Exploring Social Dynamics in Wikidata Editor Networks: A Network Analysis Approach

Akshay Saraf
Performed Task A and C
King's College London
k23039901@kcl.ac.uk

Pattarin Urapevatcharewan
Performed Task A and B
King's College London
k23036414@kcl.ac.uk

Han Gu
Helped with Task C Part A
King's College London
k23055630@kcl.ac.uk

Srinidhi Sharma
Performed Task D and Report Writing
King's College London
k23028061@kcl.ac.uk

ABSTRACT

UPDATED - September 2, 2024. This paper delves into the intricate social dynamics of Wikidata, a collaborative knowledge graph, by leveraging network analysis techniques. We construct a network representation of Wikidata editor interactions, modeling users as nodes and connections arising from shared contributions on discussion threads. The study employs various network metrics to unveil characteristic properties, distributions, and social behaviors within the Wikidata editor network. Shortest-path algorithms reveal insights into communication patterns, while comparisons with random networks provide a benchmark for understanding the uniqueness of Wikidata's social structure. The analysis extends to consider alternative connection criteria, exploring the impact on network characteristics. Through this exploration, we gain a deeper understanding of how editors engage with each other, fostering collaborative knowledge creation in projects like Wikipedia and Wikidata.

INTRODUCTION

In the ever-expanding landscape of collaborative knowledge projects, platforms like Wikipedia and Wikidata stand as pioneers, facilitating the collective creation and dissemination of information. Within these ecosystems, understanding the intricate social dynamics among contributors becomes paramount for unraveling the collaborative fabric that underpins knowledge creation. This paper embarks on a journey into the heart of Wikidata, a structured database akin to Wikipedia, to explore the social interactions among its editors. Leveraging network analysis, we unravel the underlying structure of the Wikidata editor network, seeking insights into communication patterns, social connections, and collaborative behaviors. By delving into the rich tapestry of interactions within this digital realm, our study aims to illuminate the mechanisms that drive collaborative knowledge projects and shed light on the unique characteristics that distinguish Wikidata's editor network. Through this exploration, we contribute to a wider understanding of how individuals engage and collaborate in shaping the collective knowledge that defines the digital era.

RELATED WORK

We want to include Stanley Milgram's work as it gave us the conceptual understanding of the subjedt and explored the concept of the small world, demonstrating that individuals are connected through surprisingly short paths. His experiment [5] involving email communication reveals the small-world phenomenon, suggesting that social networks are more interconnected than previously thought. Similarly, Leskovec and Horvitz may have not directly influenced our work but his investigation about the "6 degrees of separation" phenomenon using a large instant messaging network in their paper [3] is to look for. They find that communication patterns are influenced by factors such as age, language, and location, providing insights into the structure of social networks at a global scale which we indirectly use in one of our tasks. Tantardini et al. [4] compare network comparison methods, emphasizing the effectiveness of graphlet-based measures in analyzing differences between network structures. They illustrate this through real-world examples, highlighting the importance of considering diverse methods for network similarity analysis. Engineering the Emergence of Norms: A Review [2] reviews the emergence and evolution of social norms, emphasizing the transition from conventions to formal norms through conformity and sanctions. It discusses challenges and evaluates norms from individual and societal perspectives, including their formalization in legal systems. Granovetter's [1] paper discusses threshold models of collective behavior, highlighting the importance of individual thresholds in decision-making. The models challenge the idea that aggregate outcomes reflect shared norms, emphasizing the role of individual preferences in shaping collective behavior. Least Cost Rumor Community Blocking Optimization in Social Networks: The work [6] focuses on rumor community blocking optimization in social networks, aiming to achieve the desired effect with a larger running time. It explores strategies to block rumors effectively within social network communities. All these papers have helped us to shape our understanding about the particular subtopic they address.

The link to our GitHub repository:

https://github.kcl.ac.uk/k23039901/NDA_NetworkX. Our code can be found here.

METHODS

In the exploration of the Wikidata editor network, a methodology rooted in the representation of social connections using an undirected graph data structure was established. Each node in this structure corresponded to a distinct Wikidata user editor, while additional attributes captured crucial information such as thread subject (subject of the thread), page name (name of the page) and a comment date. This helped us store and access relevant data of every user and their interactions with other users.

