Educational Administration: Theory and Practice
2024, 30(1), 6258-6278
ISSN: 2148-2403
https://kuey.net/
Research Article

Analyzing Interest-Based Homophily in Online Social Networks Using Community Detection Methods

Manoj Kumar Srivastav[1*],Somsubhra Gupta[2],Subhranil Som[3]

[1]School of Computer Science, Swami Vivekananda University, Barrackpore, West Bengal, India,Email: mksrivastav2015@gmail.com
[2]School of Computer Science, Swami Vivekananda University, Barrackpore, West Bengal, India,Email: gsomsubhra@gmail.com
[3]Department of Computer Science, Bhairab Ganguly College, Kolkata, West Bengal, India,Email: subhranil.som@gmail.com

Citation: Manoj Kumar Srivastav, et.al (2024). Analyzing Interest-Based Homophily in Online Social Networks Using Community Detection Methods, *Educational Administration: Theory and Practice*, 30(1), 6258-6278
Doi: 10.53555/kuey.v30i1.9576

ARTICLE INFO	ABSTRACT

A social network is a structure made up of individuals, groups, or organizations connected by relationships. These relationships can be friendships, family bonds, or professional links. In this study, the focus is on homophily, which means people with similar interests tend to connect. The research analyzes a dataset of 1,000 users from [social media platform] to understand how shared interests affect community formation. Pearson correlation measures interest similarity between users. The Label Propagation Algorithm (LPA) and the Louvain method help detect user communities. The study finds distinct communities, where users are grouped based on common interests. The results show that homophily strongly influences how online communities form. This research provides a simple method to analyze user connections and improve community detection in social networks.

Keywords: Social networks, Homophily, Shared interests, Community detection, Pearson correlation, Label Propagation, Louvain method.

1. INTRODUCTION

Social networks help people, groups, and even countries connect through virtual platforms (Cota et al., 2019). These networks are represented as graphs where nodes represent individuals, and links show their connections (Cota et al., 2019). A key factor shaping social networks is **homophily**, which means people tend to connect with others who have similar interests, backgrounds, or characteristics (Bakshy et al., 2012; Song et al., 2016).

This study explores **interest-based homophily** and how it influences online social networks. The research uses a dataset from Kaggle, analyzing user interactions, group memberships, and page likes. The goal is to understand how common interests help people form and maintain friendships in digital spaces (Song et al., 2016; Ljepava et al., 2013). The findings show that people who share common interests are more likely to become friends and maintain strong relationships (Althoff et al., 2016). Those with similar hobbies, passions, or professional interests interact more often, helping create stable and closely connected communities (Gillani et al., 2018).

Social Network Analysis and Homophily

Social Network Analysis (SNA) is a method used to study social systems by finding patterns and relationships between individuals and groups. A key concept in SNA is **homophily**, which means people prefer to connect with those who are similar to them. This leads to clusters of similar individuals within a network. The word **homophily** comes from Greek: "homo" means "same," and "phily" means "love" or "liking." Homophily plays a big role in shaping social networks because it influences how people form relationships.

Types of Homophily

Homophily is divided into two types: **status homophily** and **value homophily** (Hughes, 1945).

- **Status homophily** occurs when people connect based on similar social status, including education, jobs, and income levels.
- **Value homophily** is based on shared beliefs, such as religion, political views, and cultural values.

A specific form of value homophily is **interest-based homophily**, where people connect based on shared hobbies, entertainment preferences, professional interests, or lifestyle choices.

Interest-Based Homophily in Social Networks

Interest-based homophily is visible in various networks:

- **Online social networks**: People join groups on platforms like Facebook, Twitter, and Instagram.
- **Professional networks**: LinkedIn connects users based on career interests.
- **Community groups**: Offline clubs and organizations bring together people with common interests.

This study focuses on how shared interests help people form communities and maintain social ties in online platforms.

To analyze interest-based homophily, this study uses three key methods:

- **Correlation Matrix Analysis**: Measures how strongly users are connected based on shared interests.
- **Machine Learning for Threshold Value Determination**: Uses supervised learning to predict when a friendship is likely to form.
- **Louvain Method for Community Detection**: Identifies tightly connected communities based on common interests.

Many studies show that **homophily** influences friendships and social groups. However, most research focuses on demographic homophily (age, gender, race), while **interest-based homophily** has not been studied as much.Granovetter (1973) introduced the idea of **weak ties**—connections between people with fewer common interests. These weak ties still help build social networks. This research expands on that idea by focusing on **interest-based ties**.Recent studies (Brahim et al., 2021; Nooribakhsh et al., 2024) have used **Label Propagation Algorithm (LPA)** to detect communities. This method spreads labels across the network to find groups with shared interests. This study applies **LPA** to examine community formation based on common interests.

Objective

- Measuring User Relationships: Evaluating the strength and direction of relationships between users based on their shared interests using a correlation matrix.
- Determining Friendship Thresholds: Identifying interaction thresholds necessary for the formation and strengthening of friendships through supervised learning models.
- Detect communities of users with common interests using the **Louvain method**.
- To analyze how shared interests influence the relationships between users in social networks.
- To explore the potential of the Label Propagation Algorithm (LPA) in detecting tightly-knit communities formed around shared interests.

2. LITERATURE REVIEW

Homophily, the tendency of individuals to form connections with those who share similar characteristics, plays a key role in shaping social networks (Ferrara et al., 2022; Masuda & Konno, 2006). It influences the formation, evolution, and dynamics of social relationships, leading to the emergence of distinct communities. Homophily can be classified into two main types: status homophily and value homophily (Cinelli et al., 2020; Valdés et al., 2020). Status homophily is based on attributes like race, age, and education, while value homophily is rooted in shared beliefs, attitudes, and interests.

The impact of homophily on network structures is significant. It facilitates the creation of tightly-knit clusters where individuals share high similarity, but it can also contribute to the formation of "echo chambers" that reinforce existing beliefs and limit exposure to diverse viewpoints (Cinelli et al., 2020). Homophily affects information flow within networks by enabling rapid diffusion within similar groups while restricting communication between different groups (Ferrara et al., 2022).

Social Network Analysis (SNA) provides a framework for studying social structures by modeling relationships as graphs, where nodes represent individuals and edges represent interactions. A core concept in SNA is homophily, which shapes online and offline communities (McPherson, Smith-Lovin, & Cook, 2001). One of the key techniques in SNA is **community detection (CD)**, which identifies groups within a network.

The **Label Propagation Algorithm (LPA)** is a widely used method for detecting communities. It spreads labels among network nodes to form clusters. Brahim et al. (2021) reviewed various community detection techniques, highlighting LPA's efficiency in large-scale networks. However, challenges remain in improving its robustness and scalability. Zhao, Wang, and Zhang (2012) proposed an enhancement to LPA by introducing **label entropy**, an information-theoretic approach that improves robustness by updating node labels based on entropy values. Their experiments showed that this modification improves the accuracy of community detection.

Raghavan, Albert, and Kumara (2007) originally introduced LPA as a semi-supervised algorithm for large-scale networks. While effective, LPA has rarely been combined with other machine learning techniques. Zhu, Lafferty, and Ghahramani (2007) applied LPA in recommender systems, demonstrating its ability to detect user communities based on shared interests. More recently, Nooribakhsh, Fernández-Diego, and González-

Ladrón-De-Guevara (2024) explored machine learning techniques to enhance interest-based community detection. Their study combined LPA with supervised learning to improve precision in identifying user communities.
Overall, research on community detection has progressed from foundational LPA models to advanced machine learning-based methods. Integrating LPA with label entropy and supervised learning improves robustness, scalability, and precision. These advancements are crucial for understanding online communities, predicting user behavior, and analyzing information flow in complex networks.

3. Methodology

3.1Dataset Overview

The dataset consists of user profiles from a social media platform, including details such as UserID, Name, Gender, Date of Birth (DOB), Interests, City, and Country. The users have a diverse range of interests, which are categorized into unique types such as Art, Beauty, Books, Business and Entrepreneurship, and many others.[https://www.kaggle.com/datasets/arindamsahoo/social-media-users]

Table 1: Social media user dataset

UserID	Gender	DOB	Interests	City	Country
1	Female	15-10-1958	'Movies', 'Fashion', 'Fashion', 'Books'	Sibolga	Indonesia
2	Female	21-07-2004	'Gaming', 'Finance and investments', 'Outdoor activities', 'Travel'	Al AbyÄ□r	Libya
3	Female	07-02-2000	'DIY and crafts', 'Music', 'Science', 'Fashion'	WÄ□dÄ« as SÄ«r	Jordan
4	Male	14-04-1985	'Outdoor activities', 'Cars and automobiles'	Matera	Italy
5	Female	18-09-1955	'Politics', 'History'	Biruaca	Venezuela
6	Male	18-06-1967	'Travel'	Belton	United States
7	Female	09-02-1969	'Outdoor activities', 'Movies'	Haslingden	United Kingdom
8	Male	30-12-1965	'Beauty', 'Nature', 'Gardening', 'Food and dining', 'DIY and crafts'	Ad-Damazin	Sudan
9	Male	16-08-1984	'Parenting and family', 'Photography', 'Finance and investments'	Tabuk	Saudi Arabia
10	Female	08-03-2003	'Gardening'	Ongole	India
995	Female	04-01-1970	'Art', 'Books', 'Sports', 'Art', 'Fitness'	PoÅ¾ega	Croatia
996	Male	23-07-1960	'Fitness', 'Health and wellness'	Korntal	Germany
997	Male	27-05-1954	'Cooking', 'Finance and investments', 'Politics', 'Business and entrepreneurship', 'Travel'	JosÃ© de Freitas	Brazil
998	Male	20-08-1989	'Books', 'Science', 'Fashion', 'Outdoor activities'	Cicero	United States
999	Female	04-08-1983	'Photography', 'Sports', 'Gaming'	Haitang	China
1000	Female	17-08-1990	'Photography', 'Gardening', 'Gaming', 'Business and entrepreneurship'	Brunswick	Australia

Table 2: ALL UNIQUE INTERESTS

I1: 'Art' , I2: 'Beauty' , I3: 'Books' ,I4: 'Business and entrepreneurship', I5: 'Cars and automobiles' ,I6: 'Cooking', I7: 'DIY and crafts',I8: 'Education and learning', I9: 'Fashion',I10: 'Finance and investments',I11: 'Fitness' I12: 'Food and dining',I13: 'Gaming',I14: 'Gardening',I15: 'Health and wellness',I16: 'History',I17: 'Movies', I18: 'Music',I19: 'Nature',I20: 'Outdoor activities',I21: 'Parenting and family',I22: 'Pets',I23: 'Photography',I24: 'Politics',I25: 'Science',I26: 'Social causes and activism',I27: 'Sports',I28: 'Technology',I29: 'Travel'

Table 3: Representation of name as user U1, U2,……,U1000

User ID
U1
U2
U3
U4
U5
U6
U7
U8
…
U995
U996
U997
U998
U999
U1000

Matrix with respect user and its interest

Interest Matrix

The matrix represents the presence (1) or absence (0) of each unique interest for each user. For instance, User U1 is interested in 'Books' but not in 'Gardening,' while User U2 is interested in 'Travel' but not in 'Books.'

