Exploring the Relationship Between Socially Observable Mobility Patterns and Social Connectivity

Chau Do and Dang Ly

1. Abstract

Despite having a high degree of freedom and variability [1], human mobility demonstrates certain geographical, temporal, and social patterns. In this work, we explore the relationship between socially observable mobility behaviour and social connectivity. Using a social media network with location-based check-in data in Finland, we show that the similarities in geographical and temporal mobility patterns are highly indicative of social connections. Moreover, we find that the degree to which a user's mobility is influenced by their social connections—measured by the likelihood of visiting locations previously visited by friends—is significantly correlated with their structural positions in the network. These findings offer insights into the interplay and influence between network positions and socially observable mobility patterns.

2. Research questions

We aim to address the following research questions:

- **RQ1**: Are the similarities in spatiotemporal mobility patterns indicative of social connections?
- **RQ2**: Is the strength of social influence on a user's mobility correlated with their structural position in the network?

3. Methodology

3.1. Data

3.1.1. Overview

We use the Gowalla dataset, retrieved from the Stanford Network Analysis Project (SNAP) [2]. Gowalla was a location-based social networking service that allowed users to connect and check in to share their locations with friends in their virtual circle. The dataset includes a friendship network of user connections, along with user check-ins containing geographic coordinates and timestamps, spanning from February 2009 to October 2010. This combination of social and spatiotemporal data makes the dataset well-suited for studying the relationship between human movement patterns and social connections.

The original dataset contains a graph of 196 591 nodes (users) and 950 327 edges (connections), along with 6 442 890 check-ins and 1 280 956 check-in locations spanning multiple countries and geographical regions. The check-in information includes the latitude and longitude of the check-in location, the ID of the user, and the ID of the location.

3.1.2. Data preprocessing

We focus our analysis on users who are consistently active in Finland, possibly representing the community of Finnish-based users on Gowalla. To this end, we first reverse-map the latitudes and longitudes of all check-ins to their regions and countries using the reverse-geocoder Python package. We next filter to include only the check-ins in Finland, resulting in 9167 check-ins from 606 users. However, this subset might

include tourists and short-term visitors, so to primarily include permanent residents in Finland, we filter to include only users whose majority of check-ins are in Finland. This results in 202 users with 260 connections and 7678 check-ins in 2376 locations. The geographical distribution of check-ins is illustrated in Fig. 1. The histogram of the number of check-ins per user are given in Fig. 2. Fig. 3 illustrates the average number of check-ins per user aggregated by weekday and 3-hour time intervals. We note the clear periodicity of the average number of check-ins, which reflects the periodicity of human activity throughout the day and the week. The degree distribution of user nodes in the social connectivity graph is given in Fig. 4.