The construction of the network involved a series of steps. Initially, users were categorized by page and thread, identifying those with social connections. An iterative process ensued, wherein each user was traversed to extract and assign relevant information, resulting in the creation of nodes for each unique user. Subsequently, edges were introduced between users who shared a social connection, signifying individuals who had commented on the same thread in the same page.

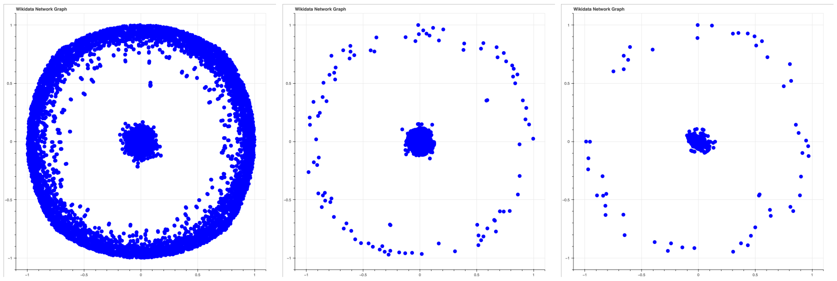
To evaluate the efficiency of the network construction process, the time required for varying sample sizes of the dataset was measured. This empirical analysis sought to untangle the algorithmic efficiency, providing insights into the time complexity concerning input size. The consideration of potential complexities, such as logarithmic, linear, or quadratic, enriched the understanding of the scalability and computational demands associated with the construction algorithm.

Concurrently, an Erdos-Renyi graph, a foundational random graph model, was generated to serve as a benchmark for subsequent analyses. This model facilitated the comparison of network metrics and distributions, offering valuable insights into the inherent features of the network. Inspiration of which was taken from Tantardini et al. [4]

Shortest path analyses were then conducted using Bellman-Ford and Dijkstra's algorithms, accommodating scenarios of both existing and non-existing paths. This approach unveiled the connectivity and efficiency nuances of the network, providing valuable insights into its structural intricacies.

Transitioning to the examination of basic network properties and metrics, parameters such as the number of nodes and edges, density, number of connected components, and the size of the largest connected components were considered. These metrics provided fundamental information about the network's size, complexity, and cohesion. Additional metrics, including average clustering coefficient and average degrees of nodes, offered insights into clustering, influence, and overall connectivity within the network.

In our network analysis, we employed both the Threshold and SIR models to gain insights into how opinions propagate among editors. A comparative study was then conducted to discern the differences between these two models. For the initialization of infected nodes, we adopt a strategic approach by selecting users from pages with the most, median, and least comments overall, rather than employing a random selection method. This allows us to explore the worst, average, and best-case scenarios for both models, offering a comprehensive understanding of their respective performances.



(a) Users (b) Administrators (c) Requests_For_Comment

Figure 1: The Three Created Network Graphs

		Cost and Time	
Time	*Size*	*Size/Time*	*Dataset*
27.95	40007.46	1431.39 Kb/sec	USERS(Large)
2.25	4847.98	2154.66 Kb/sec	ADMINISTRATORS(Medium)
0.36	797.62	2215.61 Kb/sec	REQUESTS_FOR_COMMENT(Small)

Table 1: Cost and Time

In the Threshold model, the status of a node is being determined by assessing the statuses of its neighbors, reflecting the influence of immediate connections. On the other hand, the SIR model relies on infection probabilities to ascertain the status of a node, introducing a probabilistic element into the analysis.

To enhance the efficiency of detection, a detection list is curated based on the overall centrality of users across the entire network. This evaluation combines degree centrality and proximity centrality to gauge the influence of editors within the community. A prioritized inspection list is then generated, focusing on the top 10 editors with the highest overall centrality values. These editors are considered to have significant influence and importance in the network, making them key targets for scrutiny in case of potential "spoofing." This strategic prioritization ensures that if there are instances of misinformation or manipulation, these influential editors are promptly investigated, given their potential role as central information disseminators or individuals significantly affected by such occurrences.

