Directed Network Analysis on the Slashdot Social Network Using NetworkX and Gephi

Yusuf Semih Çiftçi *Student Number: 211015056* yusufsemihcifci2@gmail.com
Duzce University *Düzce/Turkey*

1. Introduction

The primary aim of this project is to examine friendship and hostility relationships between individuals using network analysis methods, to reveal the general structure of these relationships, and to investigate what the dataset can tell us about the social structure. In particular, answers have been sought for questions such as who is friends with whom, who harbors negative relationships towards each other, and the distribution of these structures within the network.

The dataset used in this study is the Slashdot social network data provided within the Stanford Network Analysis Project (SNAP). This dataset represents the social connections defined between users on a technology-focused news website called Slashdot. Each connection is formed by one user labeling another as a "friend" or "foe." The connections are directed; a user defining another user as a friend (1) or foe (-1) makes the network directed.

The original dataset includes approximately 70,000 nodes and around 500,000 connections. However, to conduct the analysis process more efficiently, a smaller subset of the dataset containing 10,000 nodes was used in this study.

This project aims to provide significant clues for understanding the structures of positive and negative relationships in social networks and to demonstrate the power of network science techniques on such data.

2. Dataset Description

The dataset used in this study represents a directed and signed social network structure that includes user interactions on the social news platform called Slashdot. Although the original dataset contains approximately 77,000 nodes and 516,000 edges, a subset consisting of 10,044 nodes and 28,305 edges was used in this study to make the analyses more manageable and computationally efficient.

Statistic	Value
Number of Nodes	10044
Number of Edges	28305
Number of Positive Edges	21997
Number of Negative Edges	6308

The dataset consists of three columns:

* `source` (the user initiating the connection),
* `target` (the user the connection is directed towards),
* `weight` (the sign indicating whether the relationship is positive or negative: +1 = friend, -1 = foe).

The dataset was initially obtained in `.txt` format. To make it usable on Gephi, this file was converted to `.csv` format. During the conversion, comment lines (those starting with the # character) were excluded from the analysis. Although there is no missing information in the dataset, necessary checks were included in the code against the possibility of missing or corrupted data during the data processing.

```
# Read the file line by line and add edges to the graph
with open(file_path, 'r') as file:
    for line in file:
        if line.startswith("#"):
            continue

        parts = line.strip().split()
        if len(parts) != 3:
            continue

        try:
            source, target, sign = int(parts[0]), int(parts[1]), int(parts[2])
            G.add_edge(source, target, sign=sign)
        except ValueError:
            continue
```

There are no additional attributes associated with the nodes or edges in the dataset; the analysis is based solely on the positive/negative relationships between users. Below is a sample excerpt from the dataset:

source	target	weight
0	1	1
0	2	1
0	3	-1

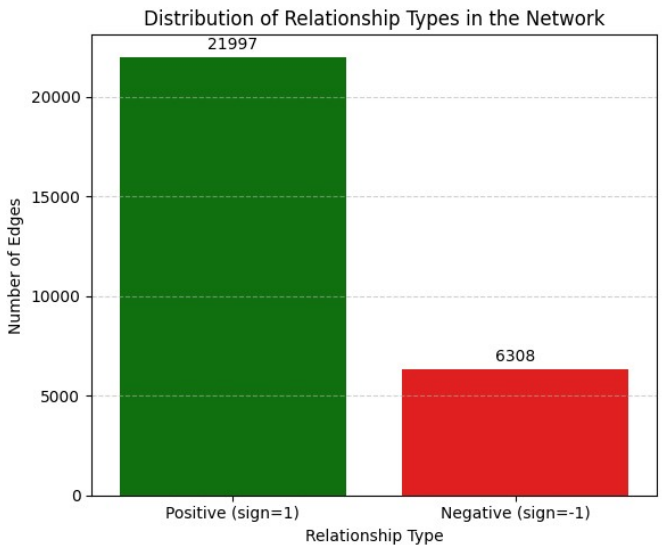
3. Methodology

In this project, network analysis was carried out both programmatically using the NetworkX library in Python and visually and statistically through the Gephi software. This allowed for a comparative evaluation of the analyses performed with different tools.