Table 4: representation of user and their interest

	I1	I2	I3	I4	I5		I25	I26	I27	I28	I29
U1	0	0	1	0	0		0	0	0	0	0
U2	0	0	0	0	0		0	0	0	0	1
U3	0	0	0	0	0		1	0	0	0	0
U4	0	0	0	0	1		0	0	0	0	0
U5	0	0	0	0	0		0	0	0	0	0
U6	0	0	0	0	0		0	0	0	0	1
U7	0	0	0	0	0		0	0	0	0	0
U8	0	1	0	0	0		0	0	0	0	0
U9	0	0	0	0	0		0	0	0	0	0
U10	0	0	0	0	0		0	0	0	0	0
U990	0	0	0	0	0		0	0	0	0	0
U991	0	0	0	0	0		0	1	1	0	0
U992	0	0	0	0	0		0	0	0	0	1
U993	0	0	0	1	0		0	0	0	0	0
U994	0	0	0	1	0		0	0	0	0	0
U995	1	0	1	0	0		0	0	1	0	0
U996	0	0	0	0	0		0	0	0	0	0
U997	0	0	0	1	0		0	0	0	0	1
U998	0	0	1	0	0		1	0	0	0	0
U999	0	0	0	0	0		0	0	1	0	0
U1000	0	0	0	1	0		0	0	0	0	0

Correlation Matrix among Users

The correlation matrix captures the strength of the relationship between pairs of users based on their interests. Values range from -1 to 1, indicating the degree of similarity or dissimilarity. For example, User U1 and User U2 have a correlation of -0.13587, suggesting a low level of similarity in their interests, whereas User
U1 and User U998 have a correlation of 0.520847, indicating a higher similarity. This data will be used to analyze user relationships, determine friendship thresholds, and identify community structures within the social network based on shared interests.

Table 5: Correlation Matrix among Users

	U1	U2	U3	U4	U5			U996	U997	U998	U999	U1000
U1	1	-0.13587	0.192487	-0.09245	-0.09245			-0.09245	-0.15504	0.520847	-0.11538	-0.13587
U2	-0.13587	1	-0.16	0.285774	-0.10887			-0.10887	0.346891	0.13	0.192487	0.13
U3	0.192487	-0.16	1	-0.10887	-0.10887			-0.10887	-0.18257	0.42	-0.13587	-0.16
U4	-0.09245	0.285774	-0.10887	1	-0.07407			-0.07407	-0.12423	0.285774	-0.09245	-0.10887
U996	-0.092450	-0.108866	-0.108866	-0.074074	-0.074074			1	-0.12423	-0.10887	-0.09245	-0.10887
U997	-0.155043	0.346891	-0.182574	-0.124226	0.236029			-0.12423	1	-0.18257	-0.15504	0.082158
U998	0.520847	0.130000	0.420000	0.285774	-0.108866			-0.10887	-0.18257	1	-0.13587	-0.16
U999	-0.115385	0.192487	-0.135873	-0.092450	-0.092450			-0.09245	-0.15504	-0.13587	1	0.520847
U1000	-0.135873	0.130000	-0.160000	-0.108866	-0.108866			-0.10887	0.082158	-0.16	0.520847	1

Measuring Homophily and Detecting Communities in Social Networks

Understanding homophily in social networks involves assessing the similarity between users based on shared interests or interactions. Two key approaches for measuring homophily are the **Jaccard similarity coefficient** and **correlation analysis**, while **community detection methods**, such as the **Louvain algorithm**, help identify clusters of closely connected users.

1. Measuring Homophily

(a) Jaccard Similarity Coefficient: Measurement of Homophily by Calculating the Average Pairwise Similarity Between Users: Homophily in social networks refers to the tendency of individuals to associate and bond with similar others. In the context of a user-interest matrix, homophily can be quantified by measuring the similarity of interests between users. One common mathematical approach to measure homophily is to calculate the average pairwise similarity between users using the Jaccard similarity coefficient.

Mathematical Explanation

Let's denote the user-interest matrix by M, where Mij is 1 if user i has interest j, and 0 otherwise. The matrix has dimension U x I, where U is the number of users and I is the number of interests.

For any two users i and j, the similarity S_{ij} can be calculated using various measures. One common measure is the Jaccard similarity coefficient, which is defined as:

$S_{ij} = |\frac{M_i \cap M_j}{M_i \cup M_j}|$ Where Mi and Mj are the sets of interests of users i and j, respectively. The numerator $|M_i \cap M_j|$ is the number of common interests, and the denominator $|M_i \cup M_j|$ is the total number of distinct interest.

Step 4: Calculate Average Homophily Score:

• Extract the upper triangle of the similarity matrix (excluding the diagonal) and calculate the mean value to get the average homophily score.

• Average Homophily score:The homophily score H for the network can be defined as the average of all pairwise similarities :

$$H = \frac{1}{\binom{U}{2}} \sum_{i=1}^{U-1} \sum_{j=i+1}^{U} S_{ij}$$

where $\binom{U}{2}$ is the binomial coefficient representing the number of unique pairs of users.

This approach ensures that the Homophily measure is robust and applicable in all situations where user interests can be represented as binary vectors.This algorithm provides a robust method to measure homophily in social networks where user interests can be represented as binary vectors.

Table 6: Similarity score based on User-Interest

	U1	U2	U3	U4	U5	U6	U7	U8 \
U1	1.000000	0.000000	0.166667	0.0	0.000000	0.00	0.250000	0.000
U2	0.000000	1.000000	0.000000	0.2	0.000000	0.25	0.200000	0.000
U3	0.166667	0.000000	1.000000	0.0	0.000000	0.00	0.000000	0.125
U4	0.000000	0.200000	0.000000	1.0	0.000000	0.00	0.333333	0.000
U5	0.000000	0.000000	0.000000	0.0	1.000000	0.00	0.000000	0.000
...	...	...	...	...	...	...	...	...
U996	0.000000	0.000000	0.000000	0.0	0.000000	0.00	0.000000	0.000
U997	0.000000	0.285714	0.000000	0.0	0.166667	0.20	0.000000	0.000
U998	0.400000	0.142857	0.333333	0.2	0.000000	0.00	0.200000	0.000
U999	0.000000	0.166667	0.000000	0.0	0.000000	0.00	0.000000	0.000
U1000	0.000000	0.142857	0.000000	0.0	0.000000	0.00	0.000000	0.125

	U9	U10	...	U991	U992	U993	U994	U995	U996 \
U1	0.000000	0.00	...	0.00	0.000000	0.142857	0.00	0.166667	0.0
U2	0.166667	0.00	...	0.00	0.200000	0.125000	0.00	0.000000	0.0
U3	0.000000	0.00	...	0.00	0.000000	0.125000	0.00	0.000000	0.0
U4	0.000000	0.00	...	0.00	0.000000	0.000000	0.00	0.000000	0.0
U5	0.000000	0.00	...	0.00	0.333333	0.000000	0.00	0.000000	0.0
...	...	...	...	...	...	...	...	...	...
U996	0.000000	0.00	...	0.00	0.000000	0.000000	0.00	0.200000	1.0
U997	0.142857	0.00	...	0.00	0.400000	0.111111	0.20	0.000000	0.0
U998	0.000000	0.00	...	0.00	0.000000	0.125000	0.00	0.142857	0.0
U999	0.200000	0.00	...	0.25	0.000000	0.142857	0.00	0.166667	0.0
U1000	0.166667	0.25	...	0.00	0.000000	0.285714	0.25	0.000000	0.0

	U997	U998	U999	U1000
U1	0.000000	0.400000	0.000000	0.000000
U2	0.285714	0.142857	0.166667	0.142857
U3	0.000000	0.333333	0.000000	0.000000
U4	0.000000	0.200000	0.000000	0.000000
U5	0.166667	0.000000	0.000000	0.000000
...	...	...	...	...
U996	0.000000	0.000000	0.000000	0.000000
U997	1.000000	0.000000	0.000000	0.125000
U998	0.000000	1.000000	0.000000	0.000000
U999	0.000000	0.000000	1.000000	0.400000
U1000	0.125000	0.000000	0.400000	1.000000

[1000 rows x 1000 columns]

For a dataset with 1000 users, the user-user Correlation Homophily Matrix is a 1000x1000 matrix that quantifies the correlation between users based on their interests or behaviors. Each cell (i, j) in this matrix represents the correlation coefficient between the interests or activities of user i and user j. Higher correlation values indicate stronger homophily, reflecting that users have more similar interests or behavior patterns. This matrix is useful for identifying clusters of users with shared characteristics and understanding the impact of homophily on network dynamics.

Table7: User-User Homophily Correlation Matrix

User-User Correlation Homophily Matrix:

	U1	U2	U3	U4	U5	U6	U7 \
U1	1.000000	0.166090	0.438632	0.114790	0.110609	0.070388	0.289277
U2	0.166090	1.000000	0.200337	0.273418	0.091314	0.306548	0.268473
U3	0.438632	0.200337	1.000000	0.119342	0.127389	0.093394	0.132383
U4	0.114790	0.273418	0.119342	1.000000	0.073579	0.050584	0.384000
U5	0.110609	0.091314	0.127389	0.073579	1.000000	0.040486	0.060510
...	...	...	...	...	...	...	...
U996	0.128540	0.131579	0.116232	0.108626	0.080906	0.071970	0.094512

```
U997  0.196141 0.452055 0.215839 0.127953 0.225383 0.253695 0.138132
U998  0.618834 0.283422 0.518962 0.260870 0.123173 0.077434 0.262332
U999  0.150579 0.296943 0.171587 0.098191 0.084211 0.061765 0.079012
U1000 0.185249 0.275926 0.200658 0.111111 0.102174 0.070423 0.110647

U8      U9      U10 ...    U991    U992     U993 \
U1    0.182566 0.144444 0.062350 ... 0.107221 0.114224 0.367521
U2    0.201005 0.321888 0.060096 ... 0.120000 0.313776 0.285024
U3    0.265442 0.150613 0.080717 ... 0.126033 0.137014 0.386777
U4    0.132231 0.109181 0.046154 ... 0.098361 0.077882 0.131295
U5    0.131078 0.112821 0.057143 ... 0.055738 0.348178 0.128205
...      ...      ...      ... ...      ...      ...      ...
U996  0.133333 0.116505 0.067416 ... 0.097792 0.091185 0.146429
U997  0.220155 0.299242 0.075789 ... 0.105973 0.440887 0.333333
U998  0.206625 0.156794 0.060738 ... 0.110664 0.126000 0.371151
U999  0.176796 0.365482 0.061404 ... 0.318885 0.092500 0.283422
U1000 0.327306 0.321503 0.297450 ... 0.130152 0.122363 0.422339

U994    U995    U996    U997    U998    U999    U1000
U1    0.092683 0.308571 0.128540 0.196141 0.618834 0.150579 0.185249
U2    0.082524 0.195462 0.131579 0.452055 0.283422 0.296943 0.275926
U3    0.075055 0.218121 0.116232 0.215839 0.518962 0.171587 0.200658
U4    0.065385 0.114471 0.108626 0.127953 0.260870 0.098191 0.111111
U5    0.060241 0.103070 0.080906 0.225383 0.123173 0.084211 0.102174
...      ...      ...      ...      ...      ...      ...      ...
U996  0.082090 0.290244 1.000000 0.131274 0.137097 0.100503 0.119748
U997  0.270936 0.198413 0.131274 1.000000 0.213518 0.175221 0.320000
U998  0.078603 0.313620 0.137097 0.213518 1.000000 0.165154 0.186795
U999  0.079179 0.294243 0.100503 0.175221 0.165154 1.000000 0.468750
U1000 0.311615 0.197952 0.119748 0.320000 0.186795 0.468750 1.000000
```