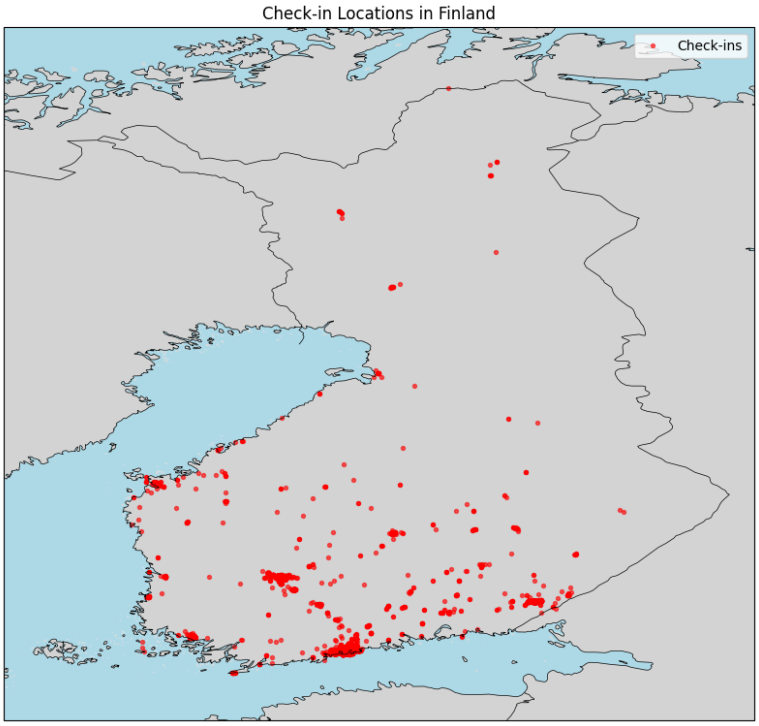

Fig. 1: The geographical distribution of check-ins in Finland

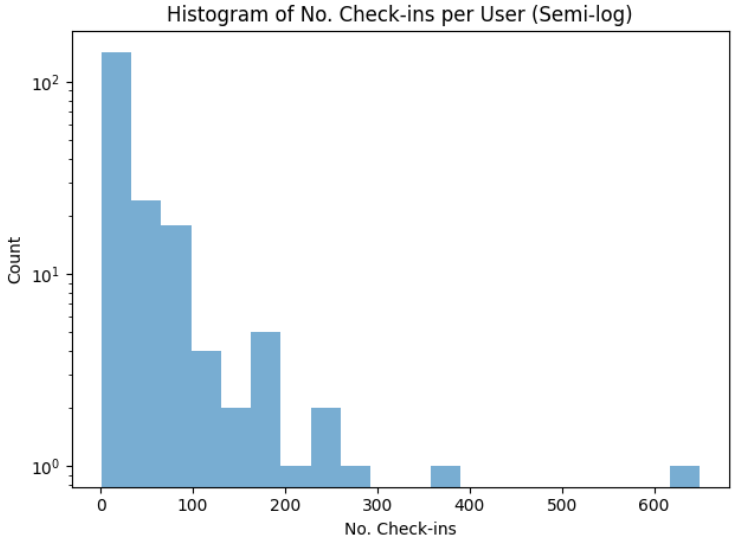

Fig. 2: Histogram of the number of check-ins per user

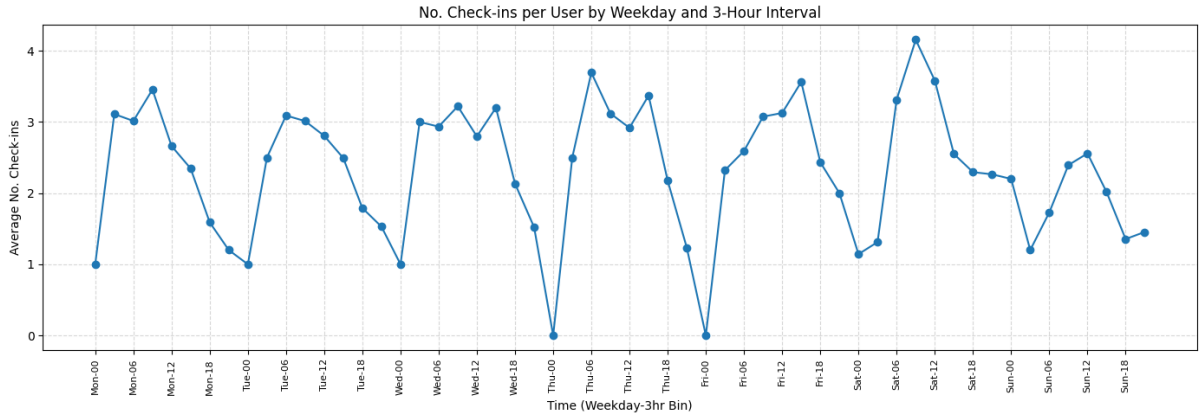

Fig. 3: Average number of check-ins per users aggregated by weekday and 3-hour intervals