3.1 NetworkX Applications

As a first step, the Slashdot social network data was created as a directed graph structure. Subsequently, various structural analyses were performed:

Basic Statistics: Number of nodes and edges, average degree, positive/negative edge distribution (21,997 positive, 6,308 negative).



Connectivity Analysis:

- Number of weakly connected components (WCC): 323
- Size of the largest WCC: 9,340 nodes
- Number of strongly connected components (SCC): 7,557
- Size of the largest SCC: 2,204 nodes

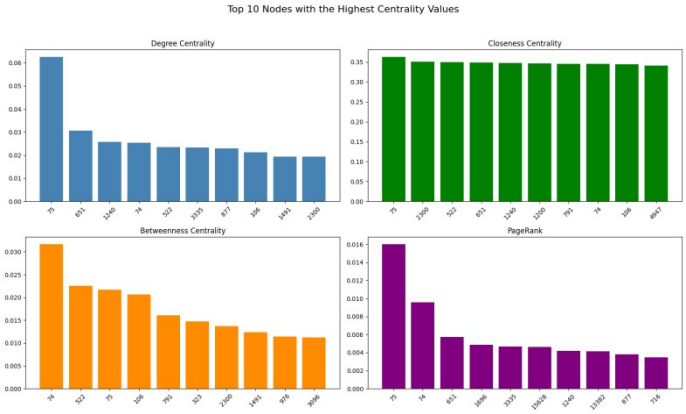
Structural Properties:

- Diameter: 12
- Average path length: 4.285
- Clustering coefficient: 0.031
- Number of triangles: 7,489
- Reciprocity: 0.174
- Assortativity: -0.110

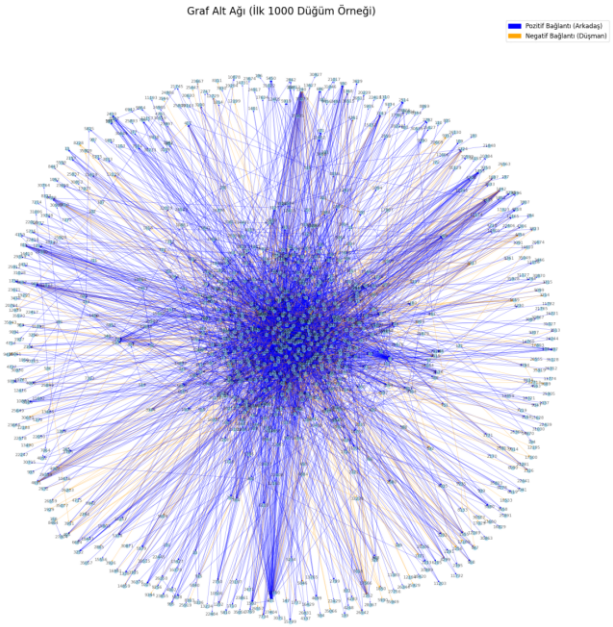
Community Detection: A total of 432 communities were detected using the Louvain algorithm. The largest community consisted of 989 nodes, while the average community size was calculated as 20.62. The modularity value is 0.510.

Social Balance Theory (Triad Analysis): 4252 triad combinations obtained from a sample of 1000 nodes were analyzed. Of these triads, 92.10% have a balanced structure, and 7.90% have an unbalanced triad structure. This indicates that the network generally has a balanced social structure.

Centrality Measures: Degree, closeness, betweenness, and PageRank centrality metrics were calculated, and it was observed that some nodes have high scores in all metrics.



3.2 Visual and Statistical Analysis with Gephi



In addition to the analyses performed with NetworkX, the data was converted to `.csv` format, imported into Gephi, and the following operations were carried out:

Layout Adjustment: Visual arrangement was performed using the ForceAtlas2 algorithm. The readability of the network was enhanced by optimizing the repulsion, gravity, and scaling parameters.

