**T.R.**

**GEBZE TECHNICAL UNIVERSITY**

**FACULTY OF ENGINEERING**

**DEPARTMENT OF COMPUTER ENGINEERING**

# DETECTING ANTI-VACCINE GROUPS USING SOCIAL MEDIA ANALYSIS

ÖMER FARUK BITIKÇIOĞLU

SUPERVISOR
PROF. DR. YUSUF SINAN AKGÜL

GEBZE
2022

**T.R.**
**GEBZE TECHNICAL UNIVERSITY**
**FACULTY OF ENGINEERING**
**COMPUTER ENGINEERING DEPARTMENT**

# DETECTING ANTI-VACCINE GROUPS USING SOCIAL MEDIA ANALYSIS

**ÖMER FARUK BITIKÇIOĞLU**

SUPERVISOR
PROF. DR. YUSUF SINAN AKGÜL

**2022**
**GEBZE**

| | |
|---|---|
| GEBZE<br>TECHNICAL UNIVERSITY | GRADUATION PROJECT<br>JURY APPROVAL FORM |

This study has been accepted as an Undergraduate Graduation Project in the Department of Computer Engineering on 20/01/2022 by the following jury.

**JURY**

Member
(Supervisor) : Prof. Dr. Yusuf Sinan Akgül

Member : Dr. Yakup Genç

# ABSTRACT

Twitter is used by many people and different kinds of groups to share and influence people. Generating a social network from a Tweet data-set and identifying meaningful community structure from it can be beneficial for many cases. For instance, finding the terrorist groups and the accounts related is very crucial to provide them to influence and add a new member to their organizations.

For political organizations, this can be used too. Nowadays, politicians and parties are trying their best to use social media efficiently to attract more people and take their votes. Especially young people are dominant in social media and, politicians sharing their promises that the young generations can be interested.

For businesses, a product can be sold more easily if they know their customer network. They can give related advertisements to related groups and take their attention much more easily than putting huge billboards on a highway. This is the world we live in today.

Instances can not last easily by counting. But we are sure that all people in the world using the internet and especially social media leave much information behind to be analyzed. Companies investing more and more money to acquire more data and analyze it meaningfully.

In our project, our motivation is to find a network of anti-vaccine people. We know that they are very actively using Twitter and propagate many messages there. It is obvious to see that they all support each other. But we can not see the entire network and relations between these people. We can not estimate who is leading to this movement.

In this study, we used traditional community detection algorithms for contemporary problems. This is the era of the Covid19 pandemic and it is one of the most important problems at the beginning of the 2020s. We want to decipher the people propagating misleading information and have an insight into how this organization is communicating, growing, interacting, and supporting each other. Only the problem is new here. Rest is mostly implemented for many different problems before. But our studies showed us, traditional methods are still useful for current problems and this is why they are still popular techniques.

# ÖZET

Twitter, insanların paylaşım yapmak ve diğerlerini etkilemek için kullandığı ve ayrıca birçok farklı türde gruplar tarafından kullanılan bir platformdur. Bir Tweet veri setinden bir sosyal ağ oluşturmak ve bundan anlamlı topluluk yapısını belirlemek birçok durumda faydalı olabilir. Örneğin terör gruplarını ve ilgili hesapları bulmak, potansiyel adayları etkilemek ve örgütlerine yeni üye katmalarını önlemek için kullanılabilir.

Siyasi örgütler için de kullanılabilir. Günümüzde politikacılar ve partiler daha fazla insanı çekmek ve oylarını almak için sosyal medyayı verimli kullanmak için ellerinden geleni yapıyorlar. Özellikle gençlerin sosyal medyaya hakim olması sebebiyle siyasiler genç nesillerin ilgisini çekebilecek vaatler paylaşıyorlar.

İşletmeler için, müşteri ağlarının bilinmesi bir ürünü daha kolay sattırabilir. İlgili gruplara ilgili reklamlar verilebilir ve bir otoyola büyük reklam panoları koymak yerine çok daha kolay dikkat çekebilirler. Bugün içinde yaşadığımız dünya budur.

Örnekler sayarak kolayca bitmez. Ancak interneti ve özellikle sosyal medyayı kullanan tüm insanların geride analiz edilecek çok fazla bilgi bıraktığından eminiz. Şirketler, daha fazla veri elde etmek ve bunları anlamlı bir şekilde analiz etmek için giderek daha fazla para yatırıyor.

Projemizde, motivasyonumuz aşı karşıtı insanlardan oluşan bir ağ bulmaktır. Twitter'ı çok aktif kullandıklarını ve orada birçok mesaj yaydıklarını biliyoruz. Hepsinin birbirini desteklediği görülüyor. Ancak bu insanlar arasındaki tüm ağı ve ilişkileri göremiyoruz. Bu harekete kimin öncülük ettiğini tahmin edemeyiz.

Bu çalışmada, güncel problemler için geleneksel topluluk algılama algoritmalarını kullandık. Bu, Covid19 pandemisinin çağıdır ve 2020'lerin başındaki en önemli sorunlardan biridir. Yanıltıcı bilgi yayan insanları deşifre etmek ve bu organizasyonun nasıl iletişim kurduğuna, büyüdüğüne, etkileşim kurduğuna ve birbirini desteklediğine dair bir fikir sahibi olmak istiyoruz. Yalnızca problemimiz yeni, geri kalanı çoğunlukla daha önce birçok farklı problem için halihazırda uygulanmaktadır. Ancak çalışmalarımız bize gösterdi ki, geleneksel yöntemler mevcut problemler için hala faydalıdır ve bu yüzden hala popüler tekniklerdir.

# ACKNOWLEDGEMENT

I am thankful to my advisor Yusuf Sinan AKGÜL for keeping track of my work and giving advice during the semester.

Besides, I also want to thank Hüseyin KÜÇÜKALİ for sharing the data-set and giving advice.

**Ömer Faruk BİTİKÇİOĞLU**

# CONTENTS

# LIST OF FIGURES

# 1. INTRODUCTION

For the last two years, we have spent a lot of energy and time with the coronavirus and the disease it causes. It spread very quickly, and we have rearranged our lives under the pandemic conditions. We listened to the words of scientists and authorized persons and tried to comply with all necessary precautions. In this way, we reduced the spread of the virus that can cause damage to both health and the economy.

