"Exploring the Six Degrees of Separation in Non-Reciprocal Social Networks"

Thomas Hamilton-Smith
School of Computer Science
The University of Adelaide
a1886319@student.adelaide.edu.au

Matthew Brennan
School of Computer Science
The University of Adelaide
matthew.brennan@adelaide.edu.

Abstract

This paper investigates the "Six Degrees of Separation" theory in Twitter's non-reciprocal social network. Analysing metrics like path length, clustering, and degree distribution, the study finds an average path length of 3.89 and a network diameter of 7. These results suggest that while most users are connected within six degrees, some paths exceed this due to one-sided connections. Compared to Facebook, Twitter's slightly greater separation highlights unique dynamics in non-reciprocal networks. This research addresses a gap in social network analysis by extending small-world theory to non-reciprocal relationship structures.

1.1 Introduction

Social network analysis (SNA) is critical for understanding the complex web of relationships in modern-day social media, where connections are often described by the "Six Degrees of Separation" social concept. The concept suggests that any two people are linked by a chain of six or fewer people. However, in today's social networks, connections are not always reciprocal. You may follow a celebrity on Instagram who doesn't follow you back, making this a one-way, direct connection. The significance of connections varies between people on these platforms, but current SNA tools often treat all connections equally.

This creates a gap, as current methods are unable to handle these direct/weighted connections effectively. This can lead to oversimplified or inaccurate insights into network influence, structure and connectivity between people. By considering and addressing this gap, this project aims to test the six degrees theory and enhance the accuracy and scalability of SNA tools, particularly in the context of large-scale social networks like Instagram and Twitter, where the correlation with real-life relationships no longer holds.

1.2 Motivation

The motivation for this paper is to research improved methods to accurately account for direct and weighted connections on social networks, which present a significant gap in SNA. Platforms such as Instagram, where connections are often one-sided and vary in importance, such as the comparison of following a friend vs a celebrity/brand, require a more sophisticated analysis method than traditional SNA tools can offer. If these nuances are not considered, any resulting analysis may lead to misinterpretations of influence, community structure and the overall dynamics of a network.

This project hypothesises that by optimising existing algorithms or developing new ones, the accuracy and scalability of SNA tools can be significantly improved to address the gap in understanding non-reciprocal connections. It also explores the possibility that the "Six Degrees of Separation" theory may not apply in non-reciprocal social networks such as Instagram and Twitter. By doing so, the research will help bridge the gap between real-life relationships and their digital counterparts.

2. Literature Review

Z. Zhang, F. Zhou and H. Ning (2022) examined the evolution of the "Six Degrees of Separation" concept in modern online social networks. Their research highlighted that technological advancements and optimisation algorithms have reduced the degrees of separation amongst users, with some platforms showing less than four degrees. This finding emphasises the need to account for these changes in network dynamics, particularly regarding influence and connectivity in digital platforms. [1]

D. J. Brass (2022) provided a comprehensive review of recent developments in SNA, focusing on the gaps (structural holes) within networks and the role of brokerage in connecting individuals across these gaps. The review emphasised the importance of understanding the nature of ties, whether they are weak or strong, positive or negative, and how these relationships, particularly non-reciprocal or asymmetric connections, can influence organisational outcomes. Brass noted that these dynamics are integral to the concept of "Six Degrees of Separation," where even the distant, non-reciprocal ties will play a crucial role in linking individuals across large networks. [2]

L. Zhang and W. Tu (2009) revisited the "Six Degrees of Separation" theory in online societies, applying mathematical modelling and empirical measurements across social platforms such as Arne Miner, Facebook, and Tencent QQ. Their study found that while the concept generally holds, the degree of separation varies based on the size and structure of the network (the number of users), suggesting that online social connections may exhibit even more minor degrees of separation compared to traditional, real-world settings. [3]

Research conducted by researchers at Facebook/Meta (2016) reported a significant decrease in the average degrees of separation between users on their site, down to just an average of 3.57. This finding accentuates the accelerated connectivity facilitated by large-scale online networks with large user counts, and advocates for re-examining traditional network analysis methodologies to account for this change. It emphasises that such a shift is necessary to better understand modern social dynamics. [4]

Despite the extensive research on social networks that rely on mutual connections, such as Facebook, a noticeable gap still exists in understanding non-reciprocal connections prevalent on platforms like Instagram. For instance, a user might follow a celebrity like Justin Bieber, without a reciprocal connection. The exploration of how many connections it would take to circle back to the original user (if at all), could provide valuable insights into the network dynamics on such platforms. This area remains underexplored, stressing the critical and urgent need for further research to address the unique characteristics of non-reciprocal social networks, and see how it relates to the "six degrees" theory.