RESULTS

We successfully constructed three Wikidata editor networks of varied sizes : USERS(Large), ADMINISTRATORS(Medium), and REQUESTS_FOR_COMMENT(Small). All three networks are represented using an undirected graph each that efficiently stores nodes and edges along with their associated attributes (thread_subject, pag_name, comment_date). The construction of the network involved grouping users by page and thread, iterating over the dataset to extract relevant information, adding nodes and edges to the graph. See Figure 1 on the bottom of the Page 2.

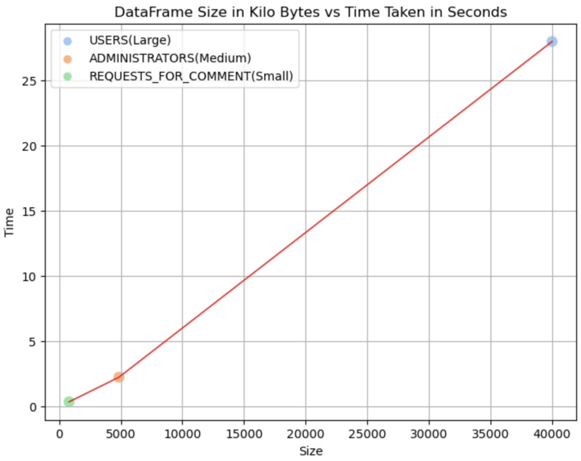


Figure 2: Size and Time Comparision

From Table 1 (bottom of the Page 2) and Figure 2, we can see that the time taken decreases as the input size increases, which indicates an INVERSE relationship. We are taking into account Blocking Optimization in Social Networks [6] to perform Therefore, it's likely that the algorithm has a logarithmic time complexity, expressed as $O(\log n)$, where n is the input size in kilo bytes.

In the exploration of node-level descriptors across networks, key metrics were analyzed to provide a comprehensive understanding of the structural characteristics and dynamics within the network.

The density of each network indicates the proportion of actual connections to possible connections. The requests for comment network has the highest density (0.0700), suggesting a relatively high level of interconnectedness. Followed by, administrators network at 0.0047. Meanwhile, the users network has a much lower density (0.0004), indicating a sparser and more decentralized network structure among users.

The presence of multiple connected components in the users network (3,262) suggests a higher level of fragmentation or isolation among editors. This may indicate the existence of distinct subgroups or communities within the larger editor community, possibly based on interests, expertise, or editing patterns. In contrast, the requests_for_comments has the least number of connected components (54), indicating a more cohesive and unified group.

The clustering coefficient measures the degree to which nodes in a network tend to cluster together. The higher clustering coefficients observed in the administrators network (0.6195) and requests_for_comments network (0.7818) suggest a higher level of local clustering or community structure among administrators and users engaging in discussions or providing feedback. This indicates the presence of tightly-knit groups or communities within these networks.

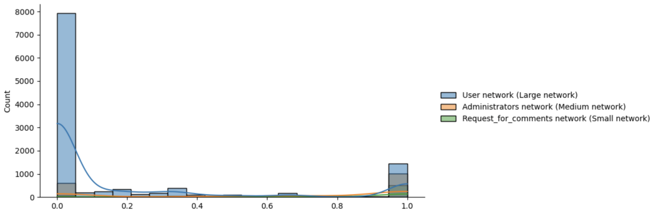


Figure 3: Comparing Clustering Coefficient

Another crucial metric investigated was the clustering coefficients distribution. See Figure 3. This analysis entailed visualizing the prevalence of clustering coefficients, aiding in the identification of cohesive subgroups or communities within the network. Nodes with higher clustering coefficients tend to form tightly interconnected groups, indicative of potential community structures. This exploration dives into the cohesive nature of the network, revealing patterns of inter-connectivity that may contribute to a refined understanding of its overall organization.

The Figure 4 illustrates the degrees distribution which involved visualizing the spread of node degrees. The degree of a node signifies the number of connections it holds, and by scanning the distribution, we gain insights into the connectivity patterns across nodes. Nodes with higher degrees may indicate heightened influence or increased activity within the network, thus spotlighting potential hubs of significance.