During this period, some people opposed the security measures put forward, found it unnecessary, and thought it absurd. Some conspiracy theories about vaccination and the precautions, in general, came out. The anti-science people gathered and spread their ideas together. No matter how hard scientists tried to convince people with scientific evidence, these groups were not satisfied. On the contrary, they continued to produce the antithesis to everything said. While scientists are trying to find a way out of this pandemic and developing medicine and vaccines, the anti-science people put themselves and the societies in danger.

Since social media acts as a communication bridge between such groupings, it provides an efficient environment for detecting communities. Communities in social media consist of user accounts that communicate and support each other and have the same interests. By obtaining and analyzing such information, surprisingly accurate results can be found and presented to the relevant researchers. Twitter is advantageous social media in this context and is also helping to identify relationships because being liked and retweeted can mean approval. Also, people belonging to this same group or opinion are likely to follow each other.

We created a social network graph with obtained relationships. In this graph, it can be seen in detail which accounts are connected and intertwined with which accounts. However, since visual detection with the human eye would not be sufficient for scientific correctness, to detect communities, algorithms are used. By using clustering algorithms, the social network graph was divided into two groups as anti-vaccine and non-anti-vaccine, albeit superficially.

# 2. LITERATURE REVIEW

In literature, there are many papers that focus on community detection in social networks and their applications. These papers use a different kinds of approaches, principles, network types, network natures, network directions, network sizes, and datasets. [1] proposed a multilevel clustering technique and used the topology of eccentric connections to detect socially unrelated accounts. [2] proposed a three-tier framework based on the node influence k-nearest neighbors (NI-KNN) algorithm for detecting the community.

[3] proposed to use the modularity maximization technique to cluster the graphs of all the views of the data. They modified the optimization function proposed in a paper in 2010 for dealing with multi-layer networks.

In [4], a graph-based approach has been constructed, and then community detection methodology is applied on the similarity graph to cluster similar tweets. They also used different graph centrality measures such as degree centrality, closeness centrality, betweenness centrality, eigenvector centrality, and page rank. They claim that their methodology performs better than some basic methods.

[5] mentions traditional methods used for community detection in graphs besides divisive algorithms, modularity-based methods, spectral algorithms, dynamic algorithms, statistical inference, and alternative methods. They also mentioned methods to find overlapping communities which is an important problem in community detection. After that, they summarized the multiresolution methods for cluster hierarchy and methods for detecting dynamic communities. It was a solid summarization of the most recent methods. Also [6] reviews the literature on community detection and published on the 10th International Conference on Ambient Systems, Networks, and Technologies which is held April 29 – May 2, 2019. [7] provides an overview of community detection problems from the perspective of bio-inspired computation.

[8] is an attempted contemporary survey of the methods of community detection and its application in the various domains of real life. It serves as an up-to-date report on the evolution of community detection and its potential applications in various domains from real-world networks.

While classical approaches for community detection usually deal only with the structure of the network and ignore the node attributes, [9] is proposed the node attributed graph which developed by using both the structure and the attributes of the network such as age, gender, interests, etc. to yield more informative and qualitative community detection results.

[10] proposed a bottom-up approach to the problem of community detection.

Network analysis and community detection are very time-consuming and require lots of memory for medium to large-scale networks. This paper tries to solve this issue and starts with the small partition of the network and constructs them separately then merges them into the bigger networks. This way community detection problem has a smaller space and time complexity.

[11] proposes an algorithm called Label Propagation Algorithm (LPA) for the community detection problem. LPA has caught attention for its advantages. Abundant literature finds LPA efficient, scalable, conceptually clear, and simple. This paper includes LPA-related proposals, including enhancements and extensions. Experiments indicate that most enhancements are beneficial in specific scenarios.

[12] proposed an approach that tracks two aspects of community evolution in retweet networks: flow low of the members in, out, and between the communities, and their influence. For community detection, they propose a two-stage approach. In the first stage, they apply an enhanced Louvain algorithm, called Ensemble Louvain, to find stable communities. In the second stage, they form influence links between these communities and identify linked super-communities.

[13] claimed that classic methods of community detection, such as spectral clustering and statistical inference are falling due to the increasing popularity of deep learning techniques. The deep learning technique has an increased capacity to handle high-dimensional graph data with strong performance. Thus, they wanted to write a survey of the current progress of deep learning and structured it into three broad research streams: deep neural networks, deep graph embedding, and graph neural networks.

[14] examines the social network analysis under the motivation of detecting terrorist groups in social media. Terrorists use social media to spread their message and recruit new members. To do this, leaders do not interact with possible recruits. They use a friend of a friend technique to influence people.

In our study, we were also motivated to find leaders of the misleading information of coronavirus vaccinations and pandemics in general. Using the friend of a friend technique could be useful for us too. Anti-vaxxers are a very dense group and we can not estimate the real leaders of the thoughts easily. So, using such algorithms can be beneficial to gain insights.

Using hashtags as an indicator of topic classification in Twitter is known by everyone but in [15] they proposed a new approach to process hashtags to find common topics and cluster them clearly. It is not social network analysis, but more like a topic classification technique.

It is hard to find weak communities that most members of it are also members of the dominant communities. In this paper [16], they proposed a new approach called HICODE (Hidden Community Detection) to both detect weak communities and enhance the technique to find dominant communities too. They claim that experiment results show that this technique outperforms most state-of-art community detection techniques and has significant importance in the field.

[17] paper analyses the spectral analysis algorithm by stressing the misused situations and insufficient results. They basically implemented the algorithm with a few lines of code explained in which conditions spectral clustering works well and analyzed experiments. To get a better understanding of the topic this paper should be the one that needs to be covered.

[18] introduced the Louvain algorithm and advance a partitioned Louvain algorithm as its improved variant. Also, they use clustering techniques for trend analysis. They use nonnegative matrix factorization (NMF), which is a convenient method to intuitively interpret and extract issues on various time scales.

[19] see the community detection problem as an optimization problem. They designed Whale Optimization Algorithm (WOA), a recently proposed meta-heuristic algorithm, to mimic the hunting behavior of humpback whales and deal with the optimization problem. In the paper, a new community detection algorithm, Whale Optimization-based Community Detection Algorithm (WOCDA), is proposed to discover communities in networks. In WOCDA, a new initialization strategy and three operations, shrinking encircling, spiral updating, and random searching, are designed to mimic the hunting behavior of humpback whales.

In the study of this project, to detect anti-vaccine groups using social media analysis we used a classical approach. We know that Twitter hashtags are tweet topic classification indicators. To find anti-vaccine tweets we needed to find anti-vaccine hashtags and we do so. Tweets are collected and analyzed. Retweet and mention are the most notable edge indicators. So, we used them to generate edges of the network. Nodes represent the accounts tweeted under these hashtags. Node attributes are not used.