(b) Correlation Analysis

Beyond set-based similarity, correlation analysis measures homophily by analyzing the statistical relationship between users' engagement patterns. Given a user correlation matrix, Pearson correlation can be used to assess the linear relationship between users' interaction behaviors. A correlation coefficient close to **1** signifies strong homophily, while values near **0** suggest little to no similarity.While homophily scores quantify similarity between users, community detection methods identify structurally cohesive groups within a network. The **Louvain method** is an efficient approach that maximizes modularity to detect tightly connected communities. The algorithm operates in two stages:

- **Local Modularity Optimization:**Each node is initially assigned to its own community, and the algorithm iteratively reassigns nodes to communities where they maximize modularity gain.
- **Community Aggregation:** The detected communities are collapsed into single nodes, and the process is repeated to refine community structures.

Unlike homophily measures, which assess pairwise user similarity, the Louvain method reveals larger structural patterns within the network, helping to uncover natural divisions and interest-based communities.In real-world networks, individuals who share common characteristics or interests naturally form communities. Within a community, members are closely connected, exhibiting strong internal ties, while their connections to individuals outside the community are relatively weak. The formation of such community structures is largely driven by **homophily**, the principle that similar nodes are more likely to form links with one another, whereas dissimilar nodes tend to be less connected. This inherent tendency causes networks to self-organize into distinct groups based on shared attributes, interactions, or affiliations.

Types of Communities in Social Networks with respect to purpose

Communities in social networks can be categorized based on their formation and structural properties:

- Interest-Based Communities – Formed by individuals with shared interests, such as online discussion groups, hobby-based forums, and fan communities.
- Social or Friendship Communities – Groups of users who are socially connected through personal relationships, such as family circles and friend groups on platforms like Facebook.
- Professional or Work-Related Communities – Networks built around professional affiliations, such as LinkedIn groups, research collaborations, and corporate networks.

• Geographical or Location-Based Communities – Users who interact based on geographical proximity, such as local neighborhood groups or city-based online communities.
• Transactional or Economic Communities – Groups formed based on business interactions, such as buyers and sellers on e-commerce platforms.

Community Detection Methods

Identifying community structures in social networks is important for analyzing group dynamics and information flow. Several methods have been developed for this purpose:
• Graph Partitioning Methods – These techniques divide the network into fixed-size clusters while minimizing inter-cluster connections. Examples include the K-means clustering and Spectral clustering approaches.
• Modularity-Based Methods – Algorithms that optimize modularity, a measure of network division quality. The Louvain method is a widely used technique in this category.
• Hierarchical Clustering – This method builds a hierarchy of communities by progressively merging or splitting clusters based on similarity.
• Density-Based Methods – Techniques such as DBSCAN (Density-Based Spatial Clustering) identify dense regions in the network as communities.
• Label Propagation Methods – These algorithms assign labels to nodes, which propagate through the network until stable communities emerge.
• Deep Learning and Machine Learning-Based Approaches – Advanced models, such as Graph Neural Networks (GNNs), leverage AI to detect communities in large-scale networks dynamically.
Understanding community structures in social networks is essential for applications such as targeted marketing, recommendation systems, and social influence analysis. By detecting and analyzing these groups, researchers can gain deeper insights into network behavior and information diffusion.
Interest-Based Homophily
In social networks, individuals tend to form connections based on shared interests, a phenomenon known as interest-based homophily. This principle explains why online communities emerge around specific topics, such as hobby groups, professional networks, and discussion forums. Interest-based homophily plays a key role in understanding content dissemination, group polarization, and information diffusion.

Selection of Community Detection Methods

For detecting communities in social networks, the following methods have been chosen due to their efficiency, scalability, and alignment with the structural properties of interest-based homophily.
Louvain Method (Modularity Optimization)
• The Louvain method is a widely used approach for detecting disjoint communities by optimizing modularity, a measure of the strength of community structure.
• This method is preferred because it is computationally efficient, making it suitable for large-scale networks.
• By identifying well-separated groups, it helps analyze strongly connected communities in social networks.
Label Propagation Method (LPA)
• The Label Propagation Algorithm (LPA) assigns labels to nodes, which spread iteratively until stable community structures emerge.
• This method is particularly useful for detecting disjoint communities in large networks due to its scalability and simplicity.
• Its variants, such as the Speaker-Listener Label Propagation Algorithm (SLPA), allow for the detection of overlapping communities, capturing the dynamic nature of social interactions.
Justification for Method Selection
• Homophily-Driven Community Formation – The chosen methods effectively capture communities formed due to interest-based homophily.
• Scalability for Large Networks – Both the Louvain method and LPA are computationally efficient, making them ideal for analyzing extensive social media datasets.
• Structural Adaptability –
○ Louvain is suitable for static networks where communities are well-defined.
○ LPA adapts well to dynamic networks, where labels propagate based on local node interactions.
The Louvain method and Label Propagation Algorithm have been selected for their ability to identify meaningful communities in social networks influenced by interest-based homophily. These methods provide a foundation for further analysis of information diffusion, user behavior, and community dynamics.

Community Detection Algorithm: The Louvain Method

The Louvain method is a widely recognized algorithm for community detection in networks, enabling the identification of densely connected groups of nodes. This section details the application of the Louvain method to uncover community structures within our dataset, which represents interactions or correlations among users. Community detection is a critical task in network analysis, aiming to reveal underlying

structures within complex networks. The Louvain method is particularly effective in identifying communities by optimizing modularity. In this study, Louvain method is applied to analyze user correlations and identify distinct communities within the dataset.
The Louvain method is designed to partition a network into communities by maximizing modularity, a measure of the strength of the division. The algorithm follows these main steps:

STEP1: Initial Partitioning
- Each node is initially assigned to its own community.
- The modularity of this initial partition is calculated as a baseline.

STEP2: Iterative Optimization
- For each node, evaluate the impact of moving it to neighboring communities by calculating the change in modularity (ΔQ).
- Nodes are reassigned to the community that results in the highest modularity gain.
- This process is repeated until no further improvement in modularity is achievable.

STEP3: Merge Communities
- Communities are merged if the combination improves modularity.
- The optimization and merging steps are repeated iteratively.

STEP 4: Final Community Structure
- The final partition represents the detected communities, highlighting groups of nodes with high intra-group connectivity.

Modularity and Modularity Gain
- Modularity: Modularity Q quantifies the strength of the division of a network into communities. It is calculated as:

$$Q=\frac{1}{2m}\sum_{i,j}[A_{i,j}-\frac{k_i k_j}{2m}]\delta(c_{i,}c_j)$$

Where, Aij is the adjacency matrix, ki and kj are the degrees of nodes i and j,
m is the total number of edges, and δ(ci,cj) is 1 if nodes i and j are in the same community and 0 otherwise.

Modularity Gain (ΔQ): The change in modularity when a node is moved to a different community is given by:

The modularity gain when merging two communities A and B is calculated using the formula:

$$\Delta Q(A \to B) = \frac{1}{2m}\left[A_{AB} - \frac{k_A \cdot k_B}{2m}\right]$$

Where:

- A_{AB} is the weight of the link between nodes A and B (the correlation value between them).
- k_A and k_B are the degrees of nodes A and B, respectively.
- m is the total number of edges in the network, calculated as the sum of all non-zero entries in the correlation matrix.

The application of the Louvain method to the dataset begins with loading the correlation matrix, which represents the relationships between users, into a suitable data structure such as a pandas DataFrame. A graph is then constructed where nodes represent users and edges correspond to the correlation values between them. Initially, each user is assigned to a separate community. The iterative optimization process follows, where modularity gain (ΔQ) is calculated for each user when considering a move to a neighboring community. Users are reassigned to communities that maximize modularity, and communities are merged iteratively to further enhance modularity until no further improvement is possible. The final community structure is then derived, identifying groups of users with strong internal correlations, which provides valuable insights into the network's structure. The work is conducted on the **user correlation matrix**, which quantifies the relationships between users based on shared interests or interactions. The correlation matrix is derived from a **user-interest matrix**, where each row represents a user, and each column represents an interest. The **Pearson correlation coefficient** is used to compute similarity scores between user pairs, generating a symmetric correlation matrix.
Using this correlation matrix, a **graph-based approach** is applied to model user relationships. Nodes represent users, and weighted edges represent the correlation values between them. To filter out weak connections, only correlations above a chosen threshold (e.g., 0.5) are considered. The **Louvain method** is then used for **community detection**, grouping users with high intra-community correlation.

This approach provides insights into **user homophily**, helping identify clusters of users with similar interests or behaviors in the network. The final detected communities highlight meaningful social structures within the dataset.

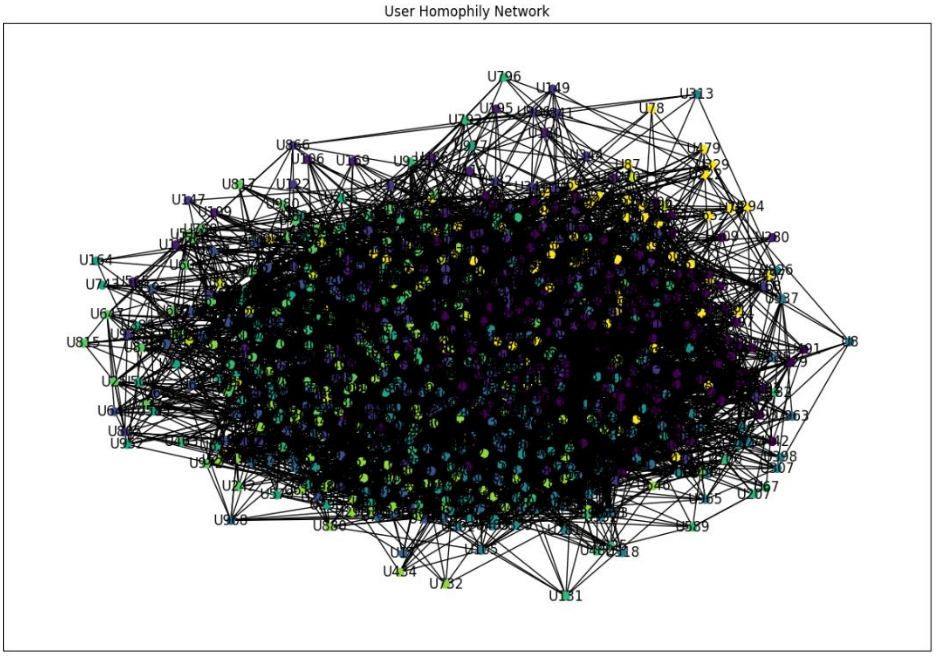

Figure 1: User Homophily network

Detected 13 communities in the dataset.

