Ecole Nationale Supérieure d'Informatique et d'Analyse des Systèmes (ENSIAS)

Ingénierie Intelligence artificielle (2IA)
Semestre 4
Unsupervised Learning prject
Report

Unveiling the Twitter-Sphere: Community Detection Analysis

Project members:
IFQIR MOHAMED
SEDJARI YASSINE

Project supervisor :
Pr. LAZAAR MOHAMED

School year 2022/2023

Abstract

Community detection in online social networks, specifically Twitter, has gained attention for uncovering underlying structures and patterns in user interactions. This report presents a comprehensive study on Twitter community detection, aiming to enhance our understanding of dynamics and characteristics, user behavior, information dissemination, and social influence. It discusses the need for community detection techniques due to the challenges of extracting meaningful information at scale. The report explores methodologies, algorithms, and factors influencing Twitter community structures. Techniques including Spectral Clustering in the Edge-based Approach (Silhouette Score: -0.1887), Spectral Clustering in the Feature-based Approach (Silhouette Score: -0.0343), Hierarchical Clustering (Silhouette Score: 0.6756), and K-Means (Silhouette Score: 0.6917) were compared. The comparative analysis revealed that K-Means outperformed the other techniques, demonstrating the highest Silhouette Score. The report validates these techniques through experiments on real-world datasets. In conclusion, this research contributes to online social network analysis by providing insights into the performance and effectiveness of different community detection techniques, with K-Means emerging as the most successful approach for Twitter community detection. It offers valuable insights for industry professionals and policymakers leveraging Twitter communities.

Résumé

La détection de communautés dans les réseaux sociaux en ligne, en particulier sur Twitter, a suscité l'attention en révélant les structures sous-jacentes et les schémas des interactions entre les utilisateurs. Ce rapport présente une étude approfondie sur la détection des communautés sur Twitter, dans le but d'améliorer notre compréhension des dynamiques et des caractéristiques, du comportement des utilisateurs, de la diffusion de l'information et de l'influence sociale. Il aborde la nécessité de techniques de détection des communautés en raison des défis liés à l'extraction d'informations significatives à grande échelle. Le rapport explore les méthodologies, les algorithmes et les facteurs influençant les structures de communautés sur Twitter. Les techniques, comprenant le regroupement spectral dans l'approche basée sur les arêtes (Score de Silhouette : -0,1887), le regroupement spectral dans l'approche basée sur les caractéristiques (Score de Silhouette : -0,0343), le regroupement hiérarchique (Score de Silhouette : 0,6755) et le K-Means (Score de Silhouette : 0,6917), ont été comparées. L'analyse comparative a révélé que le K-Means surclasse les autres techniques, en démontrant le score de Silhouette le plus élevé. Le rapport valide ces techniques à travers des expériences sur des ensembles de données réels. En conclusion, cette recherche contribue à l'analyse des réseaux sociaux en ligne en fournissant des informations sur les performances et l'efficacité des différentes techniques de détection de communautés, avec le K-Means émergent comme l'approche la plus réussie pour la détection de communautés sur Twitter. Elle offre des informations précieuses aux professionnels de l'industrie et aux décideurs qui exploitent les communautés Twitter.

Introduction

Social media platforms have revolutionized the way we connect, communicate, and share information on a global scale. Among these platforms, Twitter has emerged as a powerful tool for real-time information dissemination, fostering conversations, and engaging diverse communities. With millions of users actively participating in discussions, understanding the intricate network of interactions within the Twitter ecosystem has become a fascinating area of research.

One fundamental aspect of studying Twitter is the detection and analysis of communities within its vast user base. Communities on Twitter can be defined as groups of users who exhibit similar interests, engage in similar conversations, or share common characteristics. Identifying these communities is crucial for various applications, including targeted marketing, content personalization, and social network analysis.

The purpose of this report is to delve into the realm of Twitter community detection and explore the methods and techniques used to uncover hidden patterns and structures within the Twittersphere. By employing sophisticated algorithms and network analysis tools, we aim to shed light on the intricate web of relationships that form among Twitter users and provide valuable insights into the dynamics of online communities.

Through this study, we seek to address several key research questions. How can we identify and define communities within the vast Twitter user network? What factors contribute to the formation and evolution of these communities? How do communities interact and influence each other within the Twittersphere? Moreover, we aim to examine the implications of community detection for various

stakeholders, such as researchers, marketers, and policymakers, and highlight the potential benefits and challenges associated with this analysis.

