Community Detection in Facebook Using Visual Approach and Clustering

Bara'a Talaat Ahmed Seyala¹, Mafaz Alanezi²

¹, ² Department of Computer Science, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq.

The corresponding author’s e-mail: mafazmhalanezi@uomosul.edu.iq

ABSTRACT. The previous decade has seen the growth of interest with networks and participatory social media that have brought users jointly in many originate ways. Many users play, categorize, work and socialize online, showing new forms of cooperation, communication, and cleverness that were hard to imagine just a while ago. Social media refers to the interaction between people who create, share information and ideas in communities and virtual networks. Social media also helps reform business models, influence views and sentiments, and opens many possibilities for studying human interaction and mass conduct on an unprecedented level.

This research employs visual representation of data and cluster algorithms for discovering patterns in the Facebook network to learn some of the behaviors practiced by community members. The results can be used to find out users directions to suggest appropriate advertisements for it, and cluster algorithms can be used to collect suspicious and inappropriate communication pages to take the necessary measures to prevent them from appearing to sensitive groups in the community, and the results can also be used to direct Facebook users, especially young groups, to organize their times and control the times they spend on social media.

Keywords: Facebook, visual representation, K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

1. INTRODUCTION
Sites of social networking on the Internet present great possible for connection and interaction between users who are in many places dispersed, connecting them in different ways. It also facilitates interaction and exchange of input with other users including relatives, colleagues, family, friends, fans, and others. In addition to facilitating connections, social networking sites allow editing, slope, making profiles, and sharing private and public input [17]. Nowadays, in the myriad fields of application, we collect interconnected data in the shape of networks such as in marketing, biology, epidemiology, sociology, etc. A combined feature of social networks or information networks today, and many other scientific fields, is the structure of their society, which indicates the presence of large groups of nodes that are closely related, with separate connections between these groups. Creating these communities can be of great practical importance for understanding relevant data among them, like organizational structures, academic cooperation, and user communities in the communication network. The disclosure of societies focused on discovering and characterizing the structure of the social network has received great attention over the past few years in social sciences, such as psychology, anthropology, criminology, etc., and more recently in computer science, especially data mining [15].
Facebook users place personal information on their own pages. Some examples are the person’s name, gender, date of birth, email address, marital status, interests, hobbies, favorite sports team, favorite athletes, or favorite music. Moreover, it is possible to identify your friends on Facebook, post messages, post pictures, or any other content. From this information, it is possible to determine the interests of users, as well as gather information about users as knowledge of whether the person is a celebrity, university student, doctor, or singer, as well as knowing his relationships and friends if this information is available [7].

The operation of discovering coherent groups or groups in a network is known as community discovery. It is one of the main tasks of social network analysis. Discovering communities on social networks can be useful in many applications where group decisions are made, for example, sending an important message to a community rather than sending it to everyone in the group or recommending a group of products to a community [3].

The increased availability of data on social networks has stimulated mathematical research into the analysis of social networks. Lately, community discovery on social networks has become one of the most significant challenges in social networks. There are two general ways for communities to discover on social networks, methods that discover societies based on relationships between network users (graph-based methods) and methods based on the combined concerns of users in the network, the propinquity of the second measure of user concerns in networks Social. Most methods of community detection consider one of these sides. In fact, connection or content to get communities in social networks is very important.

All ways of community discovery have special advantages and disadvantages. The advantages may be to improve the thoroughness and quality of community discovery, to discover leaders and societies, to adjust transactions to discover societies appropriate for a specific use, and to identify convergence and external nodes and new methods of community discovery [17].