3. Methodology

This methodology outlines the steps taken to address gaps in social network analysis (SNA) for non-reciprocal networks, specifically focusing on testing the "Six Degrees of Separation" theory in the context of platforms with one-sided connections, such as Twitter.

3.1 Problem Breakdown and Data Sourcing.

To investigate degrees of separation in a large non-reciprocal network, the initial approach was to collect data from Instagram to analyse relationship structures amongst connected accounts. However, Instagram's strict privacy restrictions and limited data access required a change of approach. Instead, we sourced data from the Stanford Large Network Dataset Collection (SNAP), focusing on a Twitter dataset containing non-reciprocal connections across over 80,000 users. Twitter's directed connections (e.g., following without mutual following) effectively mirror the one-sided relationships typical on Instagram, providing a comparable framework for analysis.

While Twitter connections are generally less personal than those on Instagram, this larger dataset allowed for a robust examination of the "Six Degrees of Separation" theory within a non-reciprocal network. Twitter's scale and structure enabled a more thorough exploration of degrees of separation, addressing the gap by allowing analysis of influence and connectivity in one-sided social connections.

3.2 Data Preparation and Conversion.

After sourcing the dataset, the data was processed to fit a structured format suitable for network analysis. The raw data from the SNAP dataset was first converted into an Excel-compatible format with nodes representing user accounts and edges representing follower-following relationships. This step included:

- **Organising nodes and edges:** Each node represents a unique Twitter user ID, and directed edges indicate a follower-following relationship.
- **Importing into Gephi:** The initial CSV file was imported into Gephi for visualisation and modularity analysis which helped identify clusters
- **Exporting for Python Analysis:** Processed data was exported in a format compatible with Python for further analysis, enabling detailed metric calculations like path lengths and centrality measures.

3.3 Algorithm Development for Degree of Separation Analysis.

Custom algorithms were developed in Python to assess degrees of separation within the network. This included:

Pathfinding algorithms: These algorithms calculated the shortest path between randomly selected nodes, approximating the degrees of separation across the given network sample size, visualised by the following equation:

$$\sim Avg.\, Path\ Length = \frac{1}{Sample\ Size} \sum_{i=1}^{Sample\ Size} d(u_i, v_i)$$

- **Average Path Length and Diameter Calculations:** Python was used to compute the average path length and network diameter, automating the calculations to achieve accuracy across large data samples.

3.4 Cluster and Modularity Analysis

Gephi was utilised to analyse the network's modularity and community structure, providing insights into how users cluster within non-reciprocal networks. Key metrics included:

- **Community detection**: Gephi's modularity function identified tightly-knit clusters, or "communities," representing highly connected user groups.
- **Cluster density**: Cluster density was calculated as the proportion of edges within the same modularity class given by:

  $$Density = \frac{\#\ of\ Within\ Cluster\ Edges}{Total\ Number\ of\ Edges}$$

  This metric allowed us to quantify the network's internal connectivity, showing how densely connected communities were compared to cross-cluster connections.

This analysis provided a deeper understanding of how non-reciprocal networks organise around community structures, highlighting differences from reciprocal networks like Facebook.

3.5 Eigenvector Centrality and Influence Analysis

To identify influential nodes, we calculated Eigenvector Centrality, which scores nodes based on their connections to other high-scoring nodes. This metric, processed via Python, highlighted key "bridge" nodes facilitating connections between communities.

- **Ranking Nodes by Influence:** Nodes were ranked by eigenvector centrality and degree, revealing differences in network impact between high-degree nodes and highly influential nodes.

Eigenvector Centrality provided insights into the distribution of influence in non-reciprocal networks, allowing us to understand how certain users might act as ‘bridges’ between communities, a factor not always captured in traditional, reciprocal-based SNA tools.

3.6 Visualisation and Interpretation

Final visualisations in Gephi and Python clarified the network’s structure and connectivity patterns:

- **Cluster Structures:** Gephi’s visualisations depicted modular clusters, providing a clear view of community organisation.
- **Degree Distribution and Centrality Graphs:** Python-generated graphs illustrated degree distribution and centrality metrics, facilitating comparisons with traditional six-degree models.