To achieve these objectives, we will review existing literature on community detection in Twitter and explore different methodologies employed in previous studies. We will also discuss the limitations and ethical considerations associated with Twitter community detection, emphasizing the need for responsible research practices to safeguard user privacy and mitigate potential biases.

Ultimately, this report aims to contribute to the growing body of knowledge surrounding Twitter community detection and provide a foundation for future research in understanding the dynamics and characteristics of online communities. By unraveling the interconnectedness of Twitter users and exploring the communities that emerge within this complex network, we hope to gain a deeper understanding of the societal impact of social media and pave the way for more effective strategies for engagement, information dissemination, and community building.

List of Figures

Contents

Chapter 1

Communities in social media

1.1 Definition of communities in social media

In the context of social media, communities can be understood as groups of users who share common interests, engage in similar discussions, and exhibit patterns of interaction within a specific platform or network. These communities are characterized by the formation of relationships, the exchange of information, and the establishment of social connections among their members.

- Shared Interests and Topics:
  Communities in social media often revolve around specific interests, topics, or themes. Users with similar passions or hobbies gather within these communities to share their knowledge, experiences, and opinions. For example, there might be communities dedicated to sports, technology, fashion, or music, where users engage in conversations related to these subjects.

- Patterns of Interaction:
  Communities in social media can also be identified based on the patterns of interaction among users. Members of a community tend to interact with each other more frequently compared to interactions with users outside the community. They might reply to each other's posts, mention one another, or engage in discussions within specific hashtags or threads.

- Social Connections and Networks:
  Communities in social media are built upon social connections and networks. Users within a community often follow each other, retweet or share each other's content, and form relationships that go beyond mere interactions.

These social connections contribute to the cohesion and strength of the community, fostering a sense of belonging and shared identity.

- Size and Scale:
  Communities in social media can range from small, tightly-knit groups to large-scale communities with thousands or even millions of members. The size of a community influences its dynamics, level of activity and the diversity of opinions and perspectives within it.

It's important to note that communities in social media are not limited to geographic or physical boundaries. They transcend traditional barriers, allowing individuals from diverse backgrounds and locations to connect and form communities based on shared interests or goals.

Understanding the concept of communities in social media is essential for various stakeholders, including researchers, marketers, and platform developers. By identifying and studying these communities, researchers can gain insights into user behavior, information dissemination patterns, and the formation of online social structures. Marketers can leverage community detection to target specific audiences and tailor their strategies accordingly, while platform developers can enhance the user experience by facilitating community-building features and fostering a sense of belonging within their platforms.

In the subsequent sections, we will delve into the methods and approaches used for detecting communities in social media platforms, highlighting the techniques employed to unveil these hidden structures within the vast networks of users.

1.2 Characteristics and Attributes of Online Communities

Online communities exhibit various characteristics and possess unique attributes that differentiate them from traditional, offline communities. Understanding these characteristics and attributes is crucial for effectively studying, analyzing, and leveraging online communities. Here are some key features:

1. **Shared Interests:** Online communities are formed around shared interests, passions, or goals. Members join these communities to connect with like-minded individuals who share similar hobbies, professions, or causes. This

shared interest acts as a common thread that binds community members together.

2. **Virtual Presence:** Unlike physical communities, online communities exist in a virtual space. Members interact, communicate, and collaborate through digital platforms, such as social media platforms, forums, or specialized websites. This virtual presence enables people from different geographic locations to connect and engage with one another.

3. **Anonymity and Pseudonymity:** Online communities often provide a level of anonymity or pseudonymity to their members. This allows individuals to express themselves freely without the fear of social repercussions or judgment. Anonymity can encourage open discussions and diverse perspectives within the community.

4. **Scalability:** Online communities have the potential to scale rapidly. With the widespread adoption of the internet and social media platforms, online communities can attract a large number of participants from diverse backgrounds and locations. This scalability creates opportunities for broader reach, knowledge sharing, and collaboration on a global scale.

5. **Asynchronous Communication:** Online communities offer asynchronous communication channels, allowing members to engage in discussions and activities at their convenience. This flexibility accommodates members from different time zones and with varying schedules. It also enables more thoughtful and reflective contributions compared to real-time interactions.

6. **Data and Information Sharing:** Online communities thrive on the exchange of information and knowledge. Members contribute valuable insights, share resources, and discuss relevant topics within the community. This collective intelligence enhances the community's overall expertise and benefits individual members.

7. **Social Influence and Reputation:** Within online communities, members often establish social influence and develop reputations based on their contributions, expertise, and level of engagement. Influential members may have a significant impact on shaping discussions, setting trends, and guiding community norms.