The generated network is translated into a matrix representation (Laplacian Matrix). The Laplacian matrix is calculated using the formula L = D – A. D represents the Degree Matrix of the network. To calculate the degree matrix, we iterated over all the nodes and count the edges. A represents Adjacency Matrix that includes 1 or 0's according to the information that if two edges are connected or not. If two edges are connected then the value of the corresponding Adjacency Matrix input is 1, if not it is 0.

Laplacian Matrix used the classical approach of network analysis which is Spectral Clustering. To use spectral clustering the eigenvalues and eigenvector of the Laplacian matrix are calculated, and the second smallest eigenvalue and corresponding eigenvector take as an indicator of grouping. We have no difference from the classical approach.

# 3. PROJECT DEFINITION

In this project, our purpose is to find clusters from the given Twitter data set. This data set includes all the tweets sent during the anti-vaccine meeting in Istanbul. People sending tweets under this hashtag belong to one of these groups: Anti-vaccine or not anti-vaccine. We want to find these groups.

To find these groups, we had to examine the tweet data which includes retweet, mention, or like data. If two of the people are mentioning, retweeting, or liking their tweets it has a high chance to observe that these two people are on the same side.

From this observation, we converted the problem of finding anti-vaccine groups to finding a network of people 3.1 under this anti-vaccine hashtag. After that using that network we feed a spectral clustering algorithm to find anti-vaccine and not anti-vaccine communities.

Community detection is a problem of complex network analysis and aims to reveal an understanding of the network by extracting meaningful sub-graphs. In today's world, it gained massive importance and became a useful tool due to the increasing use of social media.

Thanks to a variety of social media applications, a huge amount of data is collected each day from the users and their interactions. Users are interacting with each other by sending posts, tweets, or other kinds of special content designed for the special application. Users of the applications give feedback to their content by using approval features like a retweet, like, or share.

By inspecting the user interactions, a social network can be generated from social media. In Twitter, if a user retweets a tweet of another user, it means that he/she approved, liked, wanted to share, or has an interest in the content of the tweet. If he/she likes the tweet, it is obvious that he/she liked the content or wanted to propagate over the main pages of their followers since the liked tweets do so. If the content is retweeted but added a comment on top of it, it is not obvious that two people have the same opinion, but they have a topic that both interested. Mention data have the same logic too. So, in the enlightenment of this information, a graph can be generated from tweets by looking at retweet, like, and mention data.

In addition to these data, Twitter has a popular feature called a hashtag. A hashtag can be thought of as a topic or subreddit feature in Reddit but the more temporary version. The topics people talk about vary in time, so do hashtags. The meaning of it is not that hashtags disappear after a certain period. They all remain still but the popularity of the hashtag changes and the tweet density over time on the hashtag increase or decreases. The other related feature of a hashtag is trends. If a hashtag has a high rate of interactions in a period, it deserves to be written in the trend topics.

In the perspective of our main goal, the problem is to analyze tweets by using the attributes mentioned above, generate a social network of Twitter users tweeted under a certain type of hashtags, and cluster them into two or more clusters to reveal underlying subdivisions and understand the organization of the desired group.

In this project, we want to analyze the group of people who decline modern science, modern medicine, and specific vaccination. In the past 2 years, these people are against Covid19 vaccines at all, wearing a mask, social distance, etc. This group also mostly propagates misleading information. The actions of the group are mostly unhealthy for society and must not be welcomed.

We have a collected dataset of 360 thousand tweets that tweeted under the #HerYerMaltepeHerYerDirenis, #HerYerMaltepeHerYerDireniş, #BuyukUyanisMitingi, #MaltepeMitingi at the period of the event occurred. This collected set of tweets must be cleaned, analyzed, and summarized for insights and observations for the sake of future projects and research.

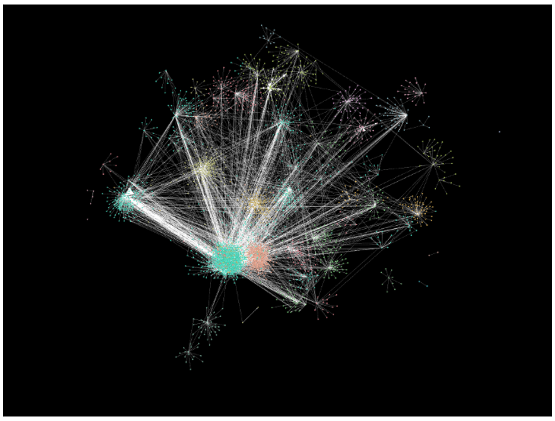


Figure 3.1: A network graph instance

The figure above shows a network graph sample. It is just for illustration and means nothing for our project. If the reader wants to imagine what a network graph can look like, this figure can help. The nodes are connected with edges and separated with colors. Colors represent the clusters the nodes belong. The stronger the edges, the stronger the connection both nodes have. A stronger connection indicates that both vertices have more properties in common. The distance can be a measure too. But the details of the network are not beyond our imagination.

# 4. REQUIREMENTS AND SYSTEM DESIGN

Tweet data must be collected from aimed hashtags. For this purpose, Twitter API, if you have a Twitter developer account or third-party applications can be used.

Social networks must be generated so that the reader can see the connections and relations between accounts. To generate such a network the edges and nodes of the network must be initialized. The network must represent Twitter accounts as nodes and relationship connections as edges of the graph. The nodes must be closed if the two nodes have more in common. There is no need to use node or edge attributes in this stage. Maybe later these can be evaluated too.

The generated network must be fed one of the clustering algorithms to find the hidden information of groupings. Thanks to these algorithms we can find groups inside the generated social network. In terms of social networks, clustering is important to understand the relations between nodes of the network. To cluster the network, we can give node attributes, edge attributes, edge count of the node, coordinates of the node (since closeness is another indicator), and most importantly connections of the nodes in the network to the clustering algorithms. 4.1

R programming language and some of its libraries have been used for this project. Libraries are ggplot2, readr, dplyr, reshape2, tidyr, formattable, RColorBrewer, lubridate, networkD3, plotly, cluster, viridis, stringr, listviewer, visNetwork.

According to Wikipedia "R is a programming language for statistical computing and graphics supported by the R Core Team and the R Foundation for Statistical Computing. Created by statisticians Ross Ihaka and Robert Gentleman, R is used among data miners and statisticians for data analysis and developing statistical software."