Community 0: 166 users
['U1', 'U3', 'U15', 'U39', 'U56', 'U58', 'U64', 'U71', 'U79', 'U80', 'U121', 'U126', 'U130', 'U132', 'U133', 'U150', 'U15
1', 'U152', 'U153', 'U158', 'U160', 'U161', 'U185', 'U213', 'U215', 'U220', 'U229', 'U234', 'U237', 'U245', 'U260', 'U
269', 'U271', 'U295', 'U301', 'U304', 'U306', 'U308', 'U325', 'U332', 'U333', 'U351', 'U355', 'U357', 'U359', 'U36
0', 'U361', 'U362', 'U368', 'U370', 'U373', 'U374', 'U375', 'U384', 'U401', 'U404', 'U410', 'U423', 'U425', 'U436',
'U437', 'U439', 'U442', 'U445', 'U447', 'U451', 'U459', 'U461', 'U473', 'U475', 'U477', 'U490', 'U491', 'U495', 'U5
02', 'U504', 'U506', 'U509', 'U510', 'U511', 'U528', 'U536', 'U540', 'U543', 'U558', 'U563', 'U568', 'U570', 'U574'
, 'U578', 'U596', 'U598', 'U603', 'U605', 'U610', 'U615', 'U617', 'U620', 'U625', 'U627', 'U631', 'U635', 'U639', 'U
645', 'U650', 'U664', 'U666', 'U671', 'U673', 'U677', 'U687', 'U688', 'U692', 'U697', 'U717', 'U720', 'U721', 'U725
', 'U726', 'U730', 'U740', 'U749', 'U751', 'U755', 'U757', 'U760', 'U765', 'U766', 'U767', 'U776', 'U777', 'U790', 'U
797', 'U798', 'U813', 'U827', 'U834', 'U835', 'U850', 'U851', 'U857', 'U861', 'U867', 'U870', 'U879', 'U882', 'U88
5', 'U889', 'U897', 'U907', 'U910', 'U911', 'U912', 'U915', 'U931', 'U936', 'U951', 'U953', 'U959', 'U970', 'U973', '
U975', 'U983', 'U989', 'U993', 'U998']

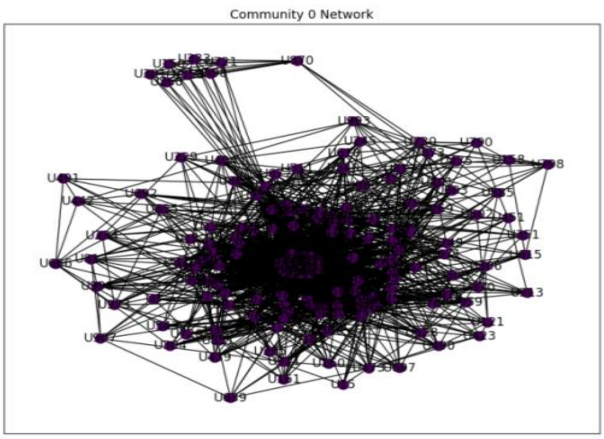

Figure2: community 0 network

Community 1: 62 users
['U2', 'U6', 'U13', 'U28', 'U41', 'U42', 'U82', 'U88', 'U92', 'U106', 'U109', 'U116', 'U120', 'U124', 'U141', 'U165', 'U
167', 'U169', 'U173', 'U180', 'U191', 'U193', 'U195', 'U228', 'U244', 'U280', 'U298', 'U336', 'U343', 'U387', 'U391',
'U394', 'U405', 'U413', 'U422', 'U432', 'U448', 'U453', 'U464', 'U486', 'U498', 'U516', 'U530', 'U538', 'U545', 'U
565', 'U577', 'U612', 'U616', 'U618', 'U679', 'U704', 'U723', 'U734', 'U762', 'U800', 'U822', 'U874', 'U884', 'U94
2', 'U992', 'U997']

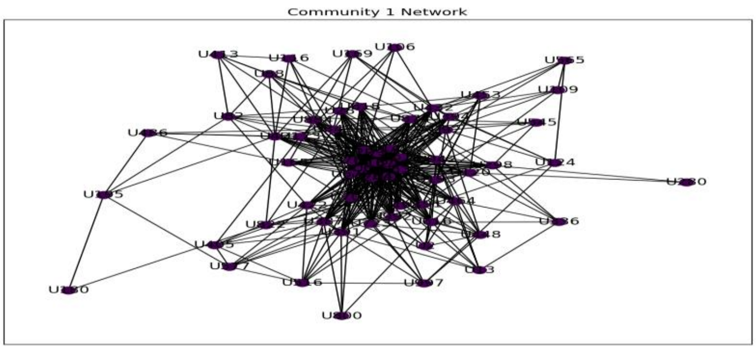

Figure3: community 1 network

Community 2: 116 users
['U4', 'U7', 'U11', 'U18', 'U20', 'U23', 'U26', 'U37', 'U44', 'U45', 'U85', 'U95', 'U97', 'U115', 'U122', 'U134', 'U139',
'U143', 'U147', 'U149', 'U166', 'U168', 'U172', 'U177', 'U184', 'U189', 'U200', 'U202', 'U204', 'U211', 'U226', 'U23
8', 'U240', 'U253', 'U261', 'U264', 'U272', 'U284', 'U285', 'U287', 'U288', 'U315', 'U321', 'U323', 'U327', 'U328',
'U339', 'U344', 'U349', 'U350', 'U369', 'U372', 'U388', 'U392', 'U397', 'U402', 'U403', 'U416', 'U435', 'U440', 'U
449', 'U456', 'U465', 'U483', 'U485', 'U487', 'U492', 'U501', 'U512', 'U520', 'U526', 'U531', 'U535', 'U557', 'U559
', 'U572', 'U582', 'U599', 'U600', 'U601', 'U607', 'U623', 'U633', 'U642', 'U643', 'U691', 'U719', 'U729', 'U733', '
U745', 'U768', 'U772', 'U789', 'U791', 'U805', 'U814', 'U824', 'U826', 'U832', 'U836', 'U841', 'U846', 'U852', 'U8
66', 'U895', 'U898', 'U901', 'U932', 'U941', 'U964', 'U976', 'U978', 'U986', 'U988', 'U990', 'U999']

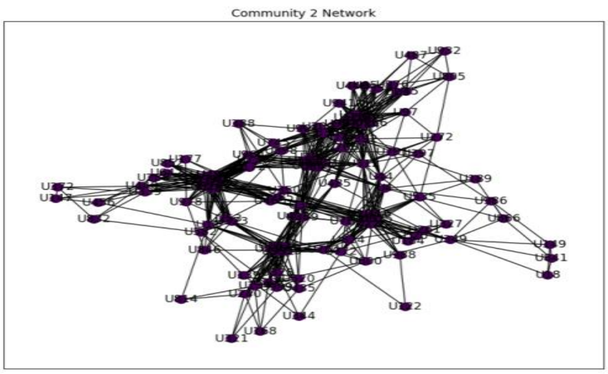

Figure3: community 2 network

Community 3: 53 users
['U29', 'U43', 'U47', 'U57', 'U59', 'U60', 'U65', 'U69', 'U94', 'U102', 'U187', 'U196', 'U197', 'U241', 'U255', 'U259',
'U267', 'U273', 'U334', 'U364', 'U377', 'U382', 'U400', 'U455', 'U476', 'U499', 'U503', 'U505', 'U515', 'U592', 'U
644', 'U654', 'U662', 'U670', 'U672', 'U722', 'U736', 'U763', 'U782', 'U801', 'U823', 'U838', 'U868', 'U878', 'U89
0', 'U900', 'U909', 'U914', 'U927', 'U939', 'U947', 'U971', 'U974']

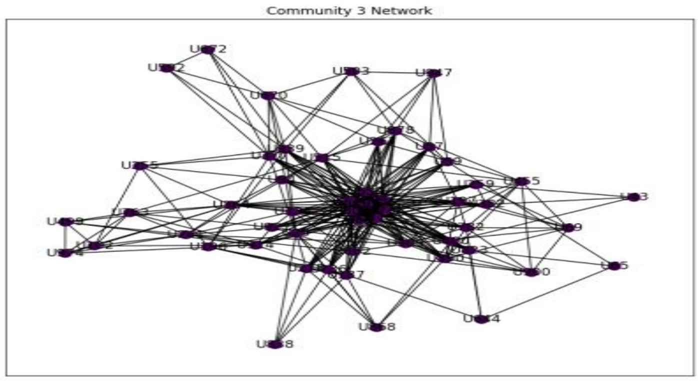

Figure4: community 3 network

Community 4: 57 users
['U17', 'U24', 'U27', 'U32', 'U74', 'U84', 'U107', 'U128', 'U136', 'U146', 'U154', 'U178', 'U190', 'U209', 'U219', 'U252', 'U268', 'U303', 'U312', 'U322', 'U335', 'U340', 'U346', 'U366', 'U411', 'U415', 'U457', 'U467', 'U537', 'U553', 'U555', 'U569', 'U606', 'U626', 'U629', 'U657', 'U661', 'U676', 'U686', 'U706', 'U724', 'U738', 'U753', 'U758', 'U769', 'U778', 'U810', 'U816', 'U833', 'U845', 'U848', 'U865', 'U886', 'U893', 'U967', 'U968', 'U969']

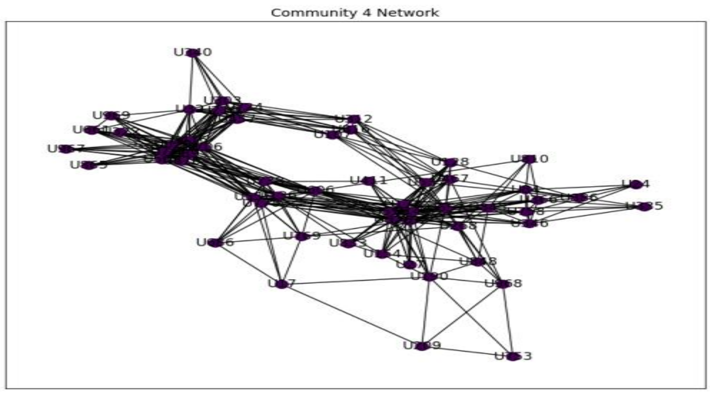

Figure5: community 4 network

Community 5: 79 users
['U8', 'U10', 'U48', 'U73', 'U75', 'U81', 'U83', 'U98', 'U99', 'U105', 'U174', 'U194', 'U214', 'U231', 'U263', 'U266', 'U277', 'U278', 'U281', 'U286', 'U292', 'U302', 'U307', 'U309', 'U313', 'U324', 'U365', 'U395', 'U398', 'U408', 'U412', 'U420', 'U430', 'U454', 'U460', 'U489', 'U494', 'U496', 'U518', 'U521', 'U523', 'U532', 'U541', 'U549', 'U550', 'U554', 'U562', 'U588', 'U611', 'U628', 'U640', 'U655', 'U663', 'U680', 'U689', 'U702', 'U703', 'U741', 'U744', 'U754', 'U771', 'U787', 'U821', 'U830', 'U837', 'U839', 'U842', 'U853', 'U905', 'U922', 'U925', 'U926', 'U937', 'U957', 'U963', 'U965', 'U981', 'U985', 'U1000']

