Facebook Groups Network Analysis

Mariam Sheta*, Mohamed Elshaarawy*, Ashrakat Saeed*, Abdelrahman Said*, Asem Bakr*, Walid Gomaa*,†
*Egypt-Japan University of Science and Technology, Alexandria, Egypt.
†Faculty of Engineering, Alexandria University, Alexandria, Egypt.
{mohamed.elshaarawy, ashrakat.saeed, abdelrahman.said, maryem.abousaad, asem.abdelhamid, walid.gomaa}@ejust.edu.eg

***Abstract*—This research project presents a comprehensive analysis of a network graph derived from an online community, focusing on the data collection process, various network analysis techniques, and the discovery of community structures. The primary objective of this study was to investigate the properties and dynamics of the network graph, shedding light on the underlying social interactions within the community. To initiate the study, data collection was conducted by gathering information from four Facebook groups, constituting the main community. A total of 140,000 members were included in the data-set. Subsequently, an additional 20,000 members were selected from the initial of 140,000 members for further analysis.**

***Index Terms*—Social Network , Vertices, Degree, Community, Clustering Coefficients**

I. Introduction

Network graph analysis offers a powerful scientific approach for investigating the social structures and dynamics within online communities. By examining key aspects such as degree distribution, path analysis, centrality measures, connected components, clustering coefficients, density calculations, network type identification and community discovery, we gain valuable insights into the underlying patterns and interactions that shape these communities. Degree distribution analysis plays a pivotal role in understanding the distribution of node degrees within the network. This analysis reveals the prevalence of highly connected members and identifies potential hubs of influence, providing a quantitative measure of the community's structure. Path analysis enables the exploration of information flow and connectivity patterns by determining the shortest paths between nodes. This analysis uncovers how information disseminates within the community, shedding light on communication dynamics and potential bottlenecks. Centrality measures, including betweenness centrality, closeness centrality and degree centrality, serve as important metrics for identifying influential members and key communication channels. By quantifying the relative importance and influence of nodes, we gain insights into the social dynamics and power structures within the community. Connected components analysis helps to identify distinct clusters or subgroups within the community. By examining the connectivity patterns, we uncover isolated factions or tightly-knit sub communities, facilitating a deeper understanding of community segmentation. Clustering coefficients and density quantify the level of local interconnectedness and cohesion within the network. These metrics allow us to assess the presence of tightly-knit sub communities and the overall structural integrity of the community. The network type identification phase aims to categorize the network based on its characteristics, such as small-world properties, scale-free behavior, or random network properties. This classification provides further insights into the community's organization and interaction patterns. Community discovery techniques are employed to unveil cohesive groups and sub communities based on shared interests, activities, or relationships. By applying advanced algorithms, we uncover underlying community structures that may not be readily apparent, allowing for a more comprehensive understanding of the community's dynamics. By undertaking this research project, we aim to contribute to the scientific understanding of online communities and their underlying social structures. The findings obtained through the comprehensive analysis of network graph data will provide valuable insights for community managers, researchers, and practitioners interested in harnessing the potential of network analysis to optimize community engagement and support decision-making processes.

II. Related Work

One notable book in this field is titled "Social Network Analysis: Methods and Applications" by Stanley Wasserman and Katherine Faust[1] is a foundational work in the field of social network analysis. The book provides a comprehensive introduction to the theory, methods, and applications of social network analysis, covering topics such as network measurement, visualization, and modeling. We found this book to be very helpful in our project on network analysis of Egyptians on Facebook, as it provided us with a solid understanding of the principles and techniques of social network analysis.

In particular, a framework for analyzing the Egyptian Facebook network was developed with the help of the book, by providing a clear understanding of key network measures such as degree centrality, betweenness centrality, and clustering coefficient.

Furthermore, insights into the limitations and challenges of social network analysis were provided by the book, such as the issue of missing data and the potential for bias in sampling. These factors were taken into account in the analysis, and appropriate strategies for addressing them were developed.