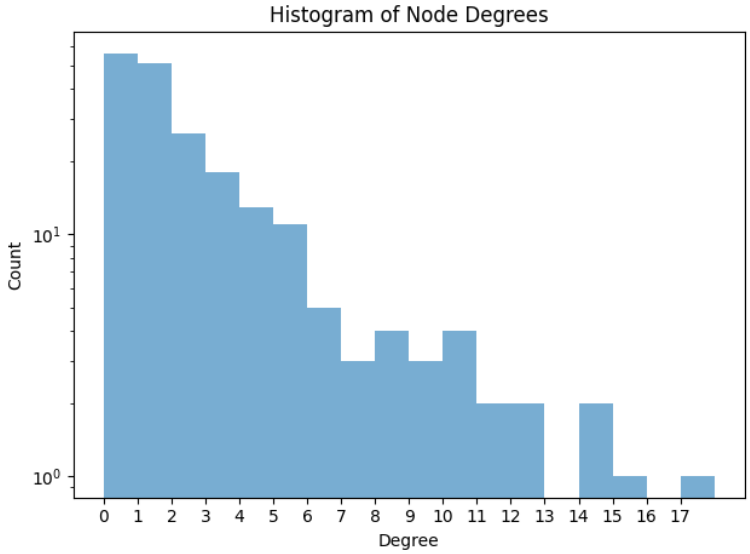

Fig. 4: Node degree distribution of the social connectivity graph

3.2. Addressing RQ1

3.2.1. Measuring the similarities in spatial and temporal check-in patterns between users

Given each user's set of check-in locations, we employ several metrics to quantify the similarities in their spatial and temporal check-in patterns.

Spatial patterns

A check-in location can be viewed as either a unique identity with the location ID, or as a geographical point with coordinates (latitude, longitude). The most straightforward way to identify the check-in location is by using the location ID. Given the check-in history of two users, we measure the overlap by calculating the Jaccard distance between the two sets of visited locations. However, this method only considers unique locations and ignores the fact that a user could visit a location multiple times. To account for this, we first compute the vector of check-in counts in shared locations for each user, then measure the directional alignment between the two vectors with the cosine distance. However, by viewing the check-in locations as unique IDs, we disregard the spatial distribution of locations. To address this, we additionally consider the coordinates of the check-in locations, effectively representing each user's check-in history as a cloud of

points. Finally, to quantify the difference between two users' check-in history, we employ the earth's mover distance (EMD) [2], a metric designed to measure the distance between two point distributions. The three metrics (Jaccard index, cosine similarity, and EMD) comprehensively characterizes the similarity between the spatial patterns of two users.

Temporal patterns
Given check-in timestamps, we consider temporal patterns of user check-ins both in terms of the time of the day and the weekday. One characteristic of the time of the day is that it wraps around – for example, 00:01 is close to 23:59 – so using simple arithmetic differences would not be suitable. To address this, we encode the time of day (ToD) using a two-dimensional circular representation. Specifically, each timestamp is converted into a point on the unit circle by mapping the number of seconds since midnight to sine and cosine components. This transformation preserves the circular nature of time. With the ToDs of check-ins represented as point clouds, the distance between the two point clouds is then measured with the EMD. Finally, to measure the similarity in weekday check-in patterns, we calculate the fraction of check-ins per weekday for each user, resulting in a distribution of check-ins for each weekday. We then use the Jensen-Shannon divergence (JSD) [3] to measure the difference between the weekday distributions of two users. JSD is a symmetric generalization of the Kullback–Leibler divergence, and can better handle the presence of categories with 0 probability (i.e. weekdays with no check-ins). The two metrics (EMD of ToDs and JSD of weekdays) comprehensively characterizes the similarity between the spatial patterns of two users.

3.2.2. Comparing the similarity between check-in patterns of connected and unconnected users

To investigate whether social connections are associated with similar check-in patterns and thus more similar mobility patterns, we compare the similarity between check-in patterns of connected and unconnected users. Specifically, we calculate the Jaccard index, cosine similarity, and EMD of check-in locations, along with the EMD of ToDs and the JSD of weekdays for each pair of connected users, totaling at 260 pairs corresponding to 260 edges. To establish a baseline, we randomly sample 260 pairs of unconnected users and calculate the same five metrics for their check-ins. In addition, we conduct a permutation-based statistical test for each distance metric and observe that the mean distances between connected users are significantly smaller than the mean distances between unconnected users for all metrics. The distributions of the five metrics are given in Fig. 5. The results of the statistical tests are given in Table 1.