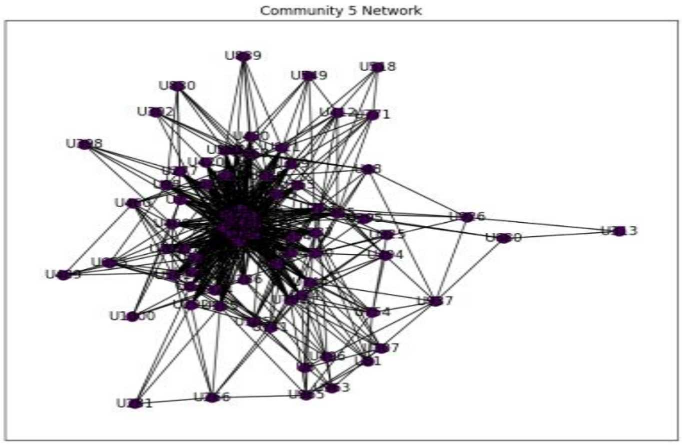

Figure6: community 5 network

Community 6: 45 users
['U33', 'U55', 'U61', 'U70', 'U100', 'U108', 'U171', 'U181', 'U188', 'U210', 'U230', 'U251', 'U274', 'U291', 'U317', 'U
389', 'U414', 'U433', 'U441', 'U450', 'U481', 'U484', 'U548', 'U576', 'U619', 'U622', 'U641', 'U649', 'U651', 'U660
', 'U674', 'U735', 'U747', 'U750', 'U774', 'U829', 'U831', 'U871', 'U913', 'U918', 'U919', 'U930', 'U956', 'U979', 'U
994']

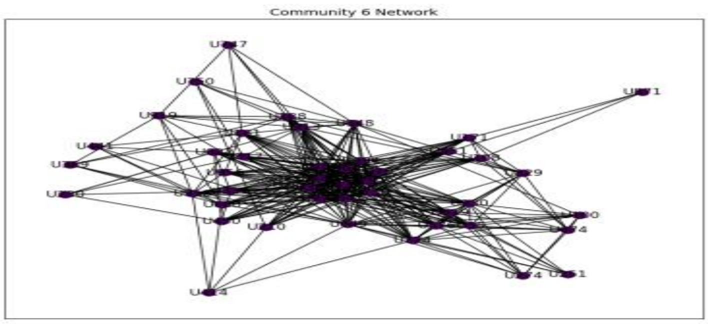

Figure7: community 6 network

Community 7: 57 users
['U5', 'U35', 'U53', 'U63', 'U117', 'U127', 'U148', 'U164', 'U175', 'U179', 'U201', 'U206', 'U212', 'U258', 'U289', 'U
290', 'U319', 'U326', 'U353', 'U371', 'U376', 'U380', 'U381', 'U669', 'U426', 'U438', 'U458', 'U462', 'U469', 'U50
0', 'U561', 'U579', 'U583', 'U584', 'U585', 'U591', 'U636', 'U646', 'U653', 'U667', 'U683', 'U695', 'U711', 'U718', '
U737', 'U743', 'U748', 'U804', 'U806', 'U807', 'U809', 'U819', 'U862', 'U896', 'U903', 'U952', 'U977']

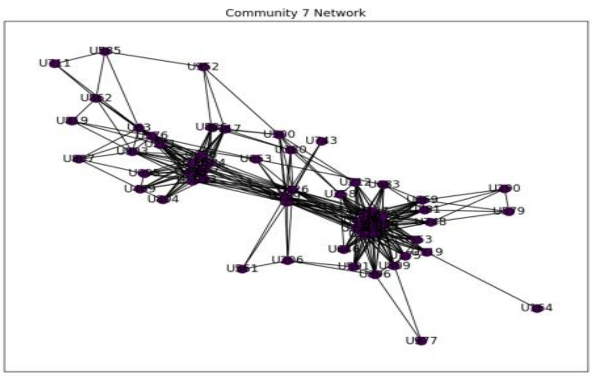

Figure8: community 7 network

Community 8: 80 users
['U19', 'U30', 'U40', 'U52', 'U67', 'U207', 'U68', 'U77', 'U90', 'U112', 'U113', 'U131', 'U135', 'U140', 'U163', 'U170', 'U182', 'U186', 'U225', 'U233', 'U235', 'U239', 'U256', 'U275', 'U283', 'U293', 'U331', 'U341', 'U342', 'U352', 'U3 56', 'U378', 'U379', 'U386', 'U393', 'U407', 'U417', 'U428', 'U444', 'U463', 'U478', 'U482', 'U508', 'U551', 'U552' , 'U580', 'U589', 'U590', 'U609', 'U621', 'U624', 'U665', 'U678', 'U699', 'U700', 'U707', 'U714', 'U727', 'U739', ' U759', 'U764', 'U770', 'U783', 'U785', 'U786', 'U792', 'U796', 'U799', 'U818', 'U825', 'U849', 'U877', 'U899', 'U9 02', 'U908', 'U934', 'U935', 'U946', 'U949', 'U984']

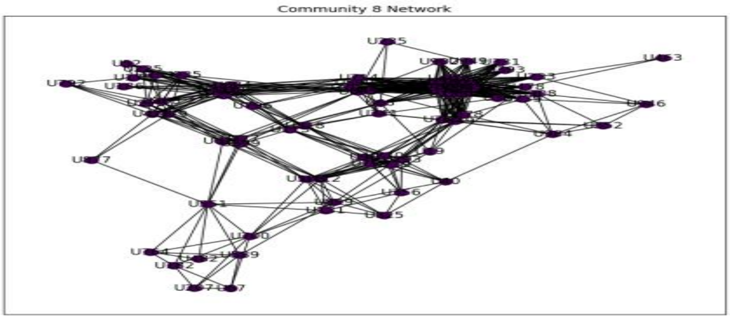

Figure9: community 8 network

Community 9: 84 users
['U12', 'U14', 'U25', 'U34', 'U36', 'U46', 'U49', 'U50', 'U51', 'U96', 'U101', 'U118', 'U125', 'U142', 'U155', 'U159', 'U 183', 'U205', 'U227', 'U232', 'U242', 'U248', 'U254', 'U279', 'U296', 'U310', 'U316', 'U330', 'U348', 'U354', 'U36 7', 'U385', 'U409', 'U424', 'U427', 'U431', 'U452', 'U493', 'U514', 'U525', 'U544', 'U560', 'U567', 'U571', 'U575', ' U593', 'U602', 'U604', 'U638', 'U647', 'U658', 'U685', 'U696', 'U705', 'U708', 'U710', 'U712', 'U715', 'U716', 'U7 31', 'U752', 'U756', 'U773', 'U779', 'U788', 'U803', 'U815', 'U817', 'U820', 'U843', 'U872', 'U876', 'U891', 'U892', 'U894', 'U921', 'U923', 'U933', 'U940', 'U948', 'U955', 'U972', 'U980', 'U996']

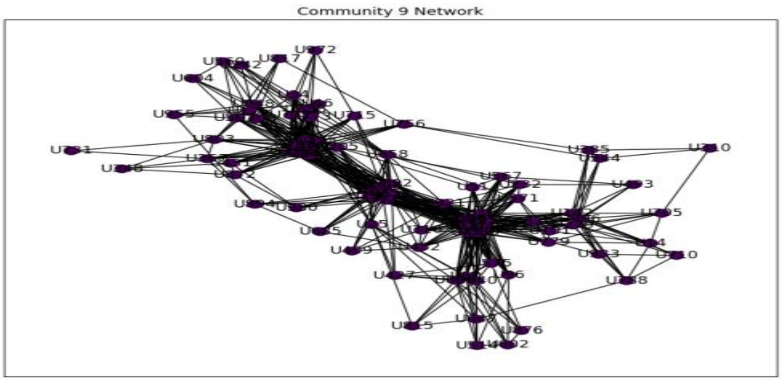

Figure10: community 9 network

Community 10: 95 users
['U9', 'U16', 'U21', 'U31', 'U38', 'U62', 'U76', 'U93', 'U104', 'U110', 'U297', 'U111', 'U114', 'U123', 'U138', 'U144', ' U156', 'U162', 'U176', 'U198', 'U199', 'U203', 'U208', 'U216', 'U217', 'U223', 'U224', 'U243', 'U246', 'U247', 'U25 0', 'U257', 'U270', 'U276', 'U282', 'U299', 'U300', 'U314', 'U318', 'U337', 'U347', 'U358', 'U363', 'U406', 'U434', 'U446', 'U468', 'U470', 'U471', 'U480', 'U513', 'U522', 'U524', 'U534', 'U546', 'U547', 'U556', 'U581', 'U586', 'U5 87', 'U595', 'U614', 'U630', 'U632', 'U675', 'U681', 'U682', 'U690', 'U728', 'U732', 'U746', 'U775', 'U780', 'U784' , 'U795', 'U808', 'U811', 'U847', 'U855', 'U863', 'U875', 'U880', 'U881', 'U883', 'U904', 'U906', 'U928', 'U938', ' U954', 'U958', 'U960', 'U961', 'U966', 'U987', 'U991']

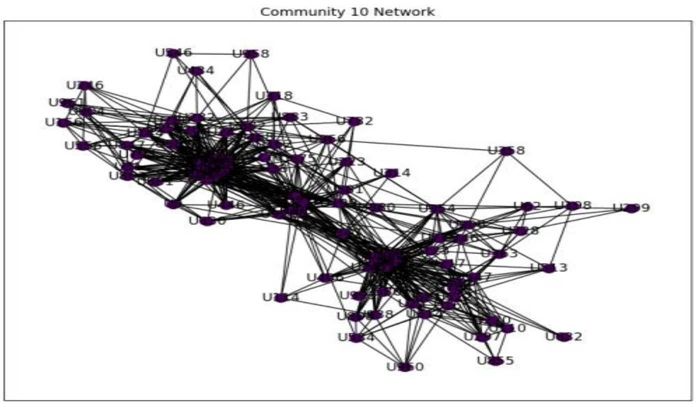

Figure11: community 10 network

Community 11: 36 users
['U22', 'U66', 'U91', 'U103', 'U119', 'U129', 'U137', 'U157', 'U249', 'U262', 'U320', 'U338', 'U383', 'U396', 'U418',
'U419', 'U429', 'U443', 'U466', 'U488', 'U517', 'U656', 'U668', 'U684', 'U694', 'U709', 'U802', 'U812', 'U828', '
U854', 'U856', 'U873', 'U887', 'U888', 'U943', 'U962']