Overall, "Social Network Analysis: Methods and Applications" proved to be a valuable resource for the project, providing a solid foundation in the principles and techniques of social network analysis and helping to develop a framework for analyzing the Egyptian Facebook network. Valuable insights into the strengths and limitations of social network analysis were offered by the book, and it helped to address some of the challenges that were encountered in the analysis. This book is highly recommended to anyone interested in social network analysis, as it provides a comprehensive and accessible introduction to the field."

Another notable book titled "What is Social Network Analysis" by John Scott[2] is a valuable resource for anyone interested in learning about social network analysis. The book provides an accessible introduction to the field, covering topics such as network theory, data collection, and network visualization. We found this book to be very helpful in our project on network analysis of Egyptians on Facebook, as it provided us with a clear and accessible introduction to the principles and techniques of social network analysis.

In particular, an understanding of the basic concepts and properties of social networks, including the idea of a network, network data types, and network properties, was facilitated by the book. This understanding was essential for developing a framework for analyzing the Egyptian Facebook network and identifying key nodes and communities within it.

Insights into data collection and measurement, including various methods for collecting network data and measuring network properties were provided by the book. This facilitated the development of appropriate strategies for collecting and analyzing data from the Egyptian Facebook network, and identification of potential sources of bias or error in the analysis.

Additionally, network visualization and analysis techniques were covered by the book, which were essential for exploring the patterns of communication and information flow within the network. Furthermore, examples of practical applications of social network analysis, such as in the study of organizations, communities, and online social networks, were provided.

Overall, "What is Social Network Analysis" was a valuable resource for the project, providing a clear and accessible introduction to the principles and techniques of social network analysis. The book facilitated the development of a framework for analyzing the Egyptian Facebook network and identifying key nodes and communities within it. The insights into data collection and measurement, network visualization and analysis, and practical applications were also highly relevant to the project. This book is highly recommended to anyone interested in social network analysis, as it provides a useful and accessible starting point for further exploration of the field.

III. Dataset

In this project, a social network analysis of Egyptians on Facebook was conducted. To build the dataset, four well-known Egyptian public groups were randomly selected.The four groups are of different interests: Food, Pets, Sarcasm, and Computer Science. A random scraping process was applied to gather approximately 140,000 members from these groups. From this pool, a random sample of 20,000 members was selected for analysis.

An edge list was constructed to represent the relationships between the selected members. Members who belonged to the same group were considered as having an edge between them. If two members were mutual members in more than one group, the weight of the edge was increased by the number of common groups between them.

The dataset consists of 20,000 nodes and 57,639,308 edges, indicating a significant amount of interconnectedness between individuals in different communities.

To extract information about the number of friends each member had, a scraping technique was used to access their public accounts. However, due to privacy settings, only 9313 accounts allowed access to their list of friends for scraping purposes. To ensure a manageable dataset, a maximum limit of 100 friends was set to be scraped from each account.

Overall, this dataset provides valuable insights into the network structure of Egyptians on Facebook, and sheds light on their interactions and connections with other individuals and communities on the platform.

IV. Methodology

In this research paper, the dataset consists of usernames for 4 Egyptian public groups on Facebook. Each member in each group is considered a node, and the edges between them indicate that they are in the same group.

The network graph is constructed using the networkx library in Python, which allows for the creation of an empty graph that is populated with nodes and edges based on the provided dataset. The edges in the network graph are weighted to represent the number of common groups shared by the connected nodes, enabling the capture of the level of association between members in terms of their group participation.