Metric	**Mean (connected)**	**Mean (unconnected)**	**P-value**
Jaccard distance	0.945 +/- 0.068	0.997 +/- 0.012	**.000**
Cosine distance	0.367 +/- 0.401	0.883 +/- 0.313	**.000**
EMD	198 +/- 125	337 +/- 168	**.000**
EMD – ToD	0.685 +/- 0.179	0.834 +/- 0.191	**.000**
JSD – weekday	0.356 +/- 0.177	0.481 +/- 0.182	**.000**

Table 1: Mean distances between connected and unconnected users and P-values of the difference

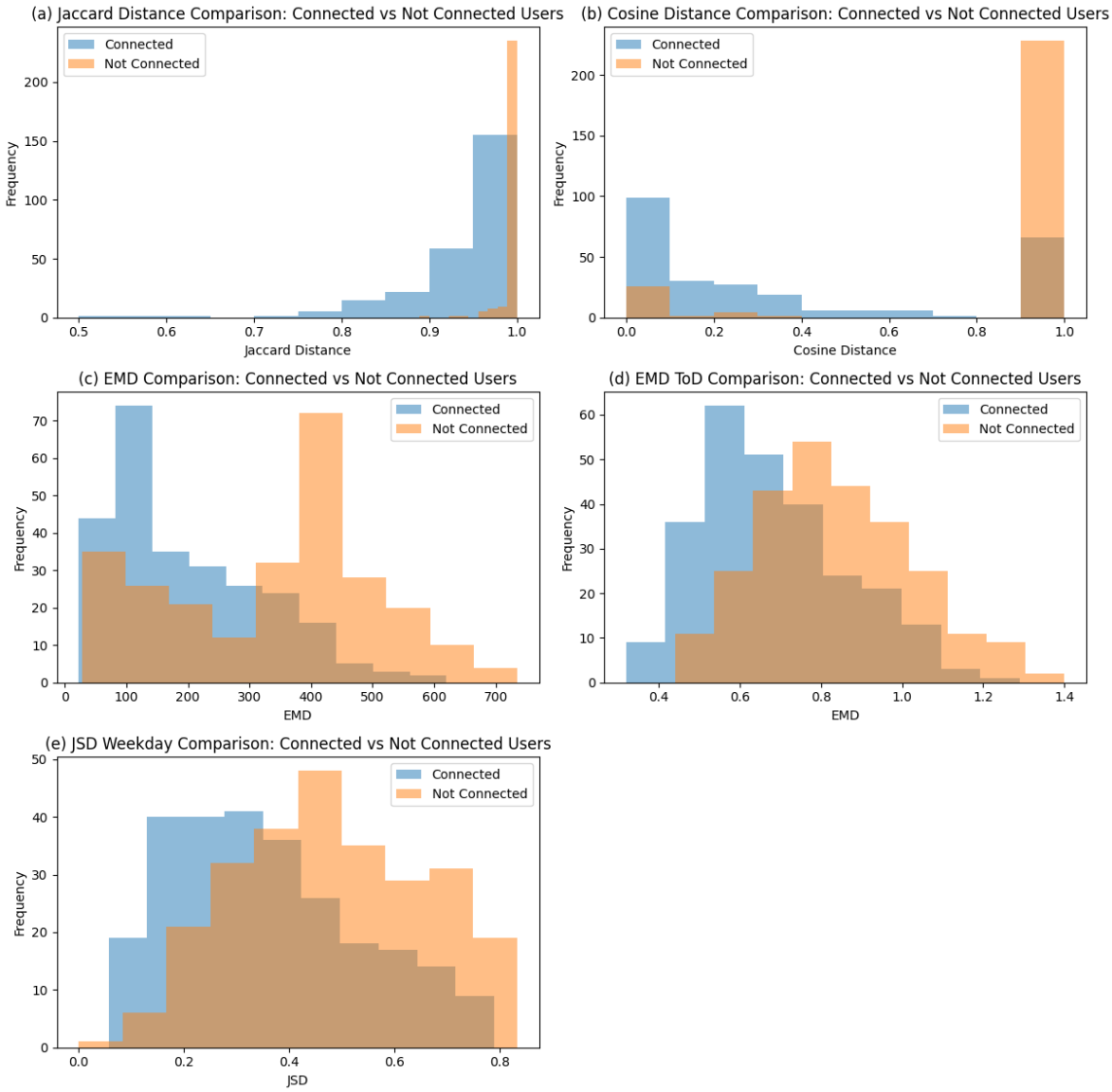

Fig. 5: Distribution of the Jaccard distance (a), cosine distance (b), EMD (c), EMD of ToD (d), and JSD of weekday (e)

3.2.3. Link prediction with distance metrics

To quantify the effects of mobility distance metrics in indicating the existence of a social connection between two users, we define a modeling task where the five distance metrics are used to predict whether two users are connected. Specifically, we employ a logistic regression model, optimized using the maximum likelihood estimation (MLE) method with the statsmodels Python package.