8. **Self-Organization:** Online communities tend to exhibit a self-organizing nature. Members autonomously create and curate content, moderate discussions, and establish community rules. This decentralized structure promotes a sense of ownership and empowerment among community members.

9. **Diverse Participation:** Online communities have the potential to foster diverse participation. Individuals from different cultural, social, and demographic backgrounds can engage and contribute within these communities. This diversity of perspectives enriches discussions and fosters inclusive environments.

10. **Longevity and Evolution:** Online communities can exhibit long-term sustainability and evolve over time. Successful communities adapt to changing needs, technological advancements, and member interests. They may undergo shifts in membership, focus, or platform while maintaining their core identity.

Understanding the characteristics and attributes of online communities is crucial for various stakeholders, including researchers, community managers, marketers, and policymakers. By grasping these unique features, they can effectively analyze, support, and harness the power of online communities for research, business, and social purposes.

Chapter 2

Dataset description

2.1 Overview

The SNAP Twitter dataset is a comprehensive collection of anonymized Twitter data that has been widely used for research in the field of social network analysis. It provides a valuable resource for studying various aspects of Twitter networks, user behavior, and information diffusion. The dataset is derived from a large-scale crawl of publicly available tweets and includes information such as user profiles, follower relationships, and tweet content.

2.2 Data Content

The dataset consists of several interconnected data files, each capturing different aspects of the Twitter network. These files include:

- **User Profiles:** This file contains information about Twitter users, such as their unique user IDs, screen names, profile descriptions, and the number of followers and followings. User profiles provide insights into the characteristics and attributes of individuals within the network.
- **Follower Relationships:** This file captures the follower relationships between Twitter users. It represents the directed edges of the Twitter network, where each line contains a pair of user IDs indicating a follower-followee relationship. This information is crucial for analyzing the structure and connectivity of the network.
- **Tweet Data:** The dataset includes a collection of tweets, which consists of the actual textual content posted by users. Each tweet is associated

with metadata such as the user ID, timestamp, retweet count, and favorite count. This information enables researchers to study tweet content, sentiment analysis, and information propagation within the network.

- **Hashtag Data:** Hashtags play a significant role in Twitter conversations and information organization. The dataset includes a file that contains hashtags associated with each tweet. Hashtags allow researchers to analyze trends, topic clustering, and user interests within the network.

In conclusion, the SNAP Twitter dataset is a comprehensive collection of anonymized Twitter data, providing researchers with a valuable resource to explore various aspects of social network analysis on Twitter. With its user profiles, follower relationships, tweet content, and hashtag data, the dataset enables researchers to gain insights into network structures, user behavior, information diffusion, and topic trends within the Twitter platform.

2.3 Dataset Statistics

The dataset used for the analysis contains a Twitter network with the following statistics:

Statistic	**Value**
Nodes in the network	81,306
Edges in the network	1,768,149
Nodes in the largest weakly connected component (WCC)	81,306 (100%)
Edges in the largest WCC	1,768,149 (100%)
Nodes in the largest strongly connected component (SCC)	68,413 (84.1%)
Edges in the largest SCC	1,685,163 (95.3%)
Average clustering coefficient	0.5653
Number of triangles	13,082,506
Fraction of closed triangles	0.06415
Diameter (longest shortest path)	7
90-percentile effective diameter	4.5

Table 2.1: Dataset statistics of the Twitter network.

The dataset consists of 81,306 nodes, representing Twitter users, and 1,768,149 edges, representing the connections or interactions between these users. The

network is analyzed based on its weakly connected component (WCC) and strongly connected component (SCC).

The largest WCC encompasses all 81,306 nodes in the network, indicating that every user is connected, directly or indirectly, to at least one other user. Similarly, the largest SCC consists of 68,413 nodes, forming a strongly connected subgraph where every user can reach every other user within the component.

The average clustering coefficient of the network is 0.5653, which indicates a moderate level of clustering or the tendency for nodes to form tightly-knit groups. This suggests the presence of communities or subgroups within the network.

The network contains a significant number of triangles, with 13,082,506 identified. A triangle represents a set of three nodes where each node is connected to the other two, indicating a higher likelihood of mutual connections and local clustering.

The diameter of the network, which is the length of the longest shortest path between any two nodes, is 7. This implies that the longest direct path between any pair of users in the network requires traversing through at most seven intermediate users.

The 90-percentile effective diameter of 4.5 indicates that the majority of the network can be reached within a relatively short path length of 4.5, demonstrating efficient communication and information flow within the network.