To gain insights into the network community and understand its structure, various analyses are performed on the constructed network graph. These analyses include:

- **Degree Distribution Analysis**:
  The distribution of node degrees in the network is examined to understand the connectivity patterns and identify highly connected nodes (referred to as hubs) or nodes with low connectivity. This analysis allows for the identification of influential members within the network. For the degree distribution analysis, the graph-tool library in Python is employed. The algorithm is applied to calculate the degree of each node in the network. This involves iterating through all the nodes and counting the number of edges connected to each node, which determines its degree. The degree distribution is subsequently obtained by tallying the occurrences of each degree value across all nodes.
  The time complexity of the degree calculation algorithm using graph-tool is $O(N)$, where N represents the number of nodes in the network. The algorithm iterates through each node once to count its edges, and the counting process incurs constant time per node. Thus, the overall time complexity is linear with respect to the number of nodes.
- **Path Analysis:**
  In the path analysis, the examination of shortest paths between nodes in the network is carried out to investigate

the accessibility and connectivity among members. The average shortest path length is then computed to determine the overall efficiency of information flow within the network.

To perform the path analysis using the graph-tool library, a well-known algorithm called Dijkstra's algorithm is utilized. This algorithm finds the shortest path between a source node and all other nodes in the network. It iteratively explores neighboring nodes, updating the distances from the source node until the shortest path to each node is determined. The algorithm continues until all reachable nodes have been visited.

The time complexity of Dijkstra's algorithm implemented in graph-tool is approximately $O(N^2)$, where N represents the number of nodes in the network. This is because the algorithm requires iterating over all nodes in the network and examining their neighboring nodes. However, graph-tool uses various optimizations to enhance performance and reduce the computational burden.

- **Centrality Analysis:**

  In the centrality analysis, various centrality measures, including degree centrality, betweenness centrality, and closeness centrality, are utilized to identify the most influential nodes within the network. These measures provide insights into the importance and prominence of individual nodes in terms of their connectivity and influence on information flow and network cohesion. For the centrality analysis, we employed the Igraph library to calculate the degree centrality, betweenness centrality, and closeness centrality measures. Each measure is computed using specific algorithms tailored for efficient centrality analysis.

  Degree Centrality: Degree centrality measures the number of edges connected to a node. In Igraph, the degree centrality is computed using an algorithm that iterates through each node and counts the number of incident edges. The time complexity of calculating degree centrality in Igraph is $O(N)$, where N represents the number of nodes in the network.

  Betweenness Centrality: Betweenness centrality quantifies the extent to which a node lies on the shortest paths between other nodes. Igraph employs the Brandes' algorithm to calculate betweenness centrality efficiently. The algorithm iterates over each node as a potential source node and determines the fraction of shortest paths passing through that node. The time complexity of the Brandes' algorithm for betweenness centrality in Igraph is approximately $O(V*(V+E))$, where N represents the number of nodes in the network.

  Closeness Centrality: Closeness centrality measures the average distance between a node and all other nodes in the network. In Igraph, the closeness centrality is computed using an optimized algorithm based on breadth-first search (BFS). The algorithm calculates the shortest path lengths from a source node to all other nodes, and then the reciprocal of the sum of these path lengths gives the closeness centrality. The time complexity of computing closeness centrality in Igraph is approximately $O(V*(V+E))$, where N represents the number of nodes in the network.

  By applying the degree centrality, betweenness centrality, and closeness centrality measures using Igraph, we gain insights into the importance and centrality of nodes within the network. The efficient algorithms implemented in igraph ensure reasonable computational performance for calculating these centrality measures.

- **Connected Components Analysis:**

  In the connected components analysis, distinct subgroups or communities within the network are identified by examining the connected components. This analysis provides insights into the level of fragmentation or cohesion present in the network, allowing the identification of clusters of individuals who share common interests or affiliations.

  For this analysis, we utilized the graph-tool library to identify the connected components within the network. The graph-tool library implements efficient algorithms for connected components analysis, enabling us to identify the distinct subgroups present in the network.