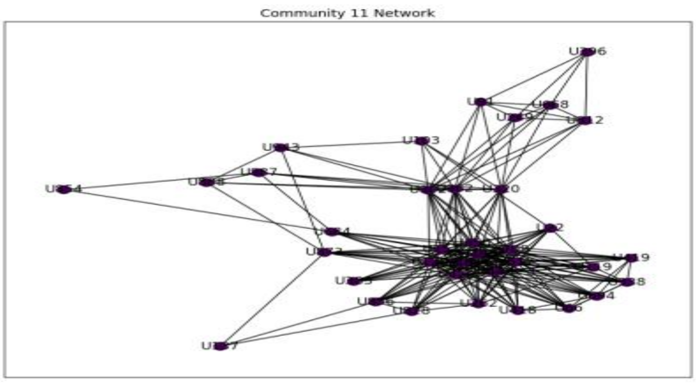

Figure12: community 10 network

Community 12: 70 users
['U54', 'U72', 'U78', 'U86', 'U87', 'U89', 'U145', 'U192', 'U218', 'U221', 'U222', 'U236', 'U265', 'U294', 'U305', 'U
311', 'U329', 'U345', 'U390', 'U399', 'U421', 'U472', 'U474', 'U479', 'U497', 'U507', 'U519', 'U594', 'U527', 'U529'
, 'U533', 'U539', 'U542', 'U564', 'U566', 'U573', 'U597', 'U608', 'U613', 'U634', 'U637', 'U648', 'U652', 'U659', 'U
693', 'U698', 'U701', 'U713', 'U742', 'U761', 'U781', 'U793', 'U794', 'U840', 'U844', 'U858', 'U859', 'U860', 'U86
4', 'U869', 'U916', 'U917', 'U920', 'U924', 'U929', 'U944', 'U945', 'U950', 'U982', 'U995']

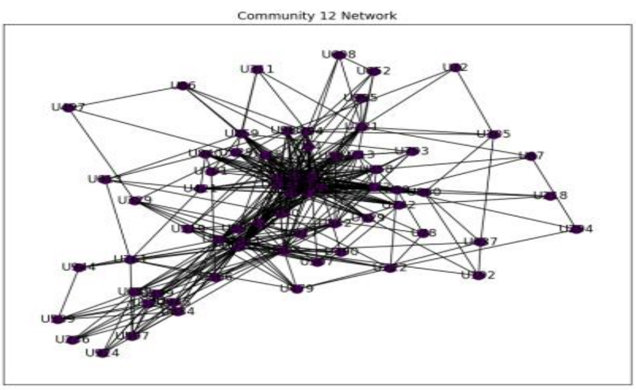

Figure13: community 0 network

Homophily in Social Networks: Label Propagation Algorithm (LPA)

This study employs the **Label Propagation Algorithm (LPA)** to detect community structures in social networks. The LPA is a semi-supervised machine learning algorithm that assigns labels to unlabeled nodes, enabling the classification and identification of communities within a network.The Label Propagation Algorithm is a community detection algorithm that works by propagating labels through a network, such that each node initially starts with a unique label. Over several iterations, nodes adopt the most frequent label among their neighbors, and this process continues until the labels converge (i.e., no node changes its label anymore).The following flowchart illustrates a Label Propagation Algorithm (LPA) for Community Detection.

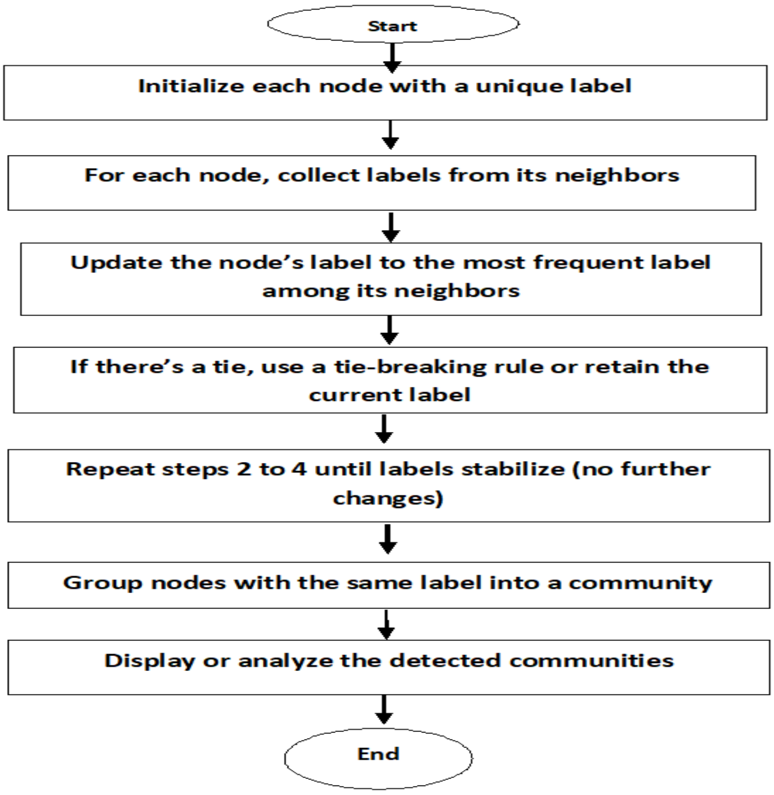

Figure 14: Flowchart illustrates a Label Propagation Algorithm (LPA)

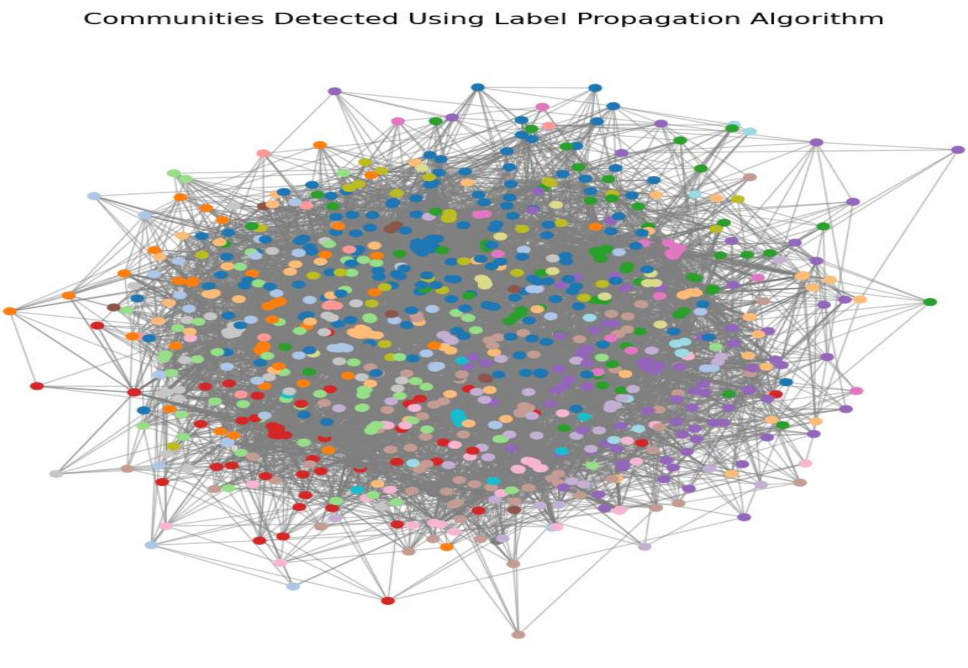

Figure 15: Social Network Visualization with Node Connections

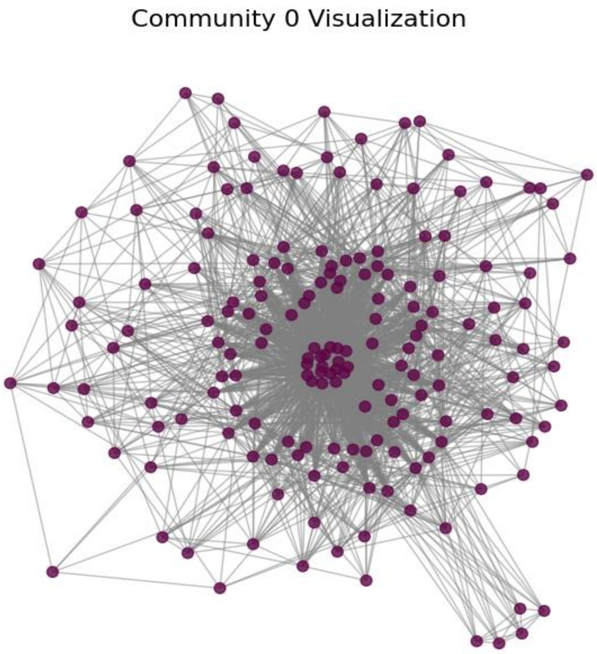

Figure16: .visualization of community 0 networks using LPA

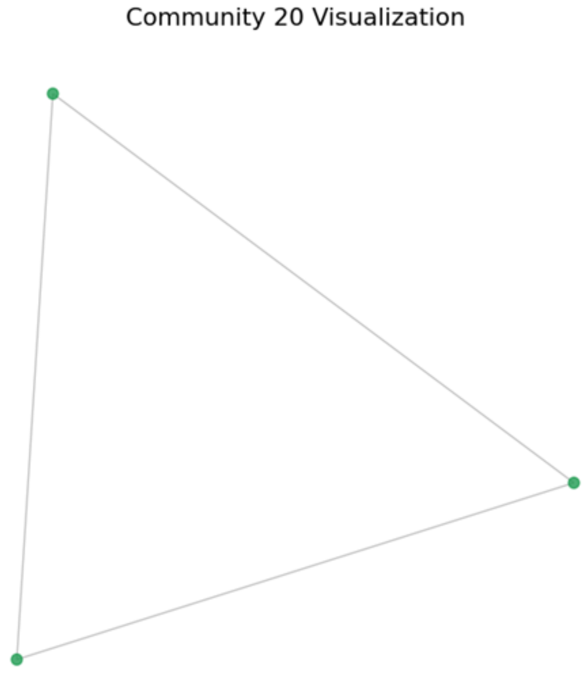

Figure17 : visualization of community 20 using LPA

Number of communities detected: 21

Community 0 (175 nodes): ['U362', 'U757', 'U158', 'U857', 'U325', 'U269', 'U475', 'U294', 'U989', 'U867', 'U213', 'U889', 'U39', 'U720', 'U215', 'U153', 'U401', 'U152', 'U664', 'U64', 'U368', 'U615', 'U850', 'U15', 'U80', 'U373', 'U578', 'U71', 'U650', 'U740', 'U766', 'U936', 'U751', 'U1', 'U58', 'U504', 'U625', 'U374', 'U953', 'U161', 'U3', 'U570', 'U509', 'U631', 'U621', 'U439', 'U540', 'U834', 'U182', 'U332', 'U308', 'U933', 'U359', 'U717', 'U132', 'U959', 'U357', 'U436', 'U894', 'U970', 'U610', 'U983', 'U767', 'U506', 'U673', 'U351', 'U56', 'U975', 'U295', 'U907', 'U185', 'U423', 'U776', 'U543', 'U666', 'U699', 'U692', 'U495', 'U410', 'U511', 'U526', 'U574', 'U229', 'U851', 'U568', 'U993', 'U912', 'U525', 'U375', 'U445', 'U688', 'U915', 'U237', 'U126', 'U437', 'U725', 'U726', 'U617', 'U951', 'U245', 'U196', 'U777', 'U531', 'U797', 'U151', 'U404', 'U943', 'U798', 'U304', 'U605', 'U973', 'U528', 'U510', 'U502', 'U491', 'U931', 'U687', 'U361', 'U671', 'U645', 'U220', 'U620', 'U635', 'U998', 'U697', 'U911', 'U633', 'U639', 'U477', 'U234', 'U459', 'U461', 'U447', 'U271', 'U79', 'U490', 'U765', 'U130', 'U238', 'U87', 'U603', 'U370', 'U473', 'U451', 'U730', 'U425', 'U910', 'U760', 'U885', 'U536', 'U790', 'U360', 'U133', 'U143', 'U749', 'U827', 'U861', 'U160', 'U627', 'U355', 'U384', 'U879', 'U897', 'U121', 'U306', 'U563', 'U727', 'U835', 'U955', 'U677', 'U596', 'U442', 'U813', 'U755', 'U882']