In summary, the dataset statistics reveal important structural characteristics of the Twitter network, including the presence of interconnected components, moderate clustering, numerous triangles representing local clustering, and efficient communication paths. These insights lay the foundation for further analysis and community detection within the network.

Chapter 3

Method 1: Detecting communities using the edges only

Figure 3.1: Graph network

3.1 Approach description

This approach to community detection is based on using a subset of edges from the Twitter network. The process can be described as follows:

1. **Creating the Node Set:** The code initializes an empty set, `nodes`, to store unique node IDs. It iterates through the given subset of edges and adds the source and target nodes to the set. This step ensures that all distinct nodes present in the network are accounted for.

2. **Determining the Number of Nodes:** After collecting the nodes in the set, the code calculates the total number of nodes, `num_nodes`, by obtaining the length of the `nodes` set. This value is crucial for creating the adjacency matrix with the appropriate dimensions.

3. **Mapping Node IDs to Indices:** A dictionary, `node_to_index`, is created to map each node ID to its corresponding index in the adjacency matrix. This mapping allows for efficient access and manipulation of matrix elements based on the node IDs.

4. **Building the Adjacency Matrix:** Using the `lil_matrix` class from the `scipy.sparse` module, an empty adjacency matrix is initialized with dimensions `(num_nodes, num_nodes)`. This matrix represents the connections between nodes in the network. The data type is specified as `np.int8` to optimize memory usage. This approach then iterates through the subset of edges again, retrieving the source and target nodes. It utilizes the `node_to_index` dictionary to obtain the corresponding indices for these nodes in the adjacency matrix. In the process, it assigns a value of 1 to the corresponding element in the matrix, indicating the presence of an edge between the two nodes.

5. **Converting to Compressed Sparse Row (CSR) Matrix:** Finally, the code converts the adjacency matrix to a compressed sparse row (CSR) matrix format using the `tocsr()` method. This format enables efficient computations and optimal memory usage for subsequent community detection algorithms.

This approach establishes the foundation for performing community detection on the Twitter network represented by the adjacency matrix. It prepares the data

for further analysis and enables the application of various community detection algorithms to reveal the underlying community structure within the network.In the next section, we are going to be tackling the training process and see the ups and downs of this method.

3.2 Training process: Spectral clustering algorithm

3.2.1 Pseudo-code

```
-----------------------------------------------------------------------
Input: Adjacency matrix, number of clusters (k)
-----------------------------------------------------------------------
Construct the graph Laplacian matrix L from the adjacency matrix.
Compute the eigenvectors corresponding to the k smallest eigenvalues of L.
Stack the eigenvectors into a matrix X.
Normalize the rows of X to have unit norm.
Apply k-means clustering to the rows of X to obtain the final
cluster assignments.
-----------------------------------------------------------------------
Output the cluster labels.
-----------------------------------------------------------------------
```

The spectral clustering algorithm is a popular method for community detection. Here's an overview of the steps involved:

Constructing the graph Laplacian matrix: The adjacency matrix is used to construct the graph Laplacian matrix, which captures the relationship between nodes in the network. There are different types of graph Laplacians, such as the unnormalized Laplacian, normalized Laplacian, and symmetric normalized Laplacian, each with its own properties.

Computing eigenvectors: The eigenvectors corresponding to the k smallest eigenvalues of the Laplacian matrix are computed. These eigenvectors capture important structural information about the network and can be used to reveal underlying community structure.

Stacking eigenvectors: The computed eigenvectors are stacked into a matrix, often referred to as the embedding matrix or feature matrix. Each row of this

matrix represents a node in the network, and the columns correspond to the different eigenvectors.

Normalizing rows: The rows of the embedding matrix are normalized to have a unit norm. This normalization step ensures that each node's representation has equal importance in the subsequent clustering process.

Applying k-means clustering: K-means clustering is applied to the rows of the normalized embedding matrix. This step groups similar nodes together into k clusters, where k is the desired number of clusters specified as input.

Output: The final cluster assignments, represented as cluster labels, are obtained from the k-means clustering process. These labels indicate which cluster each node belongs to and provide insights into the community structure of the network.

The spectral clustering algorithm leverages the spectral properties of the graph Laplacian matrix to uncover communities in networks. By computing eigenvectors and applying clustering techniques, it offers an effective approach to community detection in various types of networks.