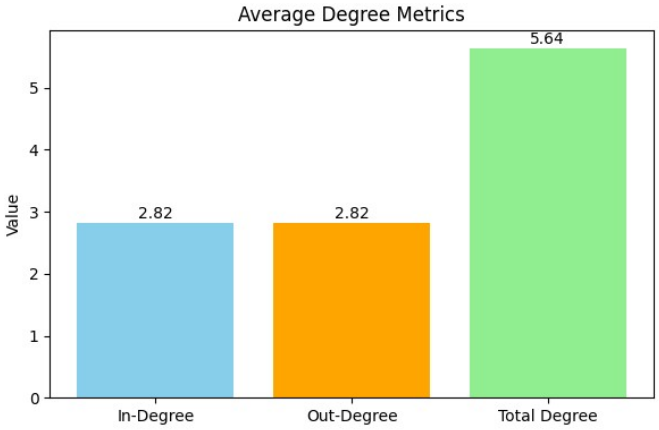
Community Detection: Communities were visualized and distinguished with colors using Gephi's Louvain algorithm. The modularity value was calculated as 0.711, indicating that the network contains quite distinct community structures.

Centrality Analysis:

- Nodes were classified by color coding according to closeness centrality. Nodes with the highest centrality are generally located in the central regions of the network.

General Network Properties:

- Average degree: 2.818
- Average weighted degree: 1.562
- Network diameter: 15
- Average path length: 5.282
- Clustering coefficient: 0.021
- Network density: 0.000



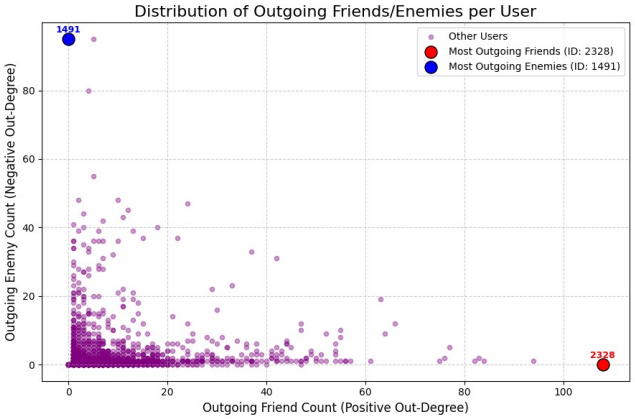
This dual analysis method enabled strong and meaningful inferences to be drawn from both numerical data and visual structures.

4. Results

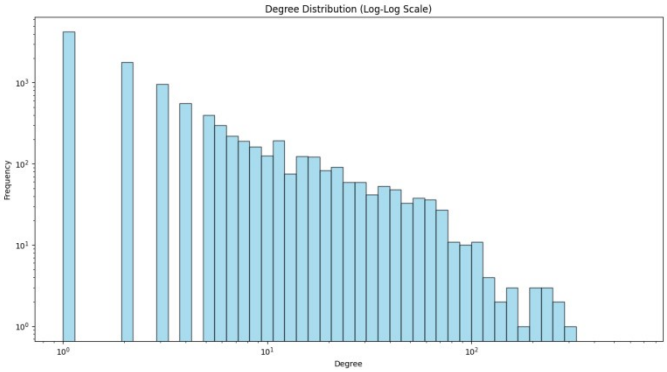
This section summarizes the key findings obtained from the analyses conducted using both NetworkX and Gephi. Numerical metrics and visual inspections are evaluated together.

4.1 Key Findings from NetworkX Analysis

Nodes and Edges: The analyzed network consists of 10,044 nodes and 28,305 directed edges.



Relationship Types: Of the connections, 21,997 are positive (friendship) and 6,308 are negative (hostility), indicating a dominance of positive relationships in the social structure.



Top 5 Most Frequent Degrees

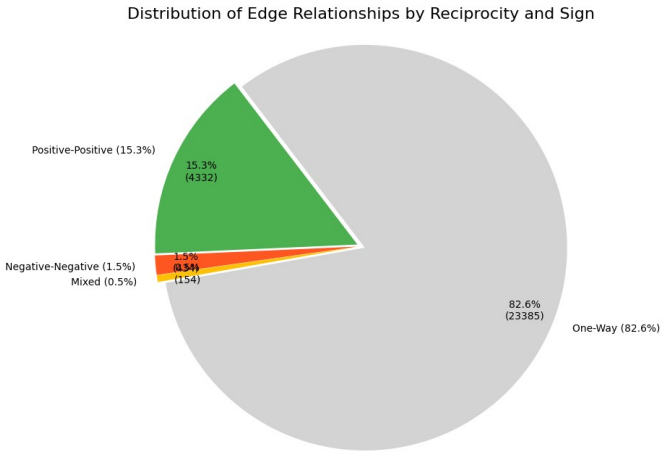
Degree	Frequency
1	4207
2	1795
3	958
4	554
5	401