These visualisations synthesised the data, presenting findings in an easily interpretable format and supporting comparisons to established social network frameworks.

4. Experimental Setup

This section provides the technical setup and resources used to ensure reproducibility. It includes information on the datasets, software tools used, and processes followed for data analysis and visualisation so that third parties can easily recreate the experiment accurately if they so choose.

4.1 Dataset

The primary dataset used in this project was the Stanford Large Network Dataset Collection (SNAP), specifically, a Twitter dataset that includes non-reciprocal connections between approximately 80,000 users. This dataset was ideal for analysing the follower-following relationships in a large-scale social network, especially for users who may not be closely connected, facilitating the exploration of the ‘Six Degrees of Separation’ theory in a non-reciprocal network. Key components of the dataset were:

- **Nodes**: Represent individual Twitter users, each with a unique ID.
- **Edges**: Directed connections representing non-reciprocal relationships (i.e., User A follows User B, but B may not follow A).

The dataset was downloaded in a raw format and processed into Excel-compatible CSV files for nodes and edges to allow for seamless integration with network analysis software.

4.2 Software and Tools

The following software and tools were used in the analysis and visualisation process:

- **Python**: The primary language for data processing and analysing, utilising key libraries:
    - **NetworkX** for constructing and analysing the network graph, calculating shortest paths, degree distribution, and eigenvector centrality.
    - **Pandas** for data processing and formatting.
    - **Matplotlib** for visualising metrics such as degree distribution and eigenvector centrality.

- **Gephi**: Used for modularity analysis and community visualisation, which allowed for identifying clusters and community structures within the network. Gephi’s advanced modularity and community detection features provided a detailed visual representation of cluster density and connectivity to help with analysis.

4.3 Experiment Process

The experiment followed structured steps to ensure reproducibility:

- **Data Import and Initial Visualisation**: The SNAP Twitter dataset was formatted into a single CSV file with source and target columns, representing directed follower-following relationships. This file was imported into Gephi and visualised using the **ForceAtlas2** layout, which provided an intuitive depiction of the network structure.

- **Modularity and Community Detection**: Gephi’s modularity algorithm was applied to identify clusters, with modularity classes colour-coded to highlight distinct communities. This visualisation allowed for an initial overview of the network’s community structure.
- **Data Export for Further Analysis**: After completing the graphic visualisation and modularity analysis in Gephi, the data was exported into two separate files: **nodes** (with modularity and other calculated attributes) and **edges** (with source and target relationships). These files were then processed in Python for detailed analysis.
- **Network Construction in Python**: Using NetworkX, the nodes and edges data were loaded into Python to reconstruct the directed graph. This graph served as the foundation for calculating additional network metrics, including degree distribution and eigenvector centrality.
- **Degree of Separation Calculation**: Custom Python scripts approximated degrees of separation by calculating shortest paths between randomly selected node pairs, efficiently validating the Six Degrees of Separation theory.
- **Eigenvector Centrality**: Eigenvector centrality scores were calculated in Python using NetworkX to identify influential nodes within the network. These scores were compared with Gephi’s modularity visualisations to confirm the influence distribution across different clusters.
- **Visualisation**: Both Gephi and Matplotlib were used for visualisation. Gephi provided community structure visuals, while Matplotlib generated plots for metrics such as degree distribution and eigenvector centrality.

4.4 Reproducibility

All scripts, data-processing steps, and visualisations are documented and stored in a **GitHub repository** here for easy access and replication. The repository contains:

- Python scripts for data processing, pathfinding, and centrality calculations.
- CSV files formatted for use with NetworkX and Gephi.
- Gephi project files with settings and configurations for modularity and community detection.
- Various other information.

This setup allows third parties to replicate the study with ease, following the same steps outlined here to analyse non-reciprocal connections within a large-scale social network for themselves.

5. Results

This section presents an in-depth analysis of the social network using various network metrics and visualisations. Insights drawn from these metrics help clarify the network’s structural dynamics, community organisation, and influential nodes, with comparisons to patterns typically seen in reciprocal social networks.

5.1 Degrees of Separation and Network Diameter

To assess the "Six Degrees of Separation" theory within this network, we calculated the average path length and approximate network diameter using a sample of 100,000 node pairs. Due to the large dataset—comprising of over 81,000 nodes and 1.7 million edges—sampling was necessary to obtain a manageable yet representative estimate. Calculating path lengths across all node pairs would have been computationally expensive, so this sample allowed us to approximate these metrics efficiently.