3.2.2 Flowchart

Here's a flowchart detailing the workflow of the spectral algorithm:

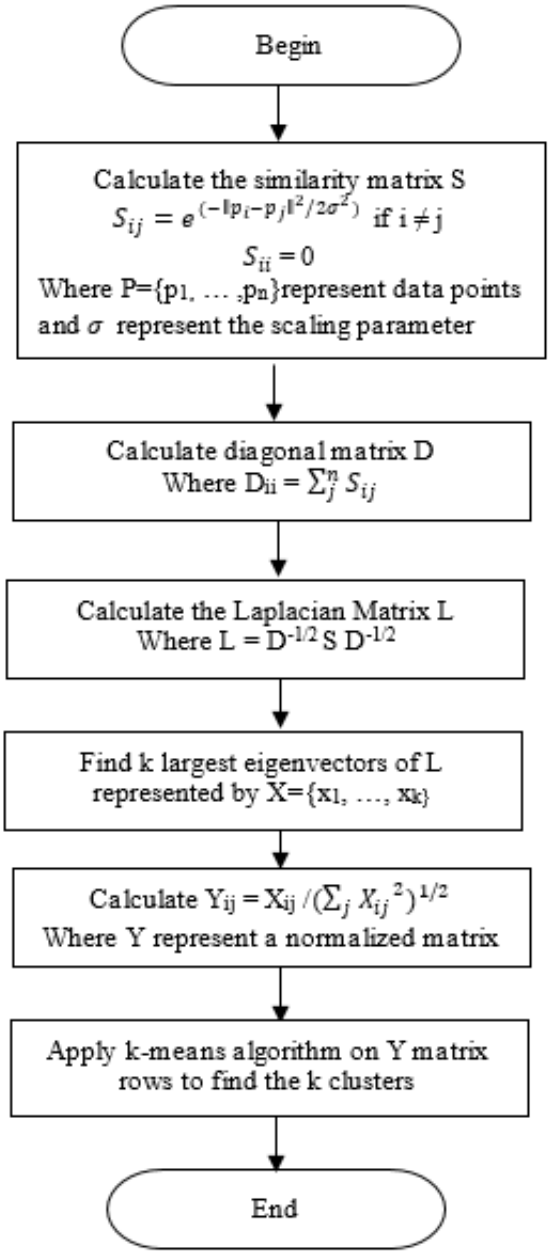

Figure 3.2: Spectral clustering algorithm flow chart

Chapter 4

Method 2: Feature-based approach

In this chapter, we are going to tackle a different approach and a more effective one; in which we will be using user attributes such as hashtags, followings, mentions .etc

4.1 Node feature extraction

The feature-based approach, utilized in this approach, aims to extract and analyze the features of nodes in a network. The process can be outlined as follows:

Node Feature Extraction: For each node in the network, the approach iterates through the available node IDs. It then identifies the file paths of the corresponding feature files: the file containing feature names (featnames) and the file containing the ego features (egofeat).

Read Ego Features: The approach opens and reads the contents of the egofeat file, which contains the features of the current node. The file is read and split to obtain a list of ego features associated with the node.

Read Feature Names: Similarly, the approach opens and reads the contents of the featnames file, which provides the names of the features. The file is read line by line, and each line is processed to extract the feature name. These feature names are collected and stored in a list.

Feature-Node Mapping: The approach then proceeds to iterate through the indices and values of the ego features that have a value of one. For each index, representing a feature, the corresponding feature name is retrieved from the list of feature names.

Feature-Node Hash Map: The feature name and the associated node ID

are stored in a hash map (denoted as hm). If the feature name does not already exist in the hash map, a new entry is created, and the node ID is added to the list of values associated with that feature name. If the feature name already exists in the hash map, the node ID is appended to the existing list of values.

The feature-based approach implemented in this approach focuses on extracting and organizing the features of nodes in the network. By mapping the feature names to the corresponding nodes, this approach facilitates further analysis and investigation of the network's characteristics based on these features.

	nodeId	#OCTAVIA	#THEHELP	#ff	@BAFTA	@FuckYesEmma	@JUDAOcombr	@astowellcom	@emmastonebr	@helpmovie	...	@chococat
968	14528221	0	0	1	0	0	0	0	0	0	...	0
969	14840869	0	0	0	0	0	0	0	0	0	...	0
970	82726142	0	0	0	0	0	0	0	0	0	...	0
971	255790981	0	0	0	0	0	0	0	0	0	...	0
972	36618690	0	0	0	0	0	0	0	0	0	...	1

@helpmovie	...	@chococat	@dannyBstyle,	@jon_blaze55	@kylepulver:	@terrycavanagh	@twitchtv.	Degree Centrality	Closeness Centrality	Betweenness Centrality
0	...	0	0	0	0	0	0	0.001030	0.238165	0.000000
0	...	0	0	0	0	0	0	0.003090	0.239990	0.001701
0	...	0	0	0	0	0	0	0.002060	0.206245	0.000958
0	...	0	0	0	0	0	0	0.009269	0.273984	0.002988
0	...	1	1	1	1	1	1	0.010299	0.294064	0.010228

Figure 4.1: Result of Transforming Graph Data to Tabular Data

4.2 The clustering phase

Now that we have each feature mapped to its corresponding node, we will proceed to train our model using different clustering approaches with the aspiration of finding the best clusters.