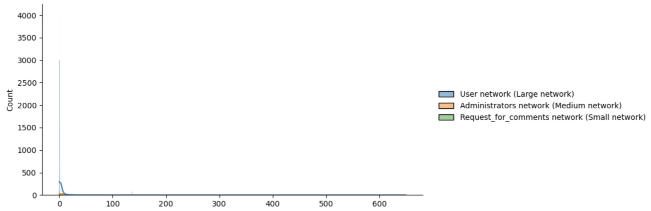
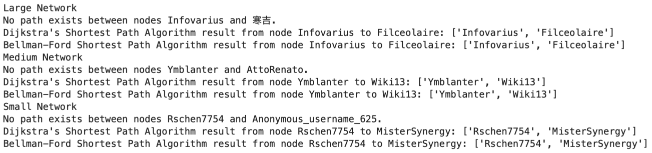


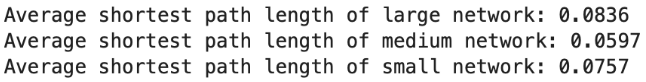
Figure 4: Comparing Degree Distribution



```
Large Network
No path exists between nodes Infovarius and 寒吉.
Dijkstra's Shortest Path Algorithm result from node Infovarius to Filceolaire: ['Infovarius', 'Filceolaire']
Bellman-Ford Shortest Path Algorithm result from node Infovarius to Filceolaire: ['Infovarius', 'Filceolaire']
Medium Network
No path exists between nodes Ymblanter and AttoRenato.
Dijkstra's Shortest Path Algorithm result from node Ymblanter to Wiki13: ['Ymblanter', 'Wiki13']
Bellman-Ford Shortest Path Algorithm result from node Ymblanter to Wiki13: ['Ymblanter', 'Wiki13']
Small Network
No path exists between nodes Rschen7754 and Anonymous_username_625.
Dijkstra's Shortest Path Algorithm result from node Rschen7754 to MisterSynergy: ['Rschen7754', 'MisterSynergy']
Bellman-Ford Shortest Path Algorithm result from node Rschen7754 to MisterSynergy: ['Rschen7754', 'MisterSynergy']
```

Figure 5: Shortest Path Algorithm

From the output in Figure 5, it can be seen that Dijkstra's Shortest Path algorithm and Bellman-Ford algorithm give us the same output. The reason for this is that our network is undirected and does not contain any negative weight.



```
Average shortest path length of large network: 0.0836
Average shortest path length of medium network: 0.0597
Average shortest path length of small network: 0.0757
```

Figure 6: Average shortest Path Algorithms

From Figure 6, it can be seen that the medium-size network (administrators) has the lowest average shortest path length (0.0597) among the three, indicating that, on average, nodes are very close to each other. This suggests highly efficient communication and connectivity within the network. Meanwhile, the large network (users) has the highest average shortest path length (0.0757) among the three, suggesting that, on average, nodes are slightly farther apart compared to the medium and small networks. This could indicate slightly less efficient communication pathways or less direct connections between nodes.

The examination of the real network in comparison to a random network has revealed several noteworthy distinctions across various criteria. See Figure 7. The degree distribution of a random network is normally distributed as expected and as seen in Figure 7, but for our networks they appear to be bimodal. Firstly, both networks share identical numbers of nodes, emphasizing a consistent foundational structure. Despite slight variations in the number of edges, the similarity in densities between the two networks suggests comparable levels of connectivity relative to the total potential connections.

However, the real network consistently diverges from the random network in specific aspects. Notably, the higher number of connected components in the real network indicates a greater degree of fragmentation and isolation compared to the random network. Despite both networks possessing numbers of connected components, the number of nodes in real network's largest component tends to be smaller, signifying a less cohesive overall structure.

The real network consistently exhibits significantly higher average clustering coefficients than its random counterpart. This observation implies a greater propensity for nodes in the real network to form tightly-knit clusters or groups, underscoring distinctive patterns of local connectivity. Although the average degree of nodes in both networks is similar, subtle variations suggest potential differences in their connectivity patterns.