Table 1: Basic Network Properties

Metric	Value
Total Nodes	81,306
Total Edges	1,768,149
Approximate Average Path Length	3.89
Approximate Diameter	7

The results, shown in Table 1, indicate an approximate average path length of **3.89**. This suggests that, on average, any user in this network can reach another user in fewer than four steps, which closely aligns with the Six Degrees of Separation theory. However, as noted earlier in the literature review, this path length is slightly longer than the average of **3.57** observed on Facebook. This discrepancy is expected given Twitter's non-reciprocal network structure, where connections are often one-sided. Despite this, the relatively short path length underscores the high interconnectedness within this network, facilitated mainly by influential nodes with high degrees and eigenvector centrality scores.

The estimated network diameter is approximately **7**, indicating that the longest shortest path between any two nodes in this sample is seven steps. This value highlights the network’s compact structure, likely shaped by influential hubs that bridge otherwise disconnected parts of the network. However, it also shows cases where the six-degree rule is exceeded, which is reasonable given the non-reciprocal and potentially ‘niche’ nature of connections in this network. Together, the average path length and diameter demonstrate the efficiency of information flow across this network, reinforcing the presence of small-world characteristics in a non-reciprocal social platform.

5.2 Network Characteristics

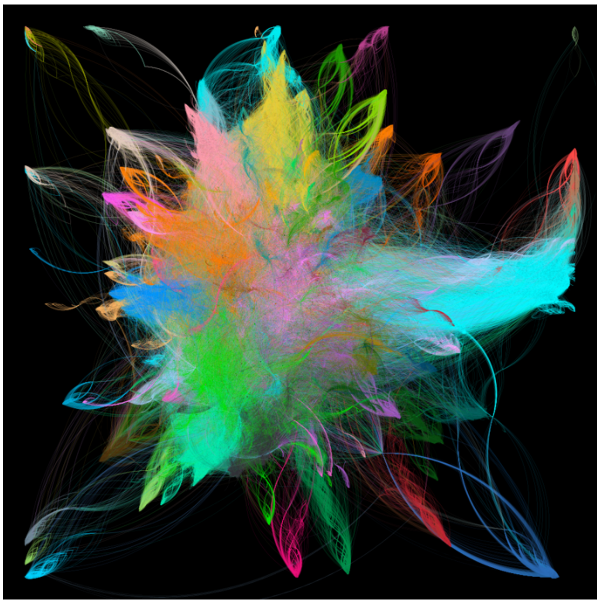
The network demonstrates a highly modular structure, with users clustering around shared interests and limited interaction across communities. Metrics such as cluster density (0.88) and modularity (0.8), as shown in Table 2, underscore this pattern, revealing a network where engagement predominantly occurs within tightly-knit groups. Each node averages 21.75 connections, indicating a densely connected network typical of non-reciprocal social platforms, where users often belong to communities centred around specific interests or affiliations.

Table 2: Community Structure & Clustering Metrics

Metric	Value
Within-Cluster Edges	1,558,733 (88.16%)
Cross-Cluster Edges	209,416 (11.84%)
Cluster Density	0.8816
Average Clustering Coefficient	0.5653
Modularity	0.8

The network’s structure is further visualised in Figure 1, where modularity classes are colour-coded, illustrating both large, central clusters, likely associated with popular topics or influential figures and smaller, isolated clusters that likely represent niche communities or local interest groups. The stark visual contrast between these dense central clusters and the more dispersed, peripheral clusters highlights the network’s heterogeneity in connectivity and engagement patterns.

Figure 1: Visualisation of network clusters.

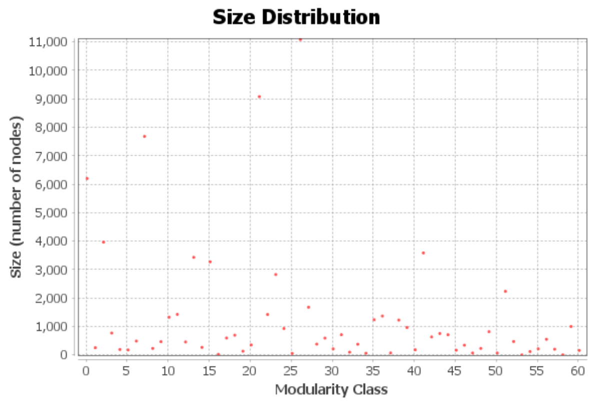


The predominance of within-cluster edges (88.16%) over cross-cluster edges (11.84%) emphasises this community-focused structure. This clustering behaviour suggests that users primarily interact within these cohesive groups, with few connections extending across clusters—an interaction pattern typical of non-reciprocal networks, where connections are more often based on shared interests rather than mutual relationships that would be found on platforms like Facebook.