First, we develop an explanatory model using the entire dataset, which consists of 260 connected pairs and 260 randomly sampled unconnected pairs. The estimated coefficients and their 95% confidence intervals are reported in Table 2. Overall, the coefficients of all the spatial distances (Jaccard distance, cosine distance, EMD) and the EMD of ToD are statistically significant, while the coefficient of the JSD of weekday is not. Moreover, all significant coefficients are negative, suggesting that greater dissimilarity in mobility patterns between users is associated with a lower likelihood of a

social link. Among these, the Jaccard distance of spatial patterns exhibits the strongest negative association.

Metric	Estimate	0.025	0.975	P-value
Jaccard distance	-1.984	-2.931	-1.037	**.000**
Cosine distance	-0.426	-0.751	-0.101	**.010**
EMD	-0.546	-0.803	-0.290	**.000**
EMD – ToD	-0.544	-0.884	-0.205	**.002**
JSD – weekday	0.073	-0.252	0.398	.661

Table 2: Estimated values, confidence intervals, and P-values of the coefficients of distance metrics in the link prediction model

Next, we develop a predictive model by equally splitting the dataset into a training set and a test set and fitting the same model on the training set. The predictive performance of the model is summarized in Table 3. The model demonstrates a strong predictive performance, suggesting that the similarity in socially observable mobility patterns is highly indicative of a social link.

Metric	Value
Accuracy	0.80
Precision	0.86
Recall	0.72
AUC	0.87
F1	0.78

Table 3: Predictive performance of the link prediction model

3.3. Addressing RQ2

3.3.1. Quantifying the influence of connections on mobility patterns

To investigate the influence of social connections on user mobility, we use the proportion of check-ins at locations previously visited by friends as a proxy for influence strength, now directly referred to as the *influence strength* for brevity. The distribution of the influence strength for each user is given in Fig. 6.

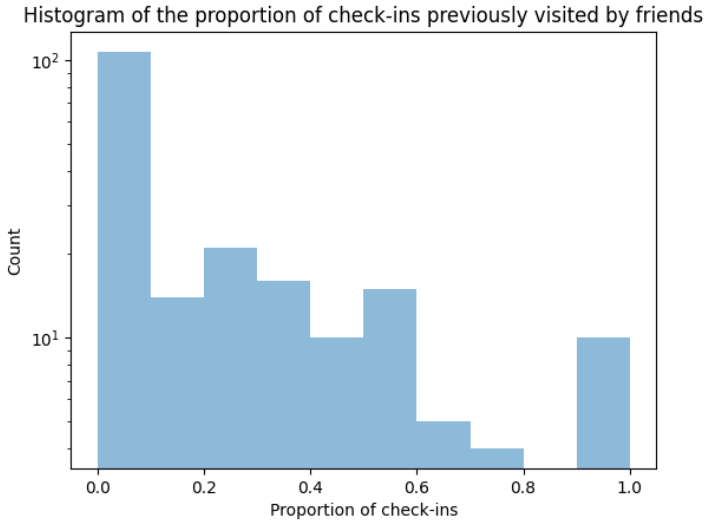

Fig. 6: Histogram of the proportion of check-ins previously visited by friends (influence strength)

3.3.2. Correlation between influence strength and network metrics

To understand the association between influence strength and the structural position of users in the network, we calculate the Spearman's correlation between the influence strength and several node metrics, including node degree, clustering coefficient, betweenness centrality, and closeness centrality. The P-values of the Spearman's rho are estimated with a permutation test. The results are provided in Table 4. All node metrics show a moderately strong, statistically significant positive correlation with the influence strength. Among the four metrics, node degree exhibits the highest Spearman's rho.