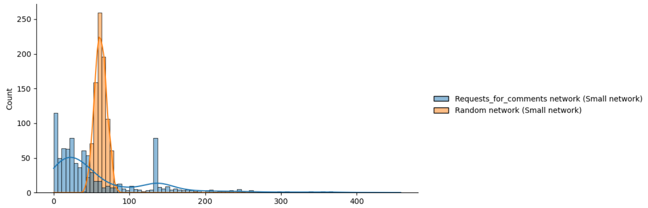


Figure 7: Comparing Degree Distribution of the Random Network and Request for Comment Network

The diameter of the largest connected component in the random network tends to be similar to that of the real network, indicating comparable maximum distances between nodes within this component. In some cases, the diameter may be equal, slightly higher, or lower in each network size. This disparity highlights distinct structural characteristics, emphasizing potential variations in the efficiency of information flow and connectivity patterns between the two networks.

Additionally, when considering the average shortest path length, the real network mostly exhibits a shorter average shortest path length compared to that of the large random network. This comparison further elucidates differences in the efficiency of information dissemination and communication within the networks. In conclusion, the analysis underscores notable differences in terms of connectivity, cohesion, and local clustering patterns. These insights contribute to a nuanced understanding of how the real network deviates from the random model across multiple structural dimensions.

Before delving into the detailed analysis of the networks, we have also created a new network and it's crucial to acknowledge a significant distinction between the old and new networks. The original network connected editors if they commented on the same thread within the same page, whereas the new network establishes connections when editors comment on the same page, irrespective of the specific thread. This shift in connectivity criteria fundamentally alters the structure of the network, influencing connectivity patterns and communication pathways among editors. With this understanding, we can now explore the characteristics and dynamics of both the old and new networks in greater detail.

The average shortest path length as seen in Figure 6 provides insights into the overall efficiency of information flow and communication within the network. The relatively low average shortest path lengths observed in all three networks

(0.0836 for users, 0.0597 for administrators, and 0.0757 for requests_for_comments) indicate that it takes only a few steps to navigate between any pair of editors. This suggests a high degree of connectivity and efficient communication pathways within these networks.

The average shortest path length remains unchanged between the original and new requests_for_comments networks, both recording a value of 0.0757. This suggests that, despite potential changes or updates in the network, the overall efficiency of communication and information flow between nodes remains consistent.

For the USERS.csv and ADMINISTRATORS.csv networks, we can observe that the number of edges increase, increasing connectivity in the new network. The new networks are more dense comparing to the density of the original networks. Additionally, there are significant reduction in fragmentation and isolation in the new networks compared to the original networks. From the number of largest connected component and average clustering coefficient values, they indicate that there are larger cohesive group of in the new networks and higher level of local clustering or community structure among administrators in the new network, respectively. The diameter of the network also decrease which means the maximum number of steps to travel to other nodes are decreased, making them easier to connect and improved overall network efficiency in the new network. The average path length in both original users network and new users network remain unchanged, while the new administrators network's average path length increases.