The other choice as a programming environment was Python. Since the starting point of this analysis, the project was built in R, it continued as it is. There is no specific reason to choose R over Python. They are both useful for the same tasks.

In this project, the data is collected from Twitter via a Twitter API developer account and a third-party application. The data had to be sterilized to work with JSON data and to analyze efficiently in terms of memory because the tweet data contain unnecessary details. After sterilization, this data is used to generate a community network and the network feeds the spectral clustering algorithm to find anti-vaccine and not anti-vaccine communities.

There are plenty of clustering algorithms. Which clustering algorithm is going to be used for this project? This is another point to be discussed. Pros and cons must be evaluated for the specific conditions. In this project, we planned to choose the Spectral Clustering algorithm. We do not have a fixed shape graph. We have a graph whose

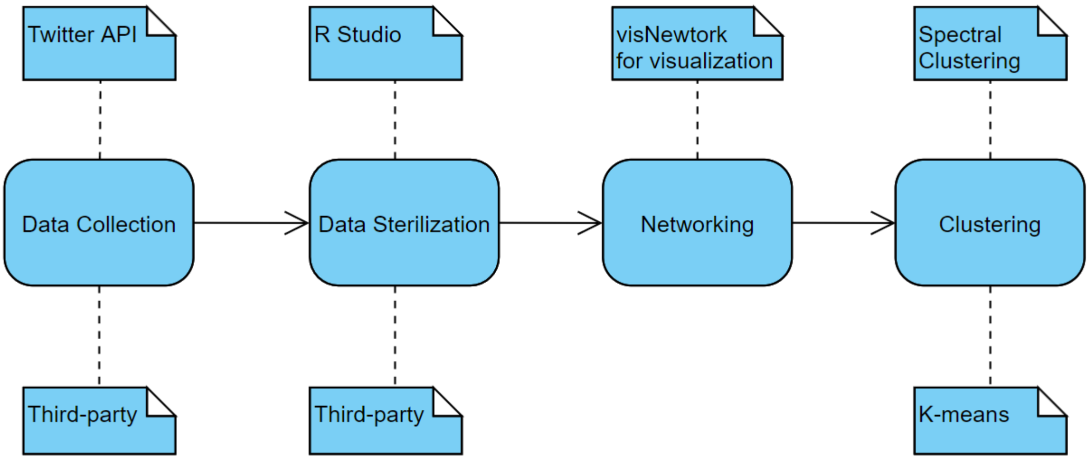


Figure 4.1: System Design

nodes can be far away but belong to the same cluster. So Spectral Clustering is useful for our project.

## 4.1. Community detection theory

In the study of social networks, community detection is simply the process of revealing information. Networks give us the data to extract and understand the relationship between the nodes and the communities inside.

*" The modern science of networks has made significant advancements in the modeling of complex real-world systems. One of the most important features in these networks is the existence of community structure. In recent years, many community detection algorithms have been proposed to unveil the structural properties and dynamic behaviors of networks "* [8]

In this project, we attempt to find the social network of the anti-vaccine gathering and the clusters of anti-vaccine and not anti-vaccine people.

## 4.2. Community detection algorithms

In the process of network analysis and clustering, there are many community detection algorithms implemented for different scenarios. We can group them into two as Overlapping Community Detection Algorithms and Disjoint Community Detection Algorithms. 4.2

The traditional approach to community detection is to use spectral clustering. The spectral clustering algorithm is also using another classical and simple clustering algorithm called K-means clustering. K-means is one of the most popular unsupervised