Connectivity Structure:

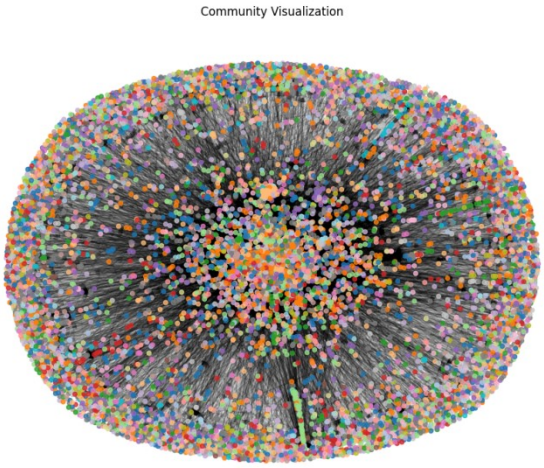
- A total of 323 weakly connected components (WCCs) were identified.
- The largest WCC includes 9,340 nodes, comprising approximately 93% of the total nodes.
- The number of strongly connected components (SCCs) is 7,557, with the largest SCC consisting of 2,204 nodes.

Structural Properties:

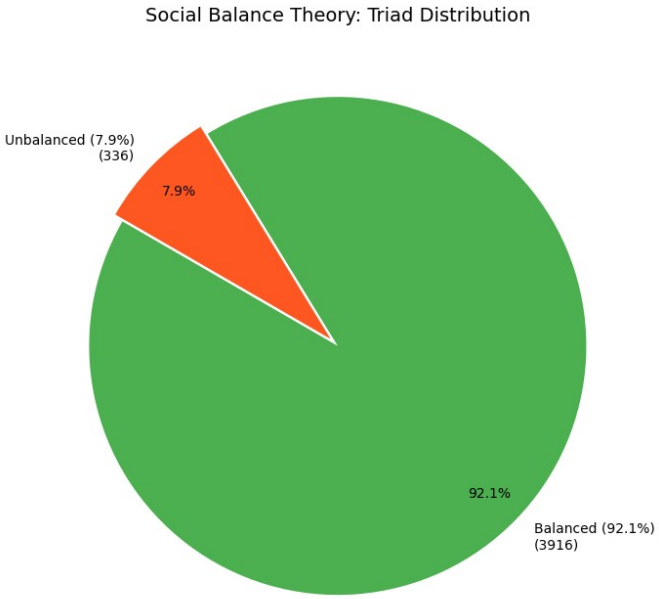
- The diameter of the network is 12, and the average path length is 4.285.
- The average clustering coefficient is low (3.1%), suggesting limited local connectivity.
- The reciprocity rate was measured at 0.174, meaning only a small fraction of the connections are bidirectional.



Communities: 432 communities were detected using the Louvain algorithm, with a modularity value of 0.510.

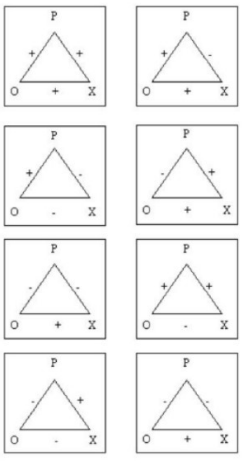


Balance Theory (Triad Analysis): 4,252 triads from a sample of 10000 nodes were analyzed. A balanced structure was observed in 92.10% of these, indicating that the social network largely contains stable relationships.



Heider's Balance Theory

Heider's Balance Theory suggests that individuals strive for cognitive consistency in their relationships with others and with attitude objects. The theory focuses on triadic relationships among three elements: a person (P), another person (O), and an object or idea (X). Each connection is either positive (+) or negative (−), and the relationship is considered balanced if the product of all three signs is positive. When imbalance occurs, it creates psychological discomfort motivating the individual to adjust their attitudes or perceptions to restore harmony. This theory helps explain how people manage conflicting social and cognitive relationships.