2. Related Works

The issue of the importance of community detection in social networks has been extensively studied where Clauset A. et al. in 2004 [4] analyzed the structure of social networks in order to discover society. In 2006 Zhou D. et al. [22] discovered the community using probability models. However, all of these ways focus only on the connection and graphical body of social networks but do not take into account the interactions and interests of users and the impact of users on social networks on the Internet. Some ways do not authorize users to enrollment in different societies, which is a problem, and to solve this challenge, researchers have proposed other ways based on the Bayesian eventuality model as proposed by Clauset A. in 2008 [5] a method for community discovery via a Bayesian probability model. Leskovec J. et al. in 2008 [16] also discovered the community by using the statistical characteristics of the social network, in 2008 Goyal A. and others [10] discovered leading figures in social networks. Adnan M. et al. In 2009 [2] also used repeated patterns analysis in social networks. In 2009 Satuluri V. and Parthasarathy S. [19] used pictorial cluster methods to discover a society, just as Khorasgani R. and others in 2010 [15] created an algorithm that revealed potential leaders in social networks. Eliassi-Rad T. and others in 2010 [9] created a hybrid method for community discovering for complex social networks or between more than one network, as this model allows the interaction of members of societies, but this way considerably focuses on the network graphic structure and not on interactions or preferences users. Also, Kanawati R. did this in 2011 [13]. In 2016 Bedi P. and Sharma C. compared several community discovery algorithms, as well as discussing some community discovery applications [7]. In 2017 Moosavi S. and others discovered the community by discovering connections between the community nodes as well as content information, as this method relies on recurring patterns in societies [17]. In 2017 Gulagiz F. and Shahin S. compared the types of hierarchical and non-hierarchical cluster algorithms using different data sets and compared using several aspects such as similarity of results, number of steps, processing time, etc. [11], In 2019 Sethuraman R. and Subhashini R. discovered popular content on social websites by using a visual representation of the data and using a very simple preprocessing of the data [20].

3. Data Visualization and cluster algorithms

3.1 Data Visualization

The human mind naturally tends to understand images more easily than other data such as numbers and words. The more data, the greater the need to represent it in a way that makes it easy to understand quickly. Also, depicting data shows patterns that make analyzing these data more easily than viewing them as numbers and equations [14].
1) Define the visual representation of the data: Data imaging is the study of structured data representation, including attributes and variables of the information unit [14].

2) Steps to visualize data: To visualize the data, the following steps must be followed [19]:

   • **The first step in the imaging process is Mapping:** How information is visualized or how information is encoded in a visual form. It can be achieved when there is an accurate relationship between the data and the visible objects to be described.

   • **The second step of the imaging process is called Selection:** It is the choice of the required data from among the available data according to the objective to be represented. This step is the most important because choosing false data misleads the user to make critical decisions and may cause heavy losses (financial, time, etc.) where unnecessary data should be included.

   • **The third step is Presentation:** It means how to manage and organize information in the available space on the screen effectively. After an intuitive assignment and a clear and accurate selection of data elements, it is important to present them in a more meaningful and understandable way.

3) Methods of visualizing data: There are several ways to visualize data, including [14]:

   • **Pie Chart:** The data is represented in a circular motion, where each piece of the circle represents the ratio of the element to the total ratio of the sum of the elements.

   • **Bar Chart:** The data is represented by bars or columns.

   • **Scatter Plot:** It represents data in separate points and it appears scattered on the drawing.

   And there are other ways to represent data, such as a bubble scatter-like representation, a Histogram, Line chart, and an Area chart, among others.

4. Clustering

   It is the operation of grouping data into multiple groups, such as clusters, so that the objects in each cluster or cluster are very similar but completely different from the objects in other clusters. Similarities are evaluated based on attribute values that describe objects and often include distance scales [1].

1) Basic steps for the cluster: The cluster passes, regardless of the algorithms used, mainly for the following steps [18]:

   • **Feature Selection:** Most cluster models assume that the vector vectors of dimensions n represent all data elements.

   • **Similarity Measure:** Similarity Measure plays an important role in the clustering process as many values are grouped into several clusters.

   • **Clustering Algorithm Selection:** The choice of specific cluster algorithms depends on the desired properties of clustering.

   • **Result Validation:** Detecting whether the results are logical or not.

   • **Results Interpretation:** Results may not be complete unless they are clearly explained.