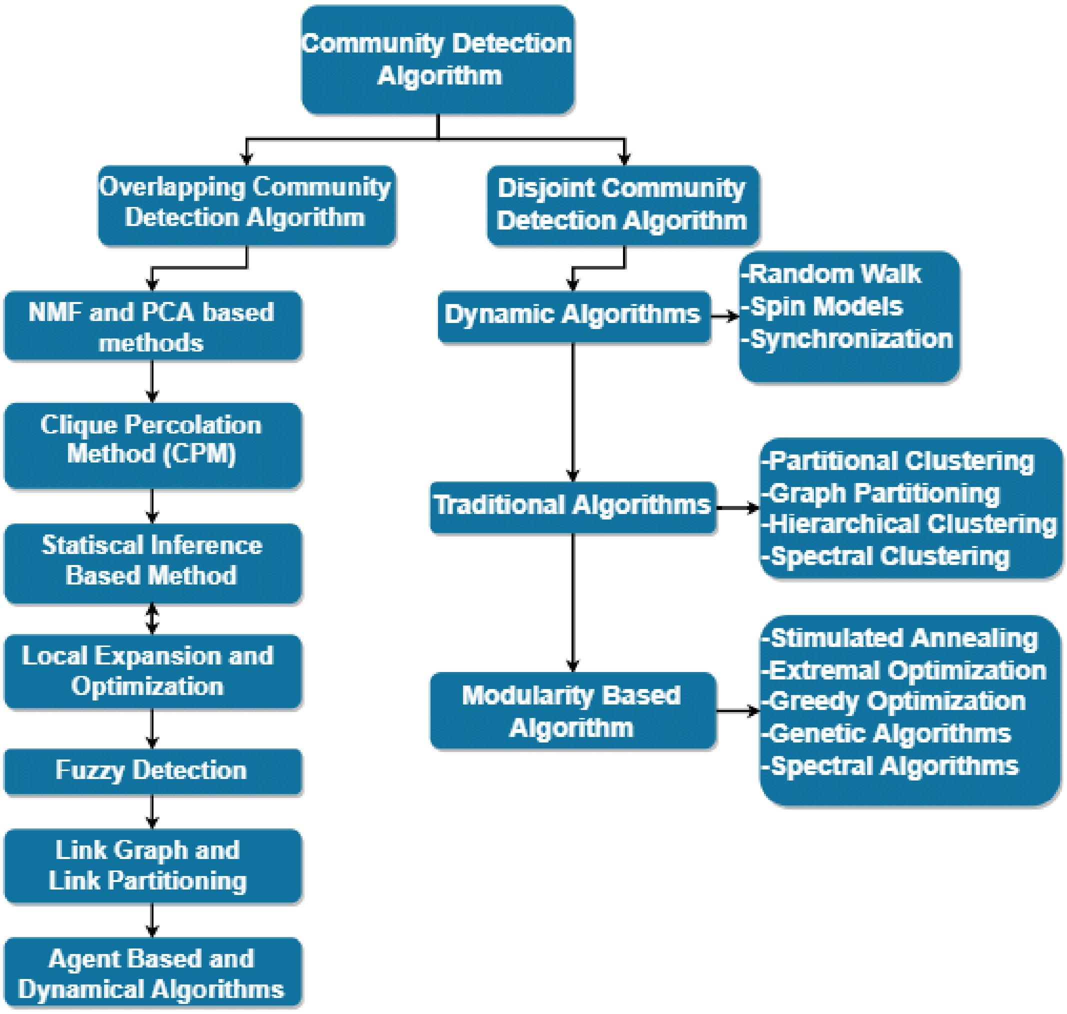

Community Detection Algorithm
Overlapping Community Detection Algorithm
Disjoint Community Detection Algorithm
NMF and PCA based methods
Clique Percolation Method (CPM)
Statiscal Inference Based Method
Local Expansion and Optimization
Fuzzy Detection
Link Graph and Link Partitioning
Agent Based and Dynamical Algorithms
Dynamic Algorithms
-Random Walk
-Spin Models
-Synchronization
Traditional Algorithms
-Partitional Clustering
-Graph Partitioning
-Hierarchical Clustering
-Spectral Clustering
Modularity Based Algorithm
-Stimulated Annealing
-Extremal Optimization
-Greedy Optimization
-Genetic Algorithms
-Spectral Algorithms

Figure 4.2: Taxonomy of Community Detection Algorithms

[20]

machine learning algorithms.

Spectral Clustering is a more advanced and again unsupervised machine learning technique to differentiate clusters and often more efficiently than the other algorithms. It is very simple to implement and can be solved by linear algebra methods efficiently. It doesn't only calculate the euclidean distance and position of the node as k-means does, but also takes into consideration the affinity. It is useful for detecting clusters in a network that has a complicated shape.

The Algorithm 1 shows the pseudo-code of the k-means clustering algorithm.

The Algorithm 2 shows the pseudo-code of the spectral clustering algorithm. [17]

**Algorithm 1** K-means clustering

```
1. Initialize k means with random values
2. For a given number of iterations:
3. Iterate through items:
4. Find the mean closest to the item
5. Assign item to mean
6. Update mean
```

**Algorithm 2** Spectral clustering

Given a set of points $S = \{s_1, \ldots, s_n\}$ in $\mathbb{R}$ we want to cluster into k subsets

```
1. Form the affinity matrix $A \in \mathbb{R}^{n \times n}$ defined by $A_{ij} = exp(-||s_i - s_j||^2/2\sigma^2)$ if $i \neq j$, and $A_{ij} = O$.
2. Define D to be the diagonal matrix whose $(i, i)$-element is the sum of A's $i$-th row, and construct the matrix $L = D^{-1/2}$ AD$^{-1/2}$.
3. Find $x_1, x_2, \ldots, x_k$, the k largest eigenvectors of $L$ (chosen to be orthogonal to each other in the case of repeated eigenvalues), and form the matrix $X = [x_1, x_2, \ldots, x_k] \in \mathbb{R}^{n \times k}$ by stacking the eigenvectors in columns.
4. Form the matrix $Y$ from $X$ by renormalizing each of $X$'s rows to have unit length (i.e. $Y_{ij} = X_{ij}/(\sum_j X_{ij}^2)^{1/2}$).
5. Treating each row of $Y$ as a point in $\mathbb{R}^k$, cluster them into k clusters via K-means or any other algorithm (that attempts to minimize distortion).
6. Finally, assign the original point $s_i$ to cluster $j$ if and only if row $i$ of the matrix $Y$ was assigned to cluster $j$.
```

In our project, Spectral Clustering has been used to cluster social networks of Twitter accounts tweeted under the anti-vaccine hashtag. It helped to find out anti-vaccine accounts. Since the graph is crowded and has not had a fixed shape it was more suitable to use this algorithm.

We have a set of nodes and their edges. Edge length depends on the node relation. It is closer if the two nodes are more related. In the next chapter, more information on the implementation process can be found.

# 5. IMPLEMENTATION DETAILS

The project is implemented in R Studio with R language. We first loaded the needed libraries with library(library_name) syntax. Libraries are ggplot2, readr, dplyr, reshape2, tidyr, formattable, RColorBrewer, lubri-date, networkD3, plotly, cluster, viridis, stringr, listviewer, visNetwork.

The data-set shared by Hüseyin Küçükali was in JSON format. To handle it in R language we first collect it with the rjson::fromJSON method and unnest intertwined data blocks to use easily.

After that, we need to know how many tweets are original tweets and how many are retweets. We calculated it using some string manipulation and added a new column indicating each tweet is retweeted or not.

We selected the user accounts which has the most followers and most tweets. This is revealing the most influential Twitter accounts in the communities. After that we clustered these popular accounts into four, using k-means clustering. These clusters are about the influence factor. Most of the results are not related to anti-vaccine people. We understand that our data-set contains unrelated accounts' tweets too. They are mostly Twitter accounts of huge media companies. Only one of the results is a popular anti-vaccine individual.

We wanted to see how many mentions do these accounts get. If an account gets many mentions it means that it has an influential potential too. And these accounts can lead to the main accounts of the communities. It is beneficial for our study since we want to find the source of the misleading information.

By using mention and retweet data we generated a social network of accounts that tweeted under the hashtag of anti-vaccine gathering in Istanbul, September 2021. This network has hidden information about anti-vaccine communities. People who belong to the anti-vaccine group are tweeting, retweeting, liking, sharing, supporting, following each other under these hashtags. If we analyze the tweets in-depth, we can find some useful information to extract clusters of anti-vaccine community and their leaders.

In our implementation of Spectral Clustering, we used coordinates of the Twitter accounts in the social network as input to find similarity matrix W. And then we computed the graph laplacianLusing this similarity matrix W. We calculated the eigenvectors of the Laplacian Matrix and clustered these vectors into two using K-means clustering. Labeled each cluster with a different color "orange" and "green".

# 6. TESTS AND RESULTS

We can see that most influential accounts around the community are news channels. 6.1 They have lots of followers and tweets. The tweets they sent reaches thousands of people. But from that information we can not estimate anti-vaccine community.