  The algorithm used in graph-tool for connected components analysis is based on a depth-first search (DFS) traversal. It starts from a randomly selected node and explores the network by visiting adjacent nodes, marking them as part of the same connected component. This process continues until all nodes in the connected component have been visited. The algorithm then selects another unvisited node and repeats the process until all nodes have been assigned to a connected component.

  The time complexity of the connected components analysis in igraph is approximately $O(N+E)$, where N represents the number of nodes in the network and E represents the number of edges. This complexity indicates that the run-time scales linearly with the size of the network, making it computationally efficient even for large-scale networks.

  By conducting the connected components analysis using igraph, we gain insights into the distinct subgroups or communities present in the network, allowing for a better understanding of the network's structure and the relationships between its members.

- **Clustering Coefficients:**

  In the clustering coefficient analysis, the extent to which nodes in the network cluster together is assessed. This analysis is performed to gain insights into the presence of tightly-knit communities or groups within the network and to understand the level of clustering and community structure. The algorithm used to calculate the clustering coefficient involves examining the local neighborhood of each node and determining the proportion of connections between its neighbors.

  To perform the clustering coefficient analysis using graph-tool, the local_clustering function is utilized. This

function calculates the clustering coefficient for each node in the network based on the described algorithm. It efficiently iterates over the nodes, evaluates their neighborhoods, and computes the clustering coefficient.
The time complexity of the clustering coefficient analysis algorithm in graph-tool is dependent on the number of nodes and the average degree of the network. In general, the algorithm has a time complexity of $O(N+E)$, where N is the number of nodes and E is the number of edges in the network. This time complexity allows for efficient computation of the clustering coefficient even for large networks.
Transitivity:
To calculate the transitivity, the algorithm iterates over each node in the network and examines its neighbors. For each neighbor pair, it checks if there is an edge connecting them. By identifying and counting the number of triangles, the algorithm determines the level of clustering or transitivity in the network.
The time complexity of the transitivity algorithm in graph-tool is dependent on the size of the network, typically denoted as n. In the case of graph-tool, the time complexity is approximately $O(n^3)$, considering the need to iterate over each node and check the connections with its neighbors.
It is important to note that graph-tool may employ optimizations or techniques to improve the efficiency of the algorithm and handle specific network characteristics. These optimizations may affect the actual time complexity observed in practice.

- **Density Analysis:**
  In the density analysis, the level of interconnectness within the network is quantified by calculating the density, which represents the ratio of actual edges to the total possible edges in the network. This analysis provides insights into the overall connectivity and interaction among the network's members. Higher density values indicate a more densely connected network, indicating a higher level of information flow and communication.
  To perform density analysis using graph-tool, the density function is utilized. This function calculates the density of the network by dividing the actual number of edges by the total number of possible edges. It efficiently computes the density by considering the graph's structure and edge connections.
  The time complexity of the density analysis algorithm in graph-tool is dependent on the number of nodes and the number of edges in the network. In general, the algorithm has a time complexity of $O(N + E)$, where N is the number of nodes and E is the number of edges. This time complexity allows for efficient computation of the density even for large networks.

- **Network Type:**
  The network type (directed or undirected) is determined to understand the nature of relationships between members. This information helps assess the flow of information within the network and identify potential biases or communication patterns.

- **Community Discovery:**
  The community discovery analysis using graph-tool is performed using the Louvain algorithm. The Louvain algorithm is an iterative algorithm that optimizes the modularity measure to identify communities within the network.
  The algorithm starts by assigning each node to its own community. It then iteratively optimizes the modularity by merging and rearranging communities. In each iteration, the algorithm evaluates the change in modularity resulting from moving a node to a different community. The node is moved to the community that yields the maximum increase in modularity. This process continues until no further improvement in modularity is possible.
  The Louvain algorithm achieves high computational efficiency by employing a two-phase approach. In the first phase, the algorithm optimizes the modularity locally within small communities, resulting in a compressed representation of the network. In the second phase, the algorithm builds a new network where the compressed communities become the nodes. The same optimization process is applied to this new network, leading to the detection of larger communities.
  The time complexity of the Louvain algorithm is dependent on the size of the network and the number of iterations required for convergence. In general, the Louvain algorithm has a time complexity of approximately $O(nlogn)$, where n is the number of nodes in the network. This makes it computationally efficient for large-scale networks.