In the next sections, we will delve deep into the conceptual and technical side of every method used in detecting communities on Twitter.

4.2.1 Graph centrality

In this section, we will be exploring our graph network and identifying the centrality of each node;thus, quantifying the importance or influence of individual nodes within the network. Centrality in a graph refers to a measure that helps identify nodes that are more central or influential in terms of their connectivity and

position in the graph structure. Centrality measures can provide insights into various aspects of a graph, such as node prominence, information flow, and structural importance.

1. Degree centrality:

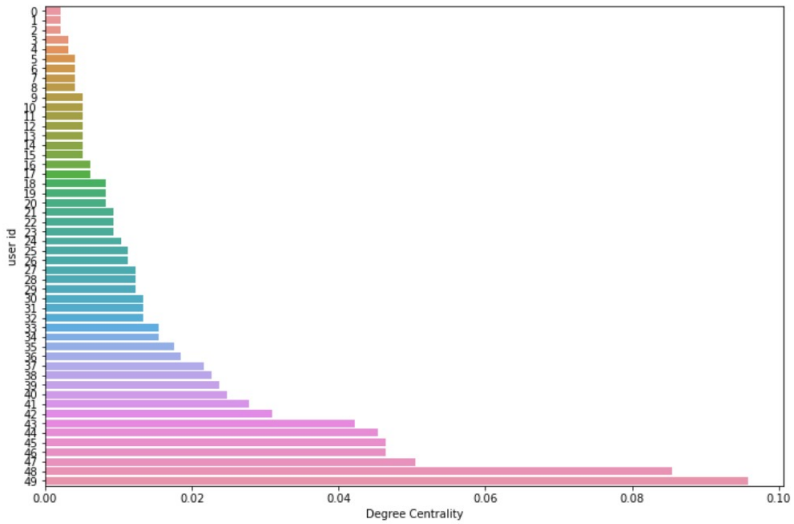

Figure 4.2: Degree centrality

2. Betweenness centrality:

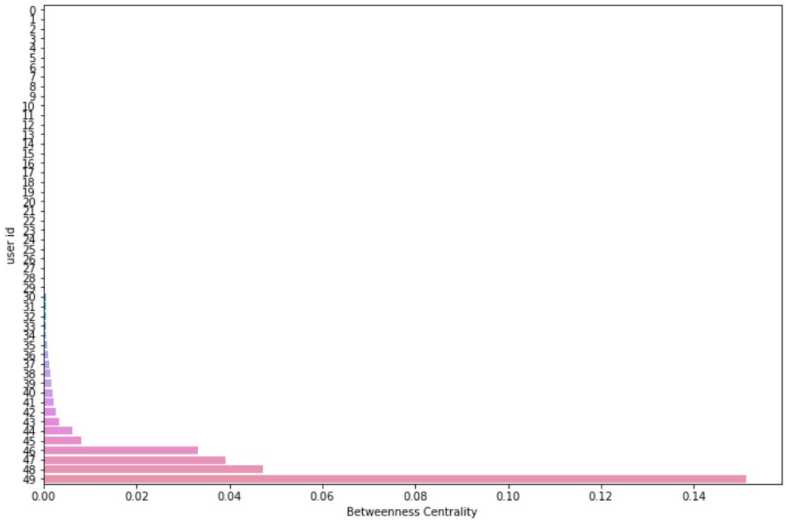

Figure 4.3: Betweenness centrality

3. Closeness centrality

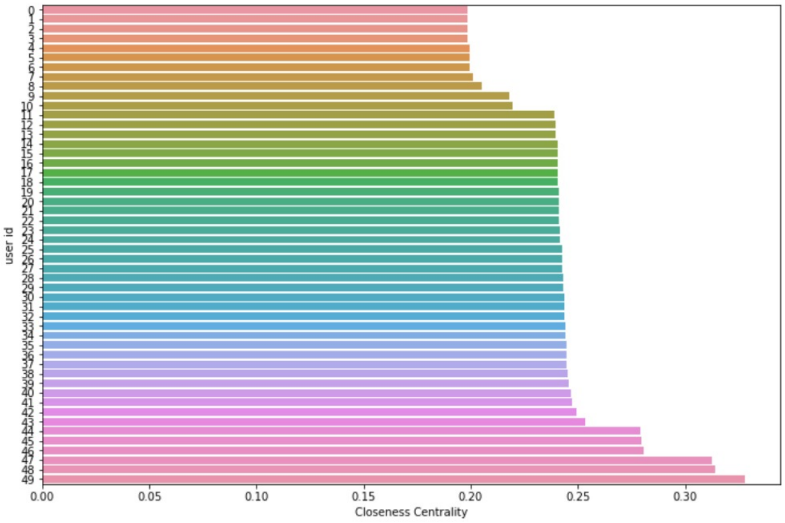

Figure 4.4: Closeness centrality

4.2.2 K-means algorithm

K-means is a popular clustering algorithm that can also be applied effectively in community detection tasks. In the context of community detection, k-means can be leveraged to group nodes into communities based on their similarity in feature space. The algorithm starts by randomly initializing a set of centroids, each representing a potential community. It then iteratively assigns each node to the nearest centroid and updates the centroids based on the mean values of the nodes assigned to them. This process continues until convergence is achieved. By optimizing the within-cluster variance, k-means efficiently identifies communities by maximizing the cohesion within each community while minimizing the separation between different communities. Despite its simplicity, k-means can effectively uncover communities in various networks, making it a valuable tool in community detection research and applications. Here's a pseudo-code detailing the flow of actions in this algorithm:

Algorithm 1 K-means algorithm for community detection

Require: Graph G, number of communities k
Ensure: Communities with the final assignment of nodes
1: Randomly assign each node to one of the k communities
2: **repeat**
3: **for all** node v in G **do**
4: Calculate the distance between v and each centroid
5: Assign v to the community corresponding to the nearest centroid
6: **end for**
7: **for all** community **do**
8: Calculate the mean value of node features for nodes in the community
9: Update the centroid with the new mean value
10: **end for**
11: **until** convergence

4.2.3 Hierarchical clustering algorithm