Relation to Triad Analysis in Network Science
In the context of network science, triad analysis involves examining groups of three connected nodes to assess the overall structural balance of a signed network. Applying Heider's Balance Theory, a triad is considered balanced when the relationships among the three nodes align in a way that maintains cognitive consistency—typically when there are either all positive links or two negative and one positive link. In this project, the high proportion of balanced triads (92.10%) reflects a stable social structure, supporting Heider's assumption that individuals tend to organize their relationships in a psychologically coherent and balanced way.

Centrality Analyses: According to degree, closeness, betweenness, and PageRank metrics, some nodes stood out in all centrality measures, suggesting that these nodes are located at critical points in the network.

Metric	Value
Number of Nodes	10,044
Number of Edges	28,305
Average Total Degree	5.636
Average In-Degree / Out-Degree	2.82
Number of Weakly Connected Components	323
Largest WCC Node Count	9,340
Number of Strongly Connected Components	7,557
Largest SCC Node Count	2,204
Diameter (WCC)	12
Average Path Length (WCC)	4.285
Average Clustering Coefficient	0.031
Network Density	0.000281
Number of Triangles	7,489
Reciprocity Ratio	0.174
Degree Assortativity Coefficient	-0.110
Louvain Modularity (Positive Edges)	0.511

4.2 Key Findings from Gephi Visualization

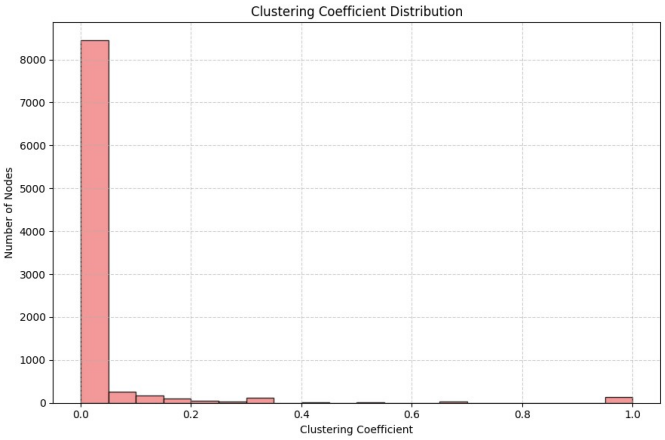
Visualization: The network was laid out using the ForceAtlas2 algorithm, and central nodes were densely clustered in the inner regions.

Community Detection: Communities identified by Gephi's Louvain algorithm are shown with different colors. The high modularity value of 0.711 indicates that the network has very strong community structures.

Centrality Visualization: Nodes were colored according to their closeness centrality, with central nodes generally concentrated in highly connected areas.

Network Properties:

- Average degree: 2.818
- Average path length: 5.282
- Network density: 0.000
- Clustering coefficient: 0.021
- Maximum diameter: 15 was observed.



The Gephi results support the numerical values obtained with NetworkX, and the overall structure of the network is presented more intuitively visually.

5. Comparison (NetworkX vs. Gephi)

In this project, network analysis was performed on the same dataset using both NetworkX and Gephi tools, and the strengths and limitations of each tool were evaluated comparatively.

NetworkX, being a Python-based library, allows for detailed mathematical calculations on the network. Thanks to its code-based approach, analyses are repeatable, controllable parametrically, and highly customizable. Analyses such as degree, centrality measurements, component analysis, triad structures, and community detection could be performed directly with Python commands. Furthermore, community detection methods like the Louvain algorithm were easily implemented, and the output files of the metrics could be obtained.

On the other hand, Gephi provided a significant advantage in terms of visual analysis and intuitive interpretation. Especially with layout algorithms like ForceAtlas2, the structures between nodes became clearer; communities, density centers, and marginal nodes were easily observed through the graph. Centrality and community analyses were visualized with color and size coding, which made Gephi much more effective, especially in reporting and presentation processes.

While both tools produced similar results, the fundamental difference between them lies in their intended uses. NetworkX is more suitable for data science, statistics, and computational analyses; Gephi excels in visual exploration, intuitive analysis, and communication.