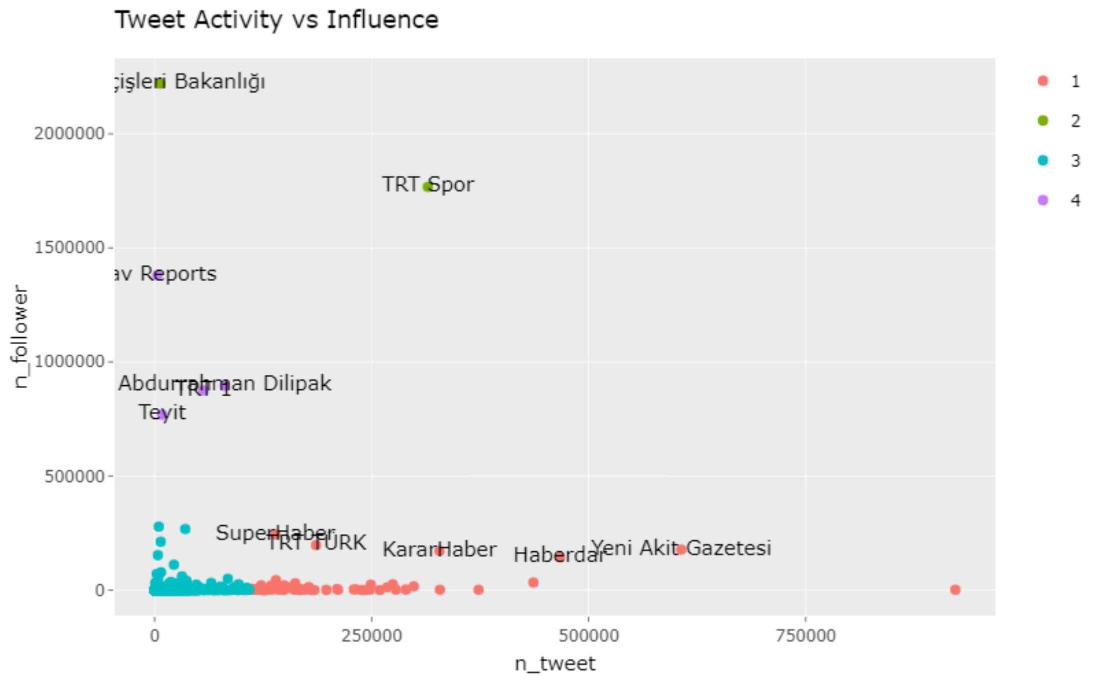
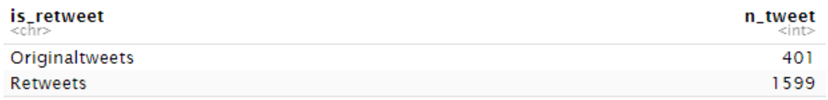


Figure 6.1: Tweet Activity vs Influence

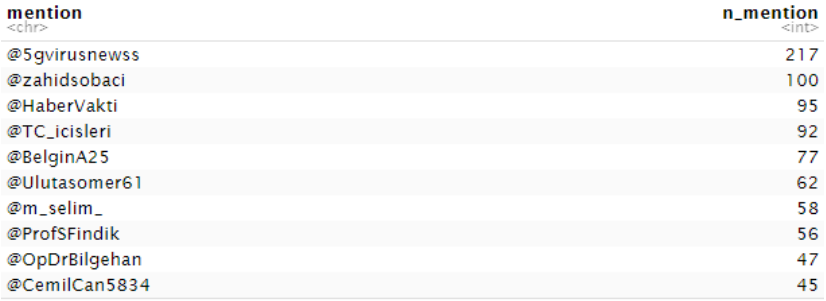
We calculated the count of retweets 6.2 in this data-set and 79.95% are retweets. Only the remaining 20.05% are pure, original tweets. It is not scientific to say like that but we can guess that not many ideas are invented by the anti-vaccine community, but these people spread the same ideas from a few origins and share them over and over again.



| is_retweet<br>&lt;chr&gt; | n_tweet<br>&lt;int&gt; |
|---|---|
| Originaltweets | 401 |
| Retweets | 1599 |

Figure 6.2: Retweet Count

Having mentions is another critique metric that shows the connection between the community network. Also having lots of mentions has a meaning of popularity. In addition to the follower count, we can consider the mention count while having assumptions about popular accounts in a social media network. 6.3



| mention<br><chr> | n_mention<br><int> |
|---|---|
| @5gvirusnewss | 217 |
| @zahidsobaci | 100 |
| @HaberVakti | 95 |
| @TC_icisleri | 92 |
| @BelginA25 | 77 |
| @Ulutasomer61 | 62 |
| @m_selim_ | 58 |
| @ProfSFindik | 56 |
| @OpDrBilgehan | 47 |
| @CemilCan5834 | 45 |

Figure 6.3: Mention Count

Using the mentioned data we can generate the network 6.4. The nodes indicate the Twitter accounts that tweeted under anti-vaccine event hashtags.

Closer the nodes, more connected to the cluster. The outsiders of the cluster are probably a member of the opposite cluster. We can not estimate it only with the human eye but we will calculate it more precisely by using Spectral Analysis of the coordinates and connections of the nodes.

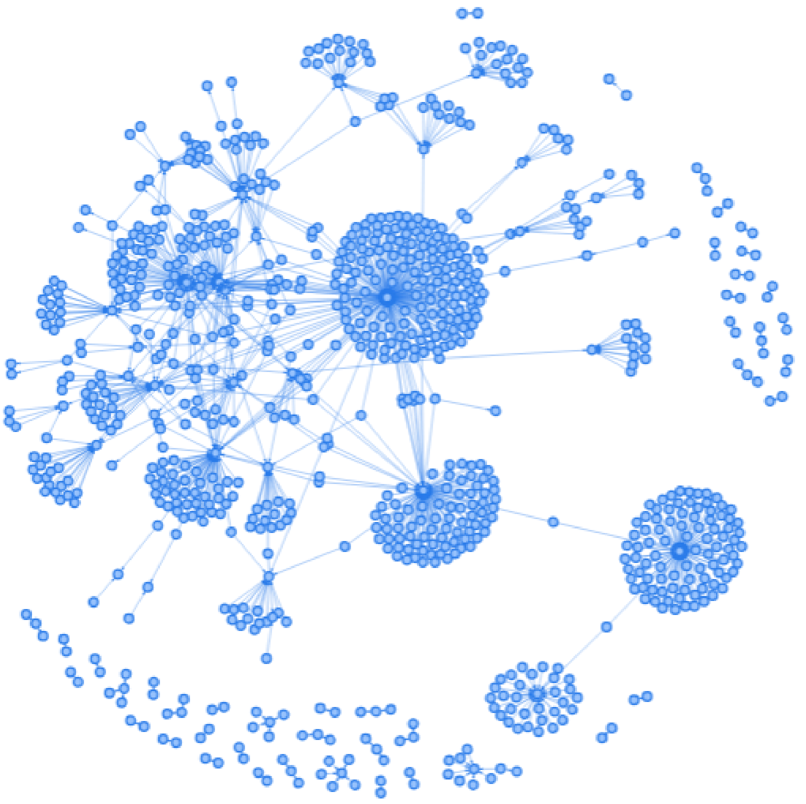


Figure 6.4: Social Network of Twitter Accounts Tweeted Under Anti-Vaccine Gathering

After analyzing the network with spectral clustering, we can see that the network is broken into two groups. The green ones are anti-vaccine people and the orange ones are not anti-vaccine people.