Node metric	Spearman's rho	P-value
Degree	0.715	**.000**
Clustering	0.551	**.000**
Betweenness	0.493	**.000**
Closeness	0.544	**.000**

Table 4: Spearman's rho and P-values of the correlation between node metrics and influence strength

To further characterize the relationship between node metrics and influence strength, we develop an explanatory linear regression model with the four node metrics as predictors and the influence strength as the label. The model is fitted using the statsmodel package. The estimated values and the confidence intervals of the coefficients are given in Table 5.

Node metric	Estimate	0.025	0.975	P-value
Degree	2.579	1.410	3.749	**0.000**
Clustering	1.575	0.768	2.382	**0.000**
Betweenness	-0.549	-1.529	0.432	0.271
Closeness	1.350	0.466	2.234	**0.003**

Table 5: Estimated values, confidence intervals, and P-values of the coefficients of node metrics in the influence strength regression model

Among the four node metrics, degree, clustering, and closeness all have statistically significant and positive regression coefficients, with degree exhibiting the strongest effect. This supports the earlier observation of the positive correlations between these metrics and the influence strength. In contrast, betweenness centrality is not a significant predictor, despite its statistically significant correlation with influence strength. A possible explanation is that the node metrics are not independent, so betweenness is correlated with other metrics (e.g. degree) that are in turn correlated with influence strength. Thus, once we control for these variables in the regression, the contribution of betweenness is no longer significant. Despite this, we note that the variance inflation factors of all metrics are sufficiently low (< 4), indicating that multicollinearity is not a concern, and including betweenness is not inherently problematic.

These findings suggest that in general, users who are more connected and central in the network are more likely to check-in at locations previously visited by their friends, indicating stronger social influence on their mobility. However, users acting as bridges

between communities (i.e. those with high betweenness centrality) are not more likely to be influenced by their friends, possibly due to having more diverse connections.

4. Discussion and conclusion

In this work, we investigate the interplay between socially observable mobility patterns and social connectivity. Utilizing a range of distance metrics to quantify the similarities between spatiotemporal mobility patterns of users, we show that such similarities are highly indicative of social connections. Moreover, we introduce the proportion of check-ins at locations previously visited by friends as a proxy for the strength of social influence on one's mobility. We further show that this influence strength is positively correlated with several network metrics, including degree, clustering, and closeness, which indicates that well-connected, central users in the network are more strongly influenced by their connections. Interestingly, while exhibiting a positive correlation, betweenness does not act as a significant predictor of the influence strength. This may suggest that users with high betweenness, who often bridge distinct communities, experience more diffuse or diluted social influence.

This work has several limitations. First, we only use a small portion of the dataset, focusing only on users inferred to be based in Finland. Therefore, our conclusions might not be applicable to other geographical regions or the dataset as a whole. Furthermore, we would like to emphasize that our analysis is focused on socially observable mobility patterns (check-in data from a social media platform), not full-ranged mobility patterns. In other words, the check-in locations are intentionally made observable from the user, and there might be a bias towards social media-suitable content. In addition, we do not have the timestamps for when connections are formed, so the social network is treated as static. This introduces some noises into our approximation of the influence strength, as connections could have formed after the overlapping check-ins occurred.

This work can be extended in several directions. First, replicating the analysis in other countries and comparing the results may offer insights into how geographical and cultural differences shape mobility patterns and their relationship to social connectivity. Additionally, extending the check-in data by matching coordinates to actual place names and categories (e.g., retail, healthcare, education, leisure) would provide semantically richer context, enabling deeper analysis beyond spatial identifiers alone.

5. References

[1] Cho, E., Myers, S. A., & Leskovec, J. (2011). *Friendship and mobility: User movement in location-based social networks*. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. doi:10.1145/2020408.2020579

[2] Flamary, R., Courty, N., Gramfort, A., et al. (2021). *POT: Python Optimal Transport*. *Journal of Machine Learning Research, 22*(78), 1–8. http://jmlr.org/papers/v22/20-451.html

[3] Lin, J. (1991). *Divergence measures based on the Shannon entropy*. *IEEE Transactions on Information Theory*, 37(1), 145–151. doi:10.1109/18.61115

6. Acknowledgements

We use OpenAI's ChatGPT 4o to proofread and polish human-written text.

7. Contributions

C.D. and D.L. jointly worked on all aspects of this project, including conceptualization, methodology, formal analysis, and writing.