Therefore, the combined use of both tools in this study provided a holistic analysis in terms of both numerical accuracy and visual intuitiveness.

6. Discussion and Interpretation

The analyses conducted have revealed the effects of directed and signed relationships between users on the social structure of the Slashdot social network. In this network, where positive relationships are predominant, the proportion of hostility connections is lower and not distributed to the extent that would disrupt the balance of the structure. In particular, the triad analysis results, with 92.10% balanced triad structures, indicate that the network has generally evolved in accordance with social balance principles. This suggests a tendency for users in social networks to seek consistency in their relationships.

The high modularity values resulting from community detection (0.510 in NetworkX and 0.711 in Gephi) show that the network has distinct, segregated sub-communities. This suggests that users are grouped according to their interests or social cliques. Furthermore, the fact that the largest communities consist of a few hundred nodes

and there are many small communities indicates that the network has a hierarchical and modular structure.

When centrality measures (degree, betweenness, closeness, PageRank) were analyzed, it was observed that certain nodes have high scores in all metrics. These nodes likely play central roles in information flow and the formation of social ties. In particular, actors such as node 522 are understood to be located at strategic points in the network.

However, it is also noteworthy that the network has a rather sparse structure. Both NetworkX and Gephi results show that the network density is quite low (approximately 0.0003). Additionally, the reciprocity rate remaining at 0.174 reveals that reciprocal relationships are limited and most connections are unidirectional. This indicates a platform structure where social relationships are largely based on unilateral declarations.

One of the limitations of the study is that the dataset represents only a static snapshot in time. The dynamic structure of the network over time could not be observed. Moreover, some of the analyses (e.g., triad analysis) were performed using a sampling method instead of the entire network. While this is sufficient to see general trends, working on the entire network may be necessary for absolute results.

Overall, the results obtained show that directed and signed social networks have complex but analyzable structures, and that both visual and statistical analyses provide complementary information.

7. Conclusion

In this study, friendship and hostility relationships between users in the Slashdot social network were examined both numerically and visually through network analysis conducted on the dataset. Within the scope of the project, detailed statistical analyses were performed using the NetworkX library in Python, followed by network visualization and intuitive evaluations using the Gephi tool.

The analyses revealed that the network is structurally largely balanced, positive relationships are dominant, and distinct community structures exist. The communities identified by the Louvain algorithm were supported by high modularity values, emphasizing the network's divided nature. Key nodes within the network were identified through centrality analyses, and it was observed that these actors play critical roles in the network's information flow.

The triad analysis provided results consistent with social balance theory, indicating that the relationships in the network are largely consistent and harmonious. Furthermore, the network's sparse structure and low reciprocity suggest that user relationships are mostly shaped by unilateral declarations.

The combined use of NetworkX and Gephi enabled a deeper analysis in both technical and visual dimensions, offering an effective approach to understanding the multi-dimensional nature of social networks. This project has once again demonstrated that directed and signed network analysis is a powerful method for making meaningful inferences about social structures.

References

i. Breiman, **Leskovec, J., Huttenlocher, D., & Kleinberg, J.** (2010). *Signed Networks in Social Media*. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.Chawla, N. V., et al. “SMOTE: Synthetic Minority Over-sampling Technique.” JAIR, 2002.

ii. SNAP Datasets – Stanford Network Analysis Project https://snap.stanford.edu/data/soc-sign-Slashdot081106.html

iii. NetworkX Documentation https://networkx.org/documentation/stable/

iv. Gephi Documentation https://gephi.org/users/

v. Newman, M. E. J. (2010). Networks: An Introduction. Oxford University Press.

vi. Wasserman, S., & Faust, K. (1994). Social Network Analysis: Methods and Applications. Cambridge University Press.

vii. Kağıtçıbaşı, Ç. (1988). İnsan ve insanlar. Evrim Basım Yayım Dağıtım.

viii. Cengiz A. (2000), Siyasal İkna, Ankara: Vadi Yayınları.

ix. Atıf Bir, A.; Suher,K. (2017). Reklam Teorileri: Reklamın İnsan Davranışına Etkileri, the Kitap İletişim Medya: İstanbul.