It is not 100% precise if you think of the nodes gathered around popular nodes. They are probably not a member of a cluster of the popular node. They just mentioned or retweeted the popular account's tweet and counted as one of the members.

Also, some of the little nodes around the graph are not fully correct because of the lack of connection information. If the data-set gets bigger, the results may produce more healthy results.

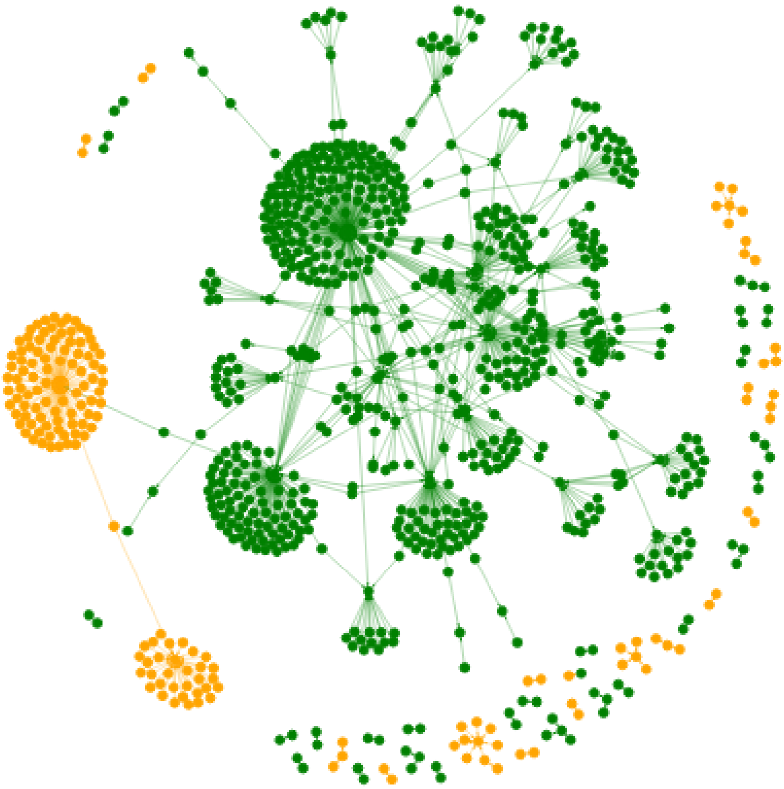
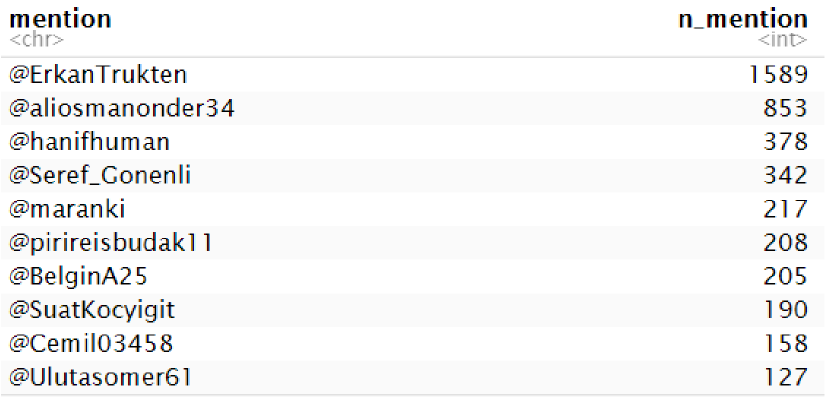


Figure 6.5: Clustered Network

We used the previous results as input for the next tests. This time we only collected the tweets of the most popular anti-vaccine accounts in terms of the previous results. We know that they are anti-vaccine since we checked manually by looking at their Twitter account and tweets. We found new accounts mentioned by these accounts. 6.6



| mention | n_mention |
|---|---|
| <chr> | <int> |
| @ErkanTrukten | 1589 |
| @aliosmanonder34 | 853 |
| @hanifhuman | 378 |
| @Seref_Gonenli | 342 |
| @maranki | 217 |
| @pirireisbudak11 | 208 |
| @BelginA25 | 205 |
| @SuatKocyigit | 190 |
| @Cemil03458 | 158 |
| @Ulutasomer61 | 127 |

Figure 6.6: New Results: Most Mentioned Accounts by Anti-Vaccine People

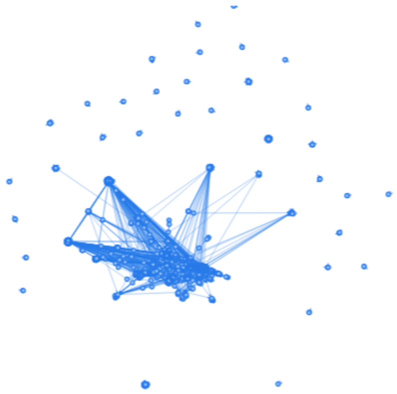


Figure 6.7: New Results: Social Network of Anti-Vaccine Accounts

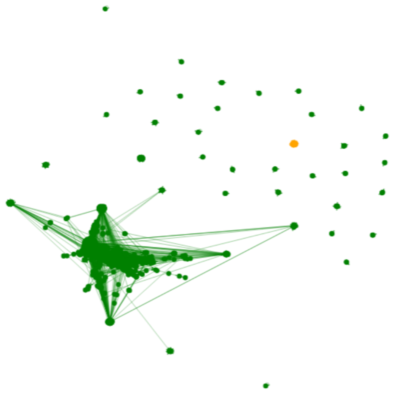


Figure 6.8: New Results: Clustered Network of Anti-Vaxers

# 7. SUCCESS CRITERIA

## 7.1. Analyzing data will take less than 1 second

At the start of the project, I was not aware of how much time would it take to analyze tons of data. We have 2000 data to be analyzed. It produces a network and runs a spectral clustering algorithm. It takes much more than 1 second.