5.3 Community Structure and Size Distribution.

The network's modular structure is evident in Figure 1, where nodes are organised into colour-coded clusters based on modularity classes, highlighting the distinct communities within this social network. Each modularity class, represented on the x-axis in Figure 2, ranges from 0 to 60 and corresponds to a unique community within the network. These modularity classes group nodes that are more densely connected to each other than to nodes outside the class, allowing us to identify clusters of users with shared connections.

Figure 2: Community Size Distribution in the Network



This plot shows the distribution of community sizes (number of nodes) across various modularity classes.

The distribution of community sizes in Figure 2 reveals significant variation across modularity classes. While some communities are densely populated, likely representing user groups with high interconnectivity, many smaller clusters suggest niche or localised interest groups. This pattern reflects the network's structure, where large communities often revolve around broadly popular topics or influential figures, while smaller clusters form around specific, specialised interests.

The predominance of large and small communities aligns with the network's non-reciprocal nature, as users gravitate toward both widely followed hubs and more niche communities. This diversity in community size, alongside high modularity and clustering metrics, reveals a network where some users actively engage in both extensive, central groups and more focused, peripheral clusters, while others interact primarily within either popular or niche communities exclusively.

5.4 Degree Distribution Analysis

The degree distribution plot in Figure 3 reveals a power-law pattern common in social networks: most nodes have only a few connections, while a small subset has a disproportionately high degree. This skewed distribution implies that most users are sparsely connected, while a few highly connected hubs play a central role in the network's structure. These hubs serve as bridges across communities, facilitating connectivity and information flow between otherwise separate clusters. Their presence is essential for maintaining the network's overall cohesion and stability.

Figure 3: Degree Distribution of Nodes in the Network

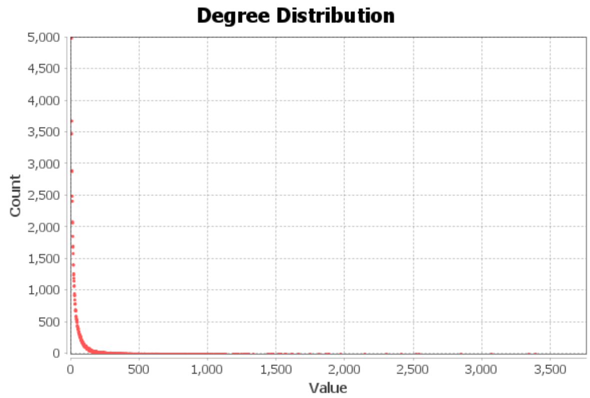


Figure 3 provides a visual representation of this distribution, with the x-axis showing node degrees (number of connections) and the y-axis indicating node frequency. The plot shows a high concentration of nodes with low degrees and a sharp decline as degree values increase, emphasising the network's dependence on a few highly connected users. This structure suggests that while most users interact within smaller circles, high-degree nodes, or “hubs,” are crucial for connecting distant clusters/parts of the network. These influential accounts facilitate communication and visibility across communities, enhancing the network’s overall connectivity and cohesion by bridging otherwise separate groups or clusters.