Community 1 (50 nodes): ['U187', 'U47', 'U736', 'U952', 'U59', 'U25', 'U102', 'U722', 'U575', 'U654', 'U939', 'U927', 'U647', 'U974', 'U50', 'U427', 'U43', 'U940', 'U60', 'U500', 'U971', 'U377', 'U303', 'U400', 'U94', 'U334', 'U241', 'U505', 'U838', 'U823', 'U914', 'U29', 'U514', 'U372', 'U662', 'U832', 'U499', 'U801', 'U259', 'U440', 'U876', 'U815', 'U602', 'U339', 'U273', 'U868', 'U696', 'U476', 'U382', 'U147']

Community 2 (44 nodes): ['U791', 'U805', 'U45', 'U128', 'U7', 'U836', 'U534', 'U501', 'U819', 'U572', 'U34', 'U272', 'U644', 'U512', 'U932', 'U681', 'U773', 'U976', 'U623', 'U941', 'U23', 'U284', 'U599', 'U392', 'U299', 'U972', 'U270', 'U95', 'U886', 'U560', 'U21', 'U768', 'U402', 'U559', 'U449', 'U242', 'U487', 'U565', 'U638', 'U712', 'U85', 'U485', 'U166', 'U97']

Community 3 (73 nodes): ['U53', 'U5', 'U414', 'U397', 'U481', 'U747', 'U100', 'U622', 'U364', 'U230', 'U117', 'U206', 'U181', 'U484', 'U450', 'U994', 'U651', 'U403', 'U328', 'U831', 'U317', 'U70', 'U636', 'U469', 'U584', 'U743', 'U520', 'U108', 'U561', 'U619', 'U380', 'U171', 'U61', 'U433', 'U201', 'U289', 'U649', 'U930', 'U674', 'U956', 'U735', 'U829', 'U548', 'U695', 'U913', 'U55', 'U806', 'U358', 'U979', 'U583', 'U710', 'U641', 'U188', 'U458', 'U653', 'U919', 'U954', 'U804', 'U576', 'U251', 'U33', 'U750', 'U210', 'U291', 'U667', 'U718', 'U660', 'U515', 'U774', 'U918', 'U389', 'U255', 'U274']

Community 4 (72 nodes): ['U99', 'U496', 'U640', 'U771', 'U81', 'U75', 'U231', 'U963', 'U981', 'U309', 'U655', 'U365', 'U702', 'U628', 'U278', 'U281', 'U277', 'U73', 'U10', 'U830', 'U611', 'U83', 'U521', 'U562', 'U263', 'U588', 'U523', 'U214', 'U8', 'U292', 'U430', 'U925', 'U48', 'U286', 'U550', 'U494', 'U821', 'U985', 'U98', 'U922', 'U839', 'U198', 'U1000', 'U518', 'U853', 'U324', 'U408', 'U266', 'U174', 'U412', 'U307', 'U744', 'U194', 'U663', 'U905', 'U741', 'U460', 'U763', 'U398', 'U754', 'U787', 'U753', 'U549', 'U532', 'U965', 'U302', 'U420', 'U837', 'U685', 'U489', 'U957', 'U541']

Community 5 (77 nodes): ['U604', 'U285', 'U593', 'U190', 'U738', 'U854', 'U553', 'U626', 'U997', 'U812', 'U606', 'U335', 'U205', 'U871', 'U684', 'U555', 'U424', 'U810', 'U848', 'U817', 'U110', 'U756', 'U91', 'U84', 'U146', 'U845', 'U2', 'U758', 'U731', 'U778', 'U843', 'U396', 'U411', 'U183', 'U968', 'U431', 'U252', 'U948', 'U217', 'U297', 'U668', 'U268', 'U17', 'U800', 'U715', 'U887', 'U704', 'U657', 'U716', 'U855', 'U803', 'U686', 'U522', 'U708', 'U547', 'U125', 'U872', 'U178', 'U595', 'U448', 'U346', 'U467', 'U177', 'U348', 'U136', 'U159', 'U513', 'U367', 'U254', 'U923', 'U154', 'U784', 'U769', 'U249', 'U366', 'U27', 'U632']

Community 6 (53 nodes): ['U298', 'U88', 'U612', 'U538', 'U874', 'U723', 'U106', 'U82', 'U453', 'U391', 'U498', 'U92', 'U343', 'U422', 'U432', 'U413', 'U141', 'U394', 'U822', 'U193', 'U992', 'U116', 'U109', 'U195', 'U942', 'U516', 'U41', 'U167', 'U42', 'U173', 'U387', 'U734', 'U6', 'U582', 'U577', 'U884', 'U13', 'U762', 'U244', 'U120', 'U464', 'U616', 'U530', 'U618', 'U280', 'U960', 'U228', 'U679', 'U169', 'U165', 'U191', 'U336', 'U28']

Community 7 (13 nodes): ['U131', 'U770', 'U508', 'U825', 'U293', 'U799', 'U90', 'U40', 'U899', 'U356', 'U700', 'U739', 'U112']

Community 8 (98 nodes): ['U566', 'U399', 'U218', 'U114', 'U917', 'U637', 'U78', 'U724', 'U693', 'U86', 'U564', 'U680', 'U390', 'U926', 'U608', 'U794', 'U46', 'U305', 'U945', 'U311', 'U421', 'U859', 'U519', 'U265', 'U742', 'U916', 'U435', 'U529', 'U781', 'U202', 'U349', 'U597', 'U507', 'U474', 'U533', 'U180', 'U950', 'U544', 'U659', 'U479', 'U385', 'U929', 'U937', 'U497', 'U844', 'U573', 'U924', 'U72', 'U652', 'U858', 'U89', 'U219', 'U344', 'U123', 'U472', 'U634', 'U840', 'U944', 'U345', 'U168', 'U982', 'U96', 'U698', 'U864', 'U761', 'U164', 'U340', 'U594', 'U139', 'U149', 'U313', 'U713', 'U539', 'U764', 'U542', 'U986', 'U222', 'U866', 'U613', 'U920', 'U938', 'U527', 'U192', 'U145', 'U995', 'U18', 'U319', 'U236', 'U841', 'U329', 'U221', 'U189', 'U54', 'U648', 'U869', 'U793', 'U935', 'U902']

Community 9 (46 nodes): ['U37', 'U186', 'U163', 'U62', 'U557', 'U906', 'U113', 'U775', 'U818', 'U428', 'U170', 'U4', 'U363', 'U57', 'U877', 'U591', 'U369', 'U379', 'U503', 'U386', 'U444', 'U52', 'U287', 'U691', 'U275', 'U486', 'U537', 'U792', 'U672', 'U833', 'U670', 'U264', 'U643', 'U878', 'U947', 'U624', 'U225', 'U895', 'U728', 'U256', 'U783', 'U759', 'U601', 'U701', 'U122', 'U592']

Community 10 (15 nodes): ['U184', 'U312', 'U204', 'U416', 'U226', 'U772', 'U107', 'U816', 'U211', 'U315', 'U310', 'U988', 'U607', 'U20', 'U261']

Community 11 (81 nodes): ['U690', 'U300', 'U682', 'U480', 'U987', 'U546', 'U223', 'U711', 'U928', 'U38', 'U732', 'U235', 'U119', 'U961', 'U208', 'U587', 'U282', 'U675', 'U318', 'U468', 'U454', 'U104', 'U93', 'U31', 'U144', 'U554', 'U24', 'U789', 'U689', 'U243', 'U780', 'U807', 'U65', 'U267', 'U630', 'U581', 'U471', 'U860', 'U881', 'U162', 'U111', 'U903', 'U434', 'U216', 'U862', 'U376', 'U199', 'U276', 'U883', 'U9', 'U314', 'U446', 'U703', 'U63', 'U338', 'U347', 'U811', 'U105', 'U600', 'U44', 'U246', 'U257', 'U585', 'U880', 'U224', 'U327', 'U545', 'U842', 'U966', 'U156', 'U203', 'U337', 'U847', 'U904', 'U958', 'U824', 'U795', 'U556', 'U482', 'U138', 'U863']

Community 12 (21 nodes): ['U258', 'U748', 'U175', 'U212', 'U381', 'U646', 'U737', 'U179', 'U438', 'U35', 'U371', 'U326', 'U977', 'U127', 'U669', 'U148', 'U896', 'U462', 'U809', 'U290', 'U579']

Community 13 (38 nodes): ['U714', 'U707', 'U552', 'U849', 'U30', 'U478', 'U19', 'U888', 'U405', 'U393', 'U665', 'U68', 'U609', 'U233', 'U239', 'U984', 'U417', 'U796', 'U678', 'U331', 'U786', 'U283', 'U140', 'U342', 'U908', 'U590', 'U135', 'U77', 'U407', 'U946', 'U934', 'U341', 'U580', 'U785', 'U949', 'U426', 'U352', 'U378']

Community 14 (21 nodes): ['U978', 'U898', 'U115', 'U483', 'U745', 'U719', 'U253', 'U323', 'U846', 'U990', 'U465', 'U388', 'U826', 'U26', 'U852', 'U350', 'U492', 'U172', 'U964', 'U733', 'U729']

Community 15 (38 nodes): ['U488', 'U962', 'U157', 'U103', 'U395', 'U456', 'U828', 'U443', 'U429', 'U466', 'U209', 'U320', 'U455', 'U814', 'U694', 'U709', 'U262', 'U129', 'U683', 'U900', 'U856', 'U656', 'U66', 'U197', 'U418', 'U873', 'U137', 'U782', 'U22', 'U419', 'U69', 'U16', 'U517', 'U890', 'U909', 'U11', 'U746', 'U802']