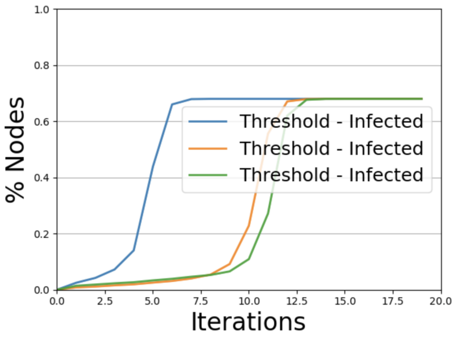


Figure 8: Threshold Model of Network 1

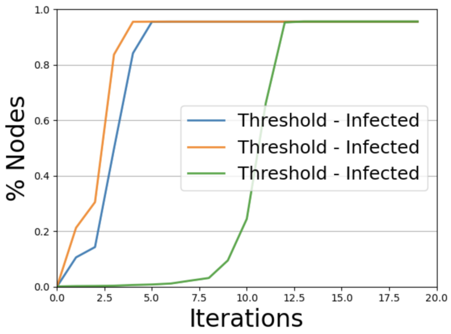


Figure 9: Threshold Model of Network 2

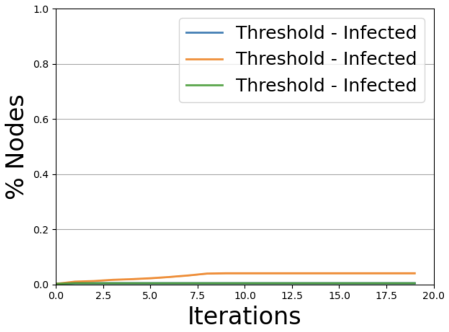


Figure 10: Threshold Model of Network 3

To comprehend Granovetter's [1] work , we employed the threshold model on network data and simulated behavior propagation starting from initially infected editors (those with the most comments). Subsequently, we observed how this behavior disseminated through the network over time and evaluated whether it reached neighboring similar editors (those with similar commenting patterns, i.e., users who commented on the same page and thread). The threshold model results revealed significant disparities in behavior spread across different networks and scenarios (worst vs. median vs. best). In the Users Network (Figure 8), the behavior spread to 70 % of users within 6 iterations in the worst-case scenario, while it took over 12 iterations in the best-case scenario. This underscores the critical role of initially infected nodes and their influence on behavior propagation throughout the network. By comparing Figure 8, Figure 9 and Figure 10 we can discern varying levels of influence and susceptibility within each network. These differences in spread suggest diverse network structures and node connectivity among nodes in each network.

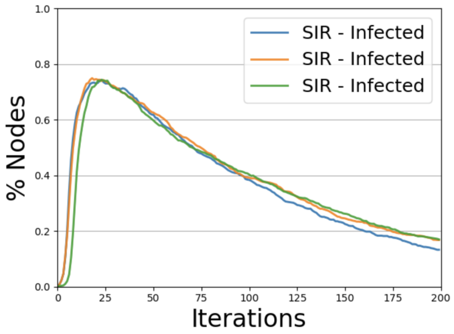


Figure 11: Diffusion Trend Comparison

Unlike Threshold model, In the SIR model, individuals are classified into three states: susceptible (S), infectious (I), and recovered (R). It assumes that individuals transition from susceptible to infectious to recovered, with no possibility of relapse. By simulating the dynamics of the SIR model on the network data, we have observed how the behavior spreads over time and how many neighboring editors become infected. By looking at the Trend Comparision (Figure 11) and Diffusion Prevalence Comparison (Figure 12) for the Request for Comment Network, we can see that the spread is almost identical

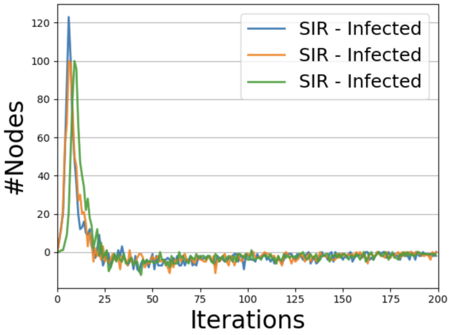


Figure 12: Diffusion Prevalence Comparison

in the Worst, medium and best case scenarios. In more than 75 % of cases, users adopt the behavior influenced by two other users, leading to a steep increase in the number of recovered users after 25 iterations. Subsequently, the spread of behavior diminishes almost entirely.

Based on the Threshold and SIR models, the Wikimedia Foundation can closely monitor user behavior and identify those most vulnerable to being affected next. By isolating these users promptly, the spread of behavior can be effectively contained. Additionally, centrality measures such as Degree Centrality and Closeness Centrality provide insights into nodes capable of efficiently disseminating information within a graph. Analyzing the network structure can help identify editors who bridge different communities or clusters, as these individuals may have a higher likelihood of spreading trolling behavior across various parts of the network. By isolating them, controlling the spread of behavior from one cluster to another becomes more manageable.

DISCUSSION

Overall, the Wikidata editor network exhibits structural characteristics that align closely with those of small-world networks rather than random or regular networks. This classification is primarily supported by key metrics such as the average clustering coefficient, the average shortest path length, and the length of diameter. The network demonstrates a high average clustering coefficient, indicating a propensity for nodes to form tightly-knit clusters or groups. Simultaneously, it showcases relatively short average shortest path lengths, suggesting that most nodes can be reached from any other node in a small number of steps.