## 7.2. At least 2000 user data will be used

This part is satisfied. But as we have more than 300 thousand data, it is not nice to analyze only 2000 data. It would have been better. But we had limited memory space and did not have powerful computers to work with. To solve this, the data-set can be cleaned more and the methods can be improved.

## 7.3. At least 4 different periods will be examined

Only one time period is used. At the start of the project, it was decided to pull data from different periods. (Tweets from different activities, meetings, gatherings). But with a limited time, this was only succeeded with only one event on the same dates.

This causes the results to have a lack of analysis of changing opinions of people on different events and periods. This study demonstrates clusters for only one period of time. We are not sure if the people changed their minds at different times and events.

## 7.4. Results are compatible with standard data and algorithm

Getting the network and feeding it into the spectral clustering algorithm is tested with standard data. So, we can conclude that it works correctly with our data too.

# 8. CONCLUSIONS

When we think of machine learning, there are no 100% correct results. The results given before are correct enough to see the connection and grouping between anti-vaccine people on Twitter.

When we collect the tweets of only anti-vaccine accounts, knowing that they are anti-vaccine accounts from the previous analysis, it is obvious that all the accounts are intertwined with each other and communicating closely. We can say that there is a strong connection and communication between these accounts.

But there are problems with our results too. The analyzed data is not enough. In different periods and with different techniques, data should be collected and analyzed again. Results would be improved with the needed time and effort. This study is analyzing only 2000, 5000, and at most 10000 data from more than 360 thousand data.

We didn't include all the data in our project since the lack of memory and poor performance of the implementation. Implementation can be improved in terms of space complexity. Also, the data-set can be divided into batches containing 5000 tweets for instance. In this way, all the data can be analyzed part by part and see the results of different parts of the data-set.

The analysis can be done for only a specific day (the day the meetings are held). So more relevant results can be gathered.

The implementation of this project only contains mention and retweet data. Also, the question of which account is following which account is another important thing to generate networks over social media.

Following, retweeting, and mentioning data can be analyzed and generate networks separately and differences can be interpreted.

For the follower network, the retweet information can be analyzed over time and we can try to see the difference dynamically.

As mentioned before in the report, node attributes are not used. We can assign follower or tweet count information to the nodes. Also, edge attributes can be used too. We can assign how many mentions, or retweets they made to the connections and make them stronger.

Spectral clustering is a beneficial technique for community detection over social networks. Because social networks are complex shaped nodes and edges can have different kinds of attributes. All things considered, this is the most suitable and most traditional method.

# BIBLIOGRAPHY

[1] I. Inuwa-Dutse, M. Liptrott, and I. Korkontzelos, "A multilevel clustering technique for community detection," *Neurocomputing*, vol. 441, pp. 64–78, 2021.

[2] P. Meena, M. Pawar, and A. Pandey, "A survey on community detection algorithm and its applications," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 6, pp. 4807–4815, 2021.

[3] I. J. Cruickshank and K. M. Carley, "Characterizing communities of hashtag usage on twitter during the 2020 covid-19 pandemic by multi-view clustering," *Applied Network Science*, vol. 5, no. 1, pp. 1–40, 2020.

[4] S. Dutta, A. K. Das, A. Bhattacharya, *et al.*, "Community detection based tweet summarization," in *Emerging Technologies in Data Mining and Information Security*, Springer, 2019, pp. 797–808.

[5] S. Fortunato, "Community detection in graphs," *Physics reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[6] E.-M. Mohamed, T. Agouti, A. Tikniouine, and M. El Adnani, "A comprehensive literature review on community detection: Approaches and applications," *Procedia Computer Science*, vol. 151, pp. 295–302, 2019.

[7] E. Osaba, J. Del Ser, D. Camacho, M. N. Bilbao, and X.-S. Yang, "Community detection in networks using bio-inspired optimization: Latest developments, new results and perspectives with a selection of recent meta-heuristics," *Applied Soft Computing*, vol. 87, p. 106 010, 2020.

[8] M. A. Javed, M. S. Younis, S. Latif, J. Qadir, and A. Baig, "Community detection in networks: A multidisciplinary review," *Journal of Network and Computer Applications*, vol. 108, pp. 87–111, 2018.

[9] P. Chunaev, "Community detection in node-attributed social networks: A survey," *Computer Science Review*, vol. 37, p. 100 286, 2020.

[10] M. Arab and M. Afsharchi, "Community detection in social networks using hybrid merging of sub-communities," *Journal of network and computer applications*, vol. 40, pp. 73–84, 2014.

[11] S. E. Garza and S. E. Schaeffer, "Community detection with the label propagation algorithm: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 534, p. 122 058, 2019.

[12] B. Evkoski, I. Mozetic, N. Ljubesic, and P. K. Novak, "Community evolution in retweet networks," *arXiv preprint arXiv:2105.06214*, 2021.

[13] F. Liu, S. Xue, J. Wu, *et al.*, "Deep learning for community detection: Progress, challenges and opportunities," *arXiv preprint arXiv:2005.08225*, 2020.

[14] T. Waskiewicz, "Friend of a friend influence in terrorist social networks," in *Proceedings on the international conference on artificial intelligence (ICAI)*, The Steering Committee of The World Congress in Computer Science, Computer . . ., 2012, p. 1.

[15] D. Gromann and T. Declerck, "Hashtag processing for enhanced clustering of tweets.," in *RANLP*, 2017, pp. 277–283.

[16] K. He, Y. Li, S. Soundarajan, and J. E. Hopcroft, "Hidden community detection in social networks," *Information Sciences*, vol. 425, pp. 92–106, 2018.

[17] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, 2002, pp. 849–856.

[18] Y.-H. Kim, S. Seo, Y.-H. Ha, S. Lim, and Y. Yoon, "Two applications of clustering techniques to twitter: Community detection and issue extraction," *Discrete dynamics in nature and society*, vol. 2013, 2013.

[19] Y. Zhang, Y. Liu, J. Li, *et al.*, "Wocda: A whale optimization based community detection algorithm," *Physica A: Statistical Mechanics and its Applications*, vol. 539, p. 122 937, 2020.

[20] T. Sangkaran, A. Abdullah, and N. Jhanjhi, "Criminal community detection based on isomorphic subgraph analytics," *Open Computer Science*, vol. 10, no. 1, pp. 164–174, 2020.