Community 16 (40 nodes): ['U642', 'U49', 'U705', 'U820', 'U232', 'U36', 'U921', 'U316', 'U155', 'U51', 'U980', 'U891', 'U142', 'U788', 'U567', 'U571', 'U892', 'U118', 'U493', 'U279', 'U124', 'U14', 'U779', 'U383', 'U248', 'U330', 'U752', 'U12', 'U296', 'U441', 'U996', 'U658', 'U463', 'U288', 'U353', 'U101', 'U409', 'U354', 'U452', 'U227']

Community 17 (17 nodes): ['U74', 'U676', 'U551', 'U457', 'U893', 'U322', 'U535', 'U661', 'U32', 'U967', 'U629', 'U706', 'U865', 'U200', 'U969', 'U569', 'U415']

Community 18 (13 nodes): ['U991', 'U875', 'U406', 'U586', 'U524', 'U614', 'U999', 'U808', 'U470', 'U247', 'U76', 'U250', 'U176']

Community 19 (12 nodes): ['U150', 'U721', 'U134', 'U260', 'U321', 'U901', 'U870', 'U240', 'U558', 'U333', 'U598', 'U301']

Community 20 (3 nodes): ['U589', 'U207', 'U67']

4. RESULTS AND DISCUSSION

A Comparative Study of Community Structures

Label Propagation: 21 communities detected
Louvain Method: 13 communities detected
The Louvain method successfully identified distinct communities within the dataset, reflecting groups of users with significant interaction patterns. The resulting community structure provides valuable insights into the network's organization and can be used for further analysis.
LPA identified **21 distinct communities** in the network. Notable findings include:

- **Community 0**: Consisted of 175 users with shared interests in books and science.
- **Community 20**: Centered around users interested in gaming and outdoor activities.

The results demonstrate that users with similar interests cluster together, reinforcing the role of homophily. LPA effectively detects these clusters, offering a scalable solution for community detection. The comparative study of community structures using the Label Propagation Algorithm (LPA) and the Louvain method revealed distinct patterns in network organization. The Louvain method identified 13 communities, capturing groups of users with significant interaction patterns. This approach successfully highlighted structurally cohesive groups within the dataset, providing valuable insights into the network's modularity and organization. By maximizing modularity, the Louvain method ensures that detected communities are well-defined and strongly interconnected. On the other hand, LPA detected 21 communities, indicating a finer level of granularity in community formation. Notable findings include Community 0, which consisted of 175 users with shared interests in books and science, and Community 20, which was centered around users engaged in gaming and outdoor activities. The results demonstrate that users with similar interests tend to cluster together, reinforcing the role of homophily in network formation. The higher number of communities detected by LPA suggests its ability to capture nuanced user interactions, making it a more scalable and adaptive solution for large-scale community detection.
Overall, both methods effectively identified community structures within the dataset. While the Louvain method provided a more consolidated view of the network by detecting fewer but well-defined communities, LPA captured more granular divisions, highlighting diverse user interests. These findings underscore the significance of homophily in shaping online communities and provide a foundation for further exploration of information diffusion and user engagement patterns.

5. CONCLUSION AND FUTURE DIRECTIONS

The study highlights the effectiveness of the **Label Propagation Algorithm (LPA)** in detecting communities within social networks, while also demonstrating the utility of the **Louvain method** for identifying structurally cohesive groups. The results reinforce the significance of **homophily**, as users with similar interests tend to cluster together. Although LPA efficiently identifies communities, alternative approaches such as the **Louvain algorithm** or **spectral clustering** could provide different perspectives on network structure. Future research should explore comparative analyses of these methods to enhance the accuracy and robustness of community detection techniques.

Challenges and Limitations

Several challenges were observed during the study:

- **Threshold Sensitivity**: The **correlation threshold (0.5)** plays a crucial role in defining community structures. A **sensitivity analysis** could help refine this parameter to optimize detection accuracy.
- **Initial Label Dependency**: The effectiveness of **LPA** may vary depending on the **initial label assignment**, potentially influencing the final community structure.
- **Scalability**: While **LPA** is computationally efficient, its application to **large-scale datasets** requires further **optimization** to handle increasing network complexity.

Addressing these challenges will contribute to improving the reliability of community detection techniques, facilitating more accurate insights into **social network structures** and **information diffusion**

patterns. Future work may also explore hybrid approaches that integrate multiple community detection methods to enhance performance across diverse datasets.

REFERENCE

1. Althoff T, Danescu-Niculescu-Mizil C, Jurafsky D. How to ask for a favor: A case study on the success of altruistic requests. In: Proceedings of the International Conference on Web and Social Media (ICWSM); 2016. p. 12-21.
2. Bhadoria, R. S., Bhoj, N., Srivastav, M. K., Kumar, R., & Raman, B. (2022). A machine learning framework for security and privacy issues in building trust for social networking. *Cluster Computing, 26*, 3907-3930. doi: 10.1007/s10586-022-03517-3
3. Bakshy E, Rosenn I, Marlow C, Adamic L. The role of social networks in information diffusion. In: *Proceedings of the 21st International Conference on World Wide Web*; 2012. p. 519-528.
4. Cota W, Ferreira SC, Pastor-Satorras R, Starnini M. Quantifying echo chamber effects in information spreading over political communication networks. *EPJ Data Sci*. 2019;8(1):35.
5. Gillani N, Yuan A, Saveski M, Vosoughi S, Roy D. Me, my echo chamber, and I: Introspection on social media polarization. In: *Proceedings of the 2018 World Wide Web Conference*; 2018. p. 823-831.
6. Ljepava N, Orr RR, Locke S, Ross C. Personality and social characteristics of Facebook non-users and frequent users. *Comput Hum Behav*. 2013;29(4):1602-1607.
7. Song C, Qu Z, Blumm N, Barabási AL. Limits of predictability in human mobility. *Science*. 2016;327(5968):1018-1021.
8. Hughes EC. Dilemmas and contradictions of status. *Am J Sociol*. 1945;50(4):353-369.
9. Feld SL. The structured use of personal associates. *Soc Forces*. 1984;62(3):640-652.
10. Cialdini RB, Goldstein NJ. Social influence: Compliance and conformity. *Annu Rev Psychol*. 2004;55:591-621.
11. Yodovich N. Cultural homophily and social networks: A study of Israelis in London. *J Ethn Migr Stud*. 2020;46(12):2595-2611.
12. Cinelli M, Morales GDF, Galeazzi A, Quattrociocchi W, Starnini M. Echo chambers on social media: A comparative analysis. *J Data Inf Qual*. 2020;12(3):1-27.
13. Ferrara E, Yang Z, Cortes C. Measuring the effects of social media homophily on the spread of misinformation during the COVID-19 pandemic. *IEEE Trans Netw Sci Eng*. 2022;9(2):1-11.
14. Masuda N, Konno N. Networks with disparate degrees of homophily and heterophily. *Phys Rev E*. 2006;74(6):066106.
15. Valdés L, Lara CM, Borge-Holthoefer J. Understanding homophily in social networks: Theories and evidence. *Soc Netw*. 2020;63:101-108.
16. Alamsyah A, Rahardjo B, Kuspriyanto. Social network analysis taxonomy based on graph representation. *arXiv preprint arXiv:2102.08888*. 2021.
17. Dhotre IA. Social network analysis. *Technical Publication*. 2021.
18. Fortunato S. Community detection in graphs. *Phys Rep*. 2010;486(3-5):75-174.
19. Ganis M, Kohirkar A. *Social media analytics*.Pearson. 2017.
20. Kouznetsov A, Tsvetovat M. *Social network analysis for startups*. 2011.
21. Salaria RS. *Programming in Python*. Khanna Book Publishing Co. (P) Ltd. 2019.
22. Scott J. *Social network analysis*. SAGE Publications. 2017.
23. Srinivasraghavan A, Joseph V. *Machine learning*. Wiley. 2020.
24. Srivastav MK, Gupta S. An approach for exploring practical phenomena in social network analysis. *J Mines Met Fuels*. 2023;71(5):583-587.
25. Srivastav MK, Nath A. A comprehensive study on complexity of relationship in social networks. In: *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*; 2015. p. 718-723.
26. Srivastav MK, Nath A. Mathematical modeling of virtual relation in social networks. In: *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*; 2015. p. 924-928.
27. Srivastav MK, Nath A. Mathematical modeling of social networks: A probabilistic approach to explore relationships in social networks. In: *2016 8th International Conference on Computational Intelligence and Communication Networks (CICN)*; 2016. p. 346-350.
28. Srivastav MK, Gupta S, Priyadharshini VM, Som S, Acharya B, Gerogiannis VC, Kanavos A, Karamitsos I. Utilizing degree centrality measures for product advertisement in social networks. In: *European, Mediterranean and Middle Eastern Conference on Information Systems*; 2023.
29. Srivastav MK, Gupta S. Analysis and prediction of friendship model in social networks like Facebook using binomial distribution and supervised learning. *SSRN Electron J*. 2022.
30. Sridhar S, Vijayalakshmi M. *Machine learning*. Oxford University Press. 2021.
31. Wasserman S, Faust K. *Social network analysis: Methods and applications*. Cambridge University Press. 1994.

32. Sahoo A. *Social-Media-Users-Dataset*. Kaggle. 2023. (Available at https://www.kaggle.com/datasets/arindamsahoo/social-media-users).
33. Abdullah M, Alissa K, Alzoubi O. New user engagement prediction using machine learning algorithms. *2023 14th International Conference on Information and Communication Systems (ICICS)*, 1–6. https://doi.org/10.1109/ICICS60529.2023.10330488
34. Alkhamees M, Alsaleem S, Al-Qurishi M. User trustworthiness in online social networks: A systematic review. *Applied Soft Computing*. 2021;103:107159. https://doi.org/10.1016/j.asoc.2021.107159
35. Barabási AL. Emergence of scaling in random networks. *Science*. 1999;286(5439):509–512. https://doi.org/10.1126/science.286.5439.509
36. Brahim L, Loubna B, Ali I. A literature survey on label propagation for community detection. *2021 Fifth International Conference on Intelligent Computing in Data Sciences (ICDS)*, Fez, Morocco, 1–7. https://doi.org/10.1109/ICDS53782.2021.9626716
37. Cheng YH, Ho HY. Social influence's impact on reader perceptions of online reviews. *J Bus Res*. 2015;68(4):883-887. https://doi.org/10.1016/j.jbusres.2014.11.046
38. Cifuentes-Faura J, Faura-Martínez U, Lafuente-Lechuga M. Mathematical modeling and the use of network models as epidemiological tools. *Mathematics*. 2022;10(18):3347. https://doi.org/10.3390/math10183347
39. Crandall DJ, Backstrom L, Cosley D, Suri S, Huttenlocher D, Kleinberg J. Inferring social ties from geographic coincidences. *Proc Natl Acad Sci USA*. 2010;107(52):22436–22441. https://doi.org/10.1073/pnas.1006155107