2) Clustering Algorithms: Cluster algorithms are divided into two categories: Hierarchical and Non-Hierarchical. Hierarchical methods have two different categories: Agglomerative and divisive. Non-hierarchical aggregation methods are also divided into four subcategories: Partitioning, Density-Based, Grid-Based, and others [11].

   a) **K-Means algorithm (Non-hierarchical):** It is the algorithm used in this research in addition to DBSCAN (Non-hierarchical) and we will talk about them. K-Means is the most common method of dividing a cluster. It was first proposed by MacQueen in 1967. In K-Means the center of every cluster is showed by the average amount of the elements in it. It divides a group of elements into several k clusters so that the similarity of the elements between the different clusters...
is low and the similarity of the elements within one cluster is high. Propinquity is measured in terms of the average amount of the elements in a cluster [12]. The K-Means process includes determining the number of clusters K, then randomly determining the number of K centers of the cluster groups and dividing the elements into the nearest cluster formation center according to the base of the nearest neighbor, then calculating the average value of each cluster and making it the new center of the raster group. The process is repeated and deducted according to the error value E and also called SSE (Sum of Squared Errors) and calculated according to the "equation (1)" [21].

\[ E = \sum_{i=1}^{k} \sum_{p \in C_i} (p - m_i)^2 \]  

(1)

Where p denotes an element of cluster Ci, and mi is the central point of cluster Ci. The smaller the E, the more similar the elements within the clusters.

The value of the central point of each cluster mi is calculated as in "equation (2)".

\[ m_i = \frac{1}{n} \sum_{j=1}^{n} p_j \]  

(2)

Where n is the number of elements in the cluster.

b) Density-Based Clustering Algorithm: DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm relies on detecting the contiguity of data points together in a two-dimensional or multi-dimensional space. DBSCAN’s definition of a group is based on the concept of density accessibility. Where point q can be reached directly from point p if it is not far from a certain distance, and if p is surrounded by points sufficiently large that it is part of one cluster [6].

DBSCAN is designed to discover randomly formed groups in any database and at the same time can distinguish between noise points. More specifically, DBSCAN accepts an Eps (ε) value based on a user-defined distance scale and a MinPts value for the minimum number of points that should occur within the Eps radius. Some concepts and terms can be defined to explain the DBSCAN algorithm as follows:

The algorithm begins with the first point in the data set, and all neighbors return this point within the Eps distance. If the total number of these neighbors is greater than MinPts - where this point is the focal point - a new group is created. The point and its neighbors are set in this new group. Then the process is repeated until all points are processed [8].

This algorithm classifies points into three groups [6]:

- **Core point**: the point that is the center of the data in the cluster. And these points form clusters around them.
- **Border point**: It is the points that are at the edges of the clusters and are close to other points and achieve the minimum requirement of points when taken as a center point, but the clusters formed from them are neglected.
- **Noise Point**: These points are far from the rest of the points or there are very few points near them. These points do not form around clusters and do not include any clusters when forming clusters because they are far.
5. **Community Detection**
The search results were obtained by Python. This language contains many libraries to implement algorithms Clusters of data and visualization algorithms for data and algorithms related to data science.

5.1. **The Used Data**
This exploratory data analysis gives insights from the Facebook dataset which consists of identifying users that can be focused more to increase the business. These valuable insights should help Facebook to take intelligent decisions to identify its useful users and provide correct recommendations to them [23].

This dataset `pseudo_facebook.csv` contains 99903 entries with 15 columns. Column names are well defined as Figure 1. So that everyone can interpret easily.

![Figure 1: Entries and columns of dataset Pseduo_facebook.csv.](image1)

5.2. **Visualization and Clustering of the Dataset**
To analyze dataset pseudo_facebook.csv, we applied several experiments:

**Experiment 1**: Figure 2 shows the violin plot of the age data frequency.

![Figure 2: Draw violin plot violin for age data.](image2)

**Experiment 2**: Figure 3 represents the percentage of users of both sexes.
Experiment 3: Figure 4 represents the number of users in each age group, where each group has a 20-year range.

