly model transport like Optimal Transport methods, it captures differences between point clouds effectively. However, its ability to interpret structural relationships or correspondences between clusters is more limited.

Algorithm 3 Jensen–Shannon Distance Analysis (Simplified)

Require: $\{(t_i, X_i)\}_{i=1}^{n},\ X_i = \{(lat, lon)\}$
Ensure: Distance matrix M, adjacent distances V, outliers
1: **if** $n < 2$ **then**
2: Terminate
3: **end if**
4: Convert each X_i into probability vector p_i
5: Compute pairwise distance matrix:

$$M_{ij} = \sqrt{\frac{1}{2}D_{\text{KL}}(p_i \parallel m) + \frac{1}{2}D_{\text{KL}}(p_j \parallel m)},$$

6: Compute adjacent distances:

$$V_i = \sqrt{\frac{1}{2}D_{\text{KL}}(p_{i-1} \parallel m) + \frac{1}{2}D_{\text{KL}}(p_i \parallel m)}$$

7: Compute robust Z-score:

$$z_i = 0.6745 \cdot \frac{V_i - \text{median}(V)}{\text{MAD}(V)}$$

8: Detect outliers: $|z_i| \geq Z_{\text{thr}}$
9: Compute row sums: $S_i = \sum_j M_{ij}$
10: Output $M, V, \text{outliers}, S$

Finally, the Jensen-Shannon (JS) distance, Hellinger distance, and Total Variation distance are categorized as f-divergences. These measures are based on the ratio of probability densities, p(x)/q(x), and have a strong foundation in information theory. They focus solely on differences in the shape of distributions and do not incorporate geometric information. As a result, they may fail to detect meaningful spatial shifts when the overall shape remains similar. On the other hand, they are computationally efficient, numerically stable, and well-suited for tasks such as distribution comparison and model evaluation.

This algorithm begins by receiving a sequence of time-indexed spatial datasets, where each dataset consists of latitude and longitude observations. In lines 1 to 3, it first checks whether there are at least two datasets available; if not, the process terminates because no comparison can be made.

This JS algorithm begins by receiving a sequence of time-indexed spatial datasets, where each dataset consists of latitude and longitude observations. In lines 1 to 3, it checks whether at least two datasets are available. If there is only one dataset, the process stops because no distance comparison can be made.

VI.CONCLUSION

This study investigated distributional change-point detection in Tor Exit node distributions using six statistical distance measures: Wasserstein (EM) distance, Sinkhorn distance, Energy distance, Jensen–Shannon distance, Hellinger distance, and Total Variation distance.

Daily Tor Exit node data collected from October 24, 2025 to March 12, 2026 were converted into geographic coordinates using the GeoLite2-City database, enabling analysis of temporal changes in spatial node distributions. The experiments demonstrated that simple node-count analysis was insufficient for detecting meaningful structural changes in the Tor network. Although the total number of nodes showed relatively smooth temporal behavior without abrupt fluctuations, significant distributional changes were detected when geometric and density-based information was incorporated.

Among the six methods, the Wasserstein (EM) distance produced the clearest and most distinguishable change-point peaks, particularly during early December 2025 and mid-February 2026. Jensen–Shannon distance, Hellinger distance, and Total Variation distance also detected similar anomalous periods, although these methods exhibited more noise-sensitive behavior and generated multiple nearby spikes. In contrast, the Sinkhorn distance and Energy distance were less effective in isolating prominent structural changes in the observed period.

The robust z-score analysis further clarified the magnitude and temporal progression of these anomalies. In particular, the EM distance consistently generated large robust z-score peaks corresponding to major distributional shifts, while the density-based methods exhibited highly similar temporal patterns. These results suggest that Optimal Transport–based approaches preserve important geometric characteristics of the Tor node distribution that are not captured by conventional density-based divergences alone.

From a cybersecurity perspective, the results indicate that monitoring only the total number of Tor nodes is insufficient for practical anomaly detection. Instead, the geographic structure, spatial concentration, and temporal redistribution of nodes provide more meaningful indicators of abnormal activity. This is particularly important because short-lived Tor Exit nodes and rapidly changing node distributions may reflect operational instability, infrastructure manipulation, or potentially malicious behavior.

Overall, the study demonstrates that geometric distribution analysis using Optimal Transport provides an effective framework for detecting structural changes in anonymization networks. Future work includes incorporating graph topology information, Autonomous System (AS) relationships, relay connectivity structures, and temporal graph models in order to further improve change-point detection and cyber threat analysis in large-scale anonymization infrastructures.

References

1. R. Dingledine, N. Mathewson, and P. Syverson, "Tor: The Second-Generation Onion Router," in Proceedings of the 13th USENIX Security Symposium, San Diego, CA, USA, Aug. 2004, pp. 303-320.
2. The Tor Project, "Onionoo: The Tor Network Status Protocol." [Online]. Available: https://metrics.torproject.org/onionoo.html [Accessed: Jun. 25, 2026].
3. [The Tor Project, "CollecTor: Tor Metrics Data-Collecting Service." [Online]. Available: https://metrics.torproject.org/collector.html [Accessed: Jun. 25, 2026].
4. The Tor Project, "Vote and Consensus Status Document Formats." [Online]. Available: https://spec.torproject.org/dir-spec/consensus-formats.html [Accessed: Jun. 25, 2026].
5. G. Peyré and M. Cuturi, "Computational Optimal Transport," Foundations and Trends in Machine Learning, vol. 11, nos. 5-6, pp. 355-607, 2019.
6. M. Cuturi, "Sinkhorn Distances: Lightspeed Computation of Optimal Transport," in Advances in Neural Information Processing Systems, vol. 26, 2013.
7. G. J. Székely and M. L. Rizzo, "Energy Statistics: A Class of Statistics Based on Distances," Journal of Statistical Planning and Inference, vol. 143, no. 8, pp. 1249-1272, 2013.
8. D. M. Endres and J. E. Schindelin, "A New Metric for Probability Distributions," IEEE Transactions on Information Theory, vol. 49, no. 7, pp. 1858-1860, Jul. 2003.
9. E. Hellinger, "Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen," Journal für die reine und angewandte Mathematik, vol. 136, pp. 210-271, 1909.