Using hierarchical algorithms in community detection offers a powerful approach for uncovering hierarchical structures within complex networks. These algorithms aim to partition nodes into communities while also capturing the nested relationships between communities at different levels. Hierarchical community detection begins with each node as a separate community and gradually merges them based on their similarity or dissimilarity. This process continues until a desired stopping criterion is met, resulting in a hierarchy of nested communities. One popular hierarchical algorithm is the agglomerative approach, which starts by considering each node as a separate community and iteratively merges the most similar communities based on a predefined similarity measure. This hierarchical approach allows for the detection of communities at multiple scales, revealing both fine-grained and global community structures within the network. It provides valuable insights into the network's organization, facilitating a more comprehensive understanding of its complex dynamics and functional relationships. the agglomerative approach will be used further down the line to test and evaluate feature-based community detection. Here's a pseudo using the agglomerative approach to detect communities in a given graph network:

Algorithm 2 Agglomerative Algorithm for Community Detection

Require: Graph G
Ensure: Communities with the final assignment of nodes
1: Initialize each node as a separate community
2: **while** number of communities > 1 **do**
3: Calculate the pairwise similarity between communities
4: Find the two most similar communities based on the similarity measure
5: Merge the two communities into a single community
6: **end while**

Pairwise similarity approaches in community detection measure the similarity or dissimilarity between nodes or communities within a network. Common measures include Jaccard similarity, Cosine similarity, Adamic/Adar index, and Overlap coefficient. These measures quantify the affinity between nodes or communities, helping to identify and characterize communities in the network. As for the spectral algorithm, it has already been touched upon in the first approach.

Chapter 5

Results & Experimentation

In this chapter, we will be taking a look at the results of each approach as well as the resulting clusters.

5.1 Metrics evaluation

Method	**Clustering Algorithm**	**Silhouette Score**
Edge-based Approach	Spectral Clustering	-0.1887
Feature-based Approach	Spectral Clustering	-0.0343
	Hierarchical Clustering	0.6756
	K-Means	**0.6917**

Table 5.1: Clustering Evaluation Results

The clustering evaluation results presented in Table 5.1 show the performance of different methods using various clustering algorithms.

The edge-based approach using Spectral Clustering achieved a silhouette score of -0.1887, indicating a relatively poor clustering quality with low cohesion and high separation between clusters. This suggests that the edge connections alone may not provide sufficient information for effective community detection.