Figure 4: Number of employees by age groups, each group has a 20-year range.

Experiment 4: have counted the number of likes for each age group of 20 years as the Figure 5.

Figure 5: Number of likes by age groups, each group has a 20-year range.
**Experiment 5**: Figure 6 shows the number of likes for ages ranges of 10 years.

![Figure 6: Number of likes by age groups, each group has a 10-year.](image)

**Experiment 6**: Figure 7 shows the comparison between number of likes of females and males.

![Figure 7: Comparison between number of likes of females and males.](image)
**Experiment 7:** Figure 8 shows comparing the number of female friends compared to males.

![Figure 8: Comparing the number of female friends compared to males.](image)

**Experiment 8:** Figure 9 shows a comparison of number of likes that females and males receive for age groups.

![Figure 9: Number of likes received by gender.](image)
Experiment 9: Figure 10 represents the number of likes per day per person.

![Figure 10: Maximum likes per day](image)

**Figure 10:** represents the number of likes per day per person.

Experiment 10: Use the elbow function to find the best value for $k$, the best number of clusters for the dataset `pseudo_facebook.csv` as shown in the Figure 11.

![Figure 11: Elbow Function](image)
**Experiment 11:** Figure 12 The cluster process using Kmeans clustering, and applied on the number of friends and the number of their likes.

![Figure 12: KMeans applied on the number of friends and the number of their likes.](image1)

**Experiment 12:** Figure 13 Shows applied on the number of friends and the number of likes received.

![Figure 13: KMeans applied on the number of friends and the number of likes received.](image2)
Experiment 13: Figure 14 Shows applied on the age and the number of likes.

![Figure 14: KMeans applied on the age and the number of likes.](image)

Experiment 14: Figure 15 Shows started the cluster process using DBSCAN clustering, and we applied it to the number of friends and the number of their likes.

![Fig.15: DBSCAN applied on the number of friends and the number of their likes.](image)
**Experiment 15:** Figure 16 Applied DBSCAN on the number of friends and the number of likes received.

![Figure 16: DBSCAN Applied on the number of friends and the number of likes received.](image1)

**Experiment 16:** Figure 17 Applied DBSCAN on the age and the number of likes.

![Figure 17: DBSCAN applied on the age and the number of likes.](image2)
The Comparison Results of "Experiment 11" with "Experiment 14" and "Experiment 12" with "Experiment 15" and "Experiment 13" with "Experiment 16" showed that Kmean is better on big data.

6. Conclusion
The discovery of societies in the growing social networks today is of great importance. In this research, the basic connotation of social networks, the structure of society, and ways of grouping comparable elements are presented. The application of algorithms to discover connections in actual networks like Facebook, Twitter, LinkedIn etc. can supply a large number of information for countless aims. The discovery and analysis of societies is used in biology, sociology and many other disciplines. This information may be useful for commercial, educational or development purposes.

This paper, demonstrated community discovery using visual representation of data, used the K-Means algorithm and the DBSCAN algorithm for the cluster and observed that Kmean is better on big data than DBSCAN and that through doing many experiments on them.

Also presented the dominant groups of the community on the Facebook social network, as well as showing us whether there are important people in the community such as celebrities and others, and we discover from the community the people's trends, preferences and opinions, which makes the process of proposing appropriate content and commercial advertisements for users.

7. ACKNOWLEDGMENTS
The researchers thank the Department of Computer Science, College of Computer Science and Mathematics, University of Mosul.