This combination of high local clustering and efficient global connectivity is characteristic of small-world networks, where nodes are highly interconnected, yet the overall structure retains elements of both local clustering and global connectivity. Moreover, the Wikidata editor network's tendency to display the "small-world phenomenon," where nodes are highly clustered yet closely connected, further reinforces its classification within the small-world network group. Overall, the wiki network's structural properties and organizational dynamics closely resemble those of small-world networks, highlighting its position within this network continuum.

Additionally, the diameter of the Wikidata editor network further supports its classification as a small-world network. A smaller diameter implies shorter maximum distances between nodes, reinforcing the network's efficient communication pathways and highlighting its cohesive structure.

However, despite these insights, potential issues impacting the quality of the network must be considered. Bias stemming from sampling methods may lead to incomplete community representation, while identity-related challenges may arise from subjective decisions in establishing edges, potentially misrepresenting connections. Like Hayens mentioned in his work in A Review [2], the subjective nature of social decisions and potential biases in data collection processes introduce uncertainties, influencing the outcomes of network analyses.

To navigate these challenges and gain a nuanced understanding of link quality and identity, a combination of network metrics proves effective. Centrality measures, including betweenness and closeness centrality, help identify influential nodes, while clustering coefficients reveal cohesive subgroups. PageRank offers insights into the influence of specific editors, and community detection algorithms, such as Louvain or Girvan-Newman, unveil natural community structures. This multifaceted approach provides a comprehensive evaluation of link quality and identity representation, recognizing the dynamic and subjective nature of social networks. It is essential to acknowledge that while these metrics offer valuable insights, no single metric is foolproof, and a holistic analysis requires considering multiple perspectives.

CONCLUSION

In summary, our exploration into the social dynamics of Wikidata through network analysis has revealed distinctive characteristics of collaborative knowledge creation. By modeling editor interactions as a network, we gained insights into the structural properties and social behaviors inherent in Wikidata's collaborative environment. The application of diverse network metrics and shortest-path algorithms unveiled communication patterns and highlighted the network's uniqueness. Comparative analyses with random networks provided a benchmark, emphasizing the distinct social structure of Wikidata. Furthermore, our investigation into alternative connection criteria deepened our understanding of the network's adaptability.

In the field of collaborative knowledge projects such as Wikipedia and Wikidata, our findings contribute to a significant understanding of how editors engage and collaborate. The complex interplay of social dynamics and the adaptability of the network to different interaction criteria underscore the intricate nature of collaborative knowledge creation. This study serves as a valuable exploration into the social fabric of Network Data Analysis for Wikidata, offering insights that have implications for fostering effective collaboration and knowledge dissemination in similar online platforms.

ACKNOWLEDGEMENT

We thank Dr. Albert Meroño Peñuela for giving us the opportunity to contribute to the topic of Network Data Analysis and expand our knowledge.

REFERENCES

[1] Mark Granovetter. 1978. Threshold Models of Collective Behavior. *Amer. J. Sociology* 83, 6 (1978), 1420–1443. http://www.jstor.org/stable/2778111

[2] Chris Haynes, Michael Luck, Peter McBurney, Samhar Mahmoud, Tomáš Vítek, and Simon Miles. 2017. Engineering the emergence of norms: a review. *The Knowledge Engineering Review* 32 (2017), e18.

[3] Jure Leskovec and Eric Horvitz. 2008. Planetary-Scale Views on an Instant-Messaging Network. (2008).

[4] Ieva F. Tajoli L. et al. Tantardini, M. 2019. Comparing methods for comparing networks. *Sci Rep* 9 (2019), 17557. DOI: http://dx.doi.org/10.1038/s41598-019-53708-y

[5] Jeffrey Travers and Stanley Milgram. 1969. An Experimental Study of the Small World Problem. (1969).

[6] Jianguo Zheng and Li Pan. 2018. Least Cost Rumor Community Blocking optimization in Social Networks. In *2018 Third International Conference on Security of Smart Cities, Industrial Control System and Communications (SSIC)*. 1–5. DOI: http://dx.doi.org/10.1109/SSIC.2018.8556739