The feature-based approach, also using Spectral Clustering, obtained a slightly better silhouette score of -0.0343. While still not highly satisfactory, it suggests that incorporating node features improved the clustering results compared to the edge-based approach.

In contrast, the hierarchical clustering method demonstrated a significantly higher silhouette score of 0.6756. This indicates good clustering quality with well-separated and internally cohesive communities. The hierarchical approach

likely captured the underlying hierarchical structure of the communities within the network.

K-Means, another popular clustering algorithm, outperformed the hierarchical approach with a silhouette score of 0.6917. This indicates even better clustering quality, implying that K-Means effectively identified cohesive communities with distinct boundaries.

Overall, based on the silhouette scores, the hierarchical clustering and K-Means methods showcased superior performance in community detection compared to the edge-based and feature-based approaches using Spectral Clustering.

5.2 Resulting clusters

After the clustering process, we obtain three main themed clusters:

- The Music cluster

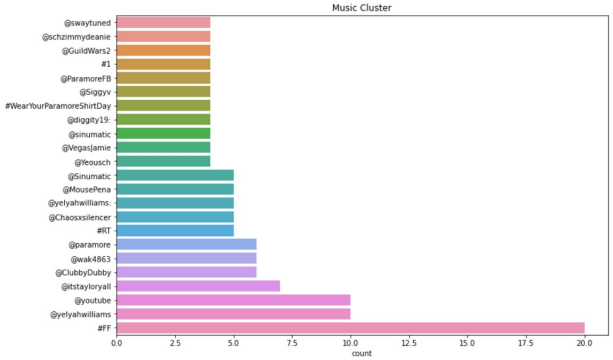

Figure 5.1: Music cluster

- The Gaming cluster

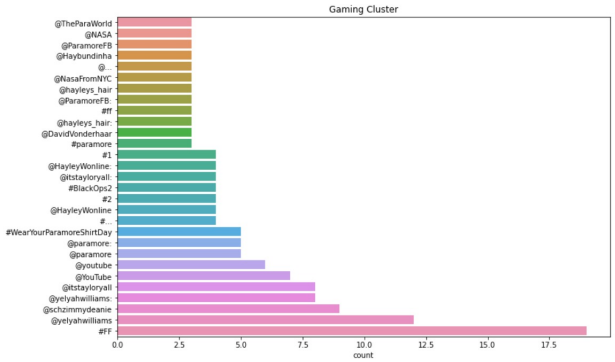

Figure 5.2: Gaming cluster

- The Social Media cluster

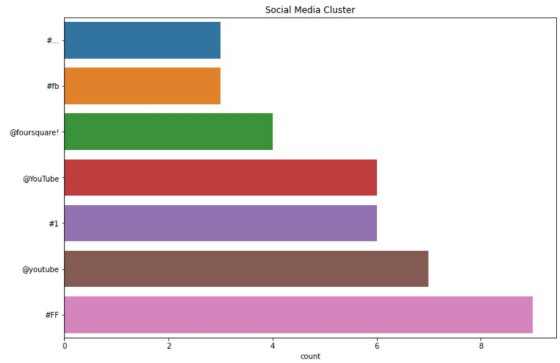

Figure 5.3: Media cluster

5.3 Conclusion

.

In conclusion, this report presents a comprehensive study on Twitter community detection using different approaches and techniques. The evaluation results revealed the performance of each method based on the Silhouette Score.

The Edge-based Approach utilizing Spectral Clustering obtained a Silhouette Score of -0.1887, indicating a lower quality of community detection. The Feature-based Approach with Spectral Clustering achieved a slightly better Silhouette Score of -0.0343. Hierarchical Clustering yielded a significantly higher Silhouette Score of 0.6756, indicating better community detection performance. However, it was K-Means that outperformed all other methods, demonstrating the highest Silhouette Score of 0.6917.

These findings suggest that K-Means is the most effective technique for community detection in the context of Twitter. It is recommended to utilize K-Means for identifying cohesive and interconnected groups within the Twitter network.

This research contributes to the understanding of Twitter community structures and provides valuable insights for researchers and practitioners in the field of online social network analysis. The results of this study can be applied in various domains, such as marketing, politics, and crisis management, to gain deeper insights into user behavior, information dissemination, and social influence within Twitter communities.

Bibliography

[1] Mr. Lazaar. *Unsupervised Learning Course*. ENSIAS, 2023.

[2] http://snap.stanford.edu/data/ego-Twitter.html