5.5 Eigenvector Centrality and Influence

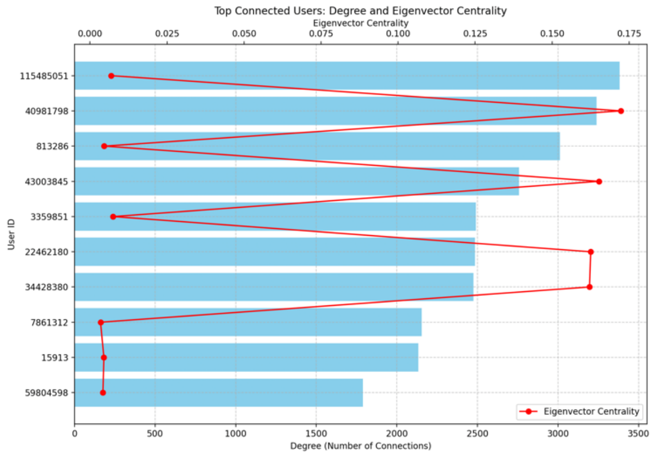
Eigenvector centrality is a measure used to identify influential nodes by considering both the number of connections a user has and the quality of those connections, particularly to other highly connected users. Unlike degree centrality, which only counts direct connections, eigenvector centrality assesses a node’s influence based on its position within the larger network structure. As shown in Table 3, the users with the highest eigenvector centralities are not always those with the highest degrees, indicating that influence in this network extends beyond sheer connectivity.

Table 3: Top 10 Connected Users by Degree with Eigenvector Centrality

Node (id)	Degree	Eigenvector Centrality
115485051	3383	0.0069
40981798	3239	0.1724
813286	3011	0.0045
43003845	2758	0.1652
3359851	2490	0.0075
22462180	2484	0.1625
34428380	2476	0.1621
7861312	2155	0.0034
15913	2133	0.0045
59804598	1789	0.0042

Table 3 highlights the Top 10 most connected users, comparing their degree and eigenvector centrality scores. Some high-degree users have relatively low eigenvector centrality, suggesting they are influential within their local clusters but lack network-wide impact. Conversely, users with high eigenvector centrality often connect to other influential nodes, extending their reach across multiple communities and enhancing their overall influence.

Figure 4: Eigenvector Centrality Distribution Amongst Top Users



The analysis reveals unique patterns of influence in this network. Nodes with high eigenvector centrality tend to bridge various parts of the network, connecting with other prominent users and amplifying their impact across clusters. Figure 4, a graphic visualisation of Table 3, illustrates the distribution of eigenvector centrality among top users, showing that only a select few nodes achieve notably high scores. These top-ranking users are critical in facilitating information flow, thereby enhancing connectivity and cohesion across otherwise distinct communities.

6. Conclusion

This study demonstrates that Twitter’s non-reciprocal social network largely adheres to the 'Six Degrees of Separation' theory, with an average path length of 3.89. While this aligns with the theory on average, the network diameter of 7 indicates that some paths exceed six degrees, highlighting Twitter’s unique structure. Compared to reciprocal networks like Facebook, where separation averages around 3.57 degrees, Twitter’s one-sided connections and reliance on bridging nodes/users result in slightly longer average distances. This research fills a critical gap by extending small-world theory to non-reciprocal networks, illustrating that efficient connectivity and small-world characteristics can persist despite the lack of mutual links.

7. References

[1]. Zhang, Z., Zhou, F., & Ning, H. (2022). When Six Degrees of Separation Meets Online Social Networks: How Low Can the Degree Be? *arXiv preprint arXiv:2209.06577v1*. Available at: https://arxiv.org/abs/2209.06577v1 [Accessed 14 Aug. 2024].

[2]. Brass, D.J. (2022). New Developments in Social Network Analysis. *Annual Review of Organizational Psychology and Organizational Behavior*, 9, pp. 225-246. Available at: https://www.annualreviews.org/doi/full/10.1146/annurev-orgpsych-012420-090628 [Accessed 15 Aug. 2024].

[3]. Zhang, L., & Tu, W. (2009). Six Degrees of Separation in Online Society. *WebSci09*, Tsinghua University, University College Cork. Available at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=1febc4cc5199d36e50ec7d3533a8ac8a32889d7c [Accessed 15 Aug. 2024].

[4]. Facebook Research, S. Edunov, S. Bhagat, M. Burke, C. Diuk, I. Onur Filiz (2016). Three and a Half Degrees of Separation. Available at: https://research.facebook.com/blog/2016/2/three-and-a-half-degrees-of-separation/ [Accessed 16 Aug. 2024].

[5]. SNAP. (n.d.). *ego-Twitter Social Network.* Stanford Large Network Dataset Collection. Available at: https://snap.stanford.edu/data/ego-Twitter.html

[6]. Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. *International AAAI Conference on Weblogs and Social Media.* Available at: https://gephi.org

[7]. Python Software Foundation. (n.d.). Python Language Reference, version 3. Available at: https://www.python.org/

[8]. Microsoft Corporation. (2024). Microsoft Excel. Available at: https://www.microsoft.com/en-us/microsoft-365/excel