REFERENCES
[1] Abbas, O. A. 2008. “Comparisons Between Data Clustering Algorithms”, International Arab Journal of Information Technology (IAJIT), Vol.5, No.3, pp. 320-325.
[2] Adnan, M., Alhajj, R., and Rokne, J. 2009, “Identifying social communities by frequent pattern mining”, paper presented at the 13th international conference information visualization, pp. 413-418.
[3] Bedi, P. and Sharma, C. 2016. “Community detection in social networks”, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol.6, No.3, pp. 115-135.
[4] Clauset, A., Moore, C. and Newman, M. 2008 “Hierarchical structure and the prediction of missing links in networks”, Nature, Vol.453, No.7191, pp.98–101.
[5] Claude, J. 2008, “Morphometrics With R”, Springer, ISBN 978-0-387-77789-4.
[6] Dnv, K., 2011, “FDBSCAN”, Journal of Software Engineering Research, Vol.16, No.1, pp. 11-14.
[7] Dam J., and Velden M., 2015, “Online profiling and clustering of Facebook users”. Decision Support Systems, Vol.70, pp. 60-72.
[8] Derya, B. and Alp, K. 2007, “ST-DBSCAN: An algorithm for clustering spatial–temporal data” ,Data & Knowledge Engineering ,Vol.60, Issue 1 ,pp. 208-221.
[9] Eliassi, T., Henderson, K., Papadimitriou, S., and Faloutsos, C. 2010, “A hybrid community discovery framework for complex networks”, Proceedings of the 2010 SIAM International Conference on Data Mining, pp. 754-765.
[10] Goyal, A., Bonchi, F. and Lakshmanan, L. 2008, “Discovering leaders from community actions”, Paper presented at the proceedings of the 17th ACM conference on information and knowledge management, pp. 499-508.
[11] Gülagiz, F. and Sahin S 2017, “Comparison of hierarchical and non-hierarchical clustering algorithms”, International Journal of Computer Engineering and Information Technology, Vol.9, No.1, pp. 6-14.
[12] Jyoti, Y. and Monika, S. 2013, “A Review of K-Mean Algorithm”, International Journal of Engineering Trends and Technology (IJETT). Vol.4, No.7, pp. 2972-2976.
[13] Kanawati, R. 2011, “LICOD: Leaders Identification for Community Detection in Complex Networks”, IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing, pp. 577-582.
[14] Khan, M. and Khan, S. 2011, “Data and information visualization methods, and interactive mechanisms: A survey”, International Journal of Computer Applications, Vol.34, No.1, pp. 1-14.
[15] Khorasgani, R., Chen, J. and Zaïane, O. 2010, “Top leaders community detection approach in information networks”. 4th SNA-KDD workshop on social network mining and analysis.
[16] Leskovec, J., Lang, K., Dasgupta, A. and Mahoney, M. 2008, “Statistical properties of community structure in large social and information networks”, paper presented at the proceedings of the 17th international conference on world wide web, pp. 695-704.
[17] Moosavi, S., Jalali, M., Misaghian, N., Shamshirband, S. and Anisi, M. 2017, “Community detection in social networks using user frequent pattern mining”, Knowledge and Information Systems, Vol.51, No.1, pp. 159-186.
[18] Rao, I. 2003, “Data mining and clustering techniques”. DRTC Workshop on Semantic Web, Vol.8.
[19] Sadiku, M., Shadare, A., Musa, S., Akujuobi, C. and Perry, R. 2016, “DATA VISUALIZATION”, International Journal of Engineering Research and Advanced Technology (IJERAT), Vol.12, pp. 2454-6135.
[20] Subhashini, R. and Sethuraman, R. 2019, “Identifying Trends in Facebook Usage: A Visual Approach”, International Journal of Recent Technology and Engineering (IJRTE), Vol. 7, p.p. 823-825, Blue Eyes Intelligence Engineering and Science Publication.
[21] Zhang, Z., Zhang J. and Xue H. 2008, “Improved K-means clustering algorithm”, Congress on Image and Signal Processing, Vol.5, pp. 169-172, IEEE.
[22] Zhou, D., Manavoglu, E., Li J., Giles, C. and Zha, H. 2006, “Probabilistic models for discovering e-communities”. 15th international conference on World Wide Web, pp. 173-182, ACM.
[23] The Official Kaggle Website: https://www.kaggle.com.