By conducting these analyses, a comprehensive understanding of the Facebook groups network, its underlying structure, influential members, communication patterns, and community dynamics is aimed to be achieved. These insights can provide valuable information for understanding the social dynamics and interactions within the Egyptian public groups on Facebook, contributing to a deeper understanding of online communities and their impact.

V. Results

1) Visualization of the Network

   - No.of Vertices :20000
   - No.of Edges: 57639308
   - No of communities: 4

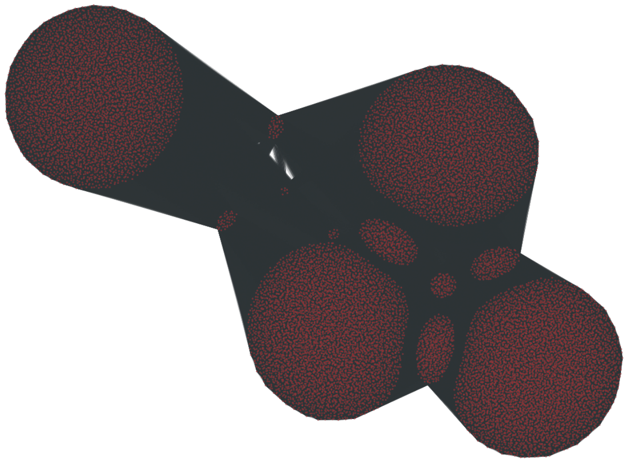

Fig. 1: Visualized Network Graph

2) **E**xploration of the Network

a) **Degree Distribution Analysis**

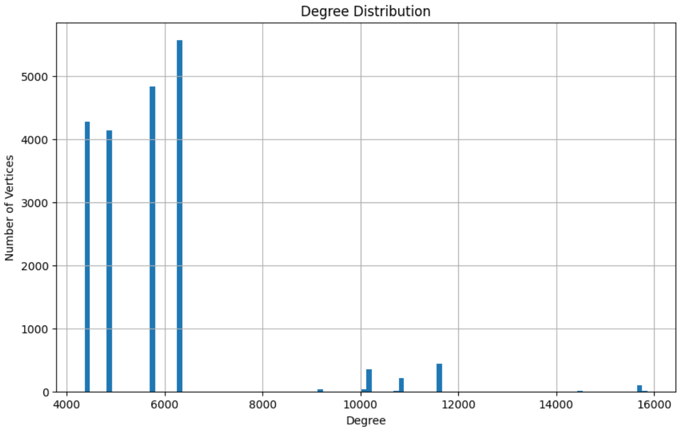

Fig. 2: Distribution Degree

Based on the graphical representation provided above, it can be deduced that a subset of approximately 500 vertices demonstrates a substantial degree of connectivity, as evidenced by their involvement in approximately 12000 edges. This observation implies a significant level of interconnectivity among these vertices, highlighting their pronounced role in facilitating information flow in the graph.

b) **Path Analysis**

The result of this analysis, indicated by a graph diameter of 2, signifies that the maximum shortest path length between any two vertices in the graph is 2. In other words, it suggests that the graph exhibits a relatively compact and tightly connected structure, where most vertices can be reached from one another in a very short number of steps. This implies a high level of connectivity and accessibility within the graph, indicating efficient communication and information transfer between vertices.

c) **Connected Components Analysis**

Number of Connected Components: 1

Connected Component 1: Size = 20000

The result of having only one connected component with a size of 20000 implies that all vertices in the graph are interconnected and form a single cohesive unit. There are no isolated or disconnected subsets of vertices within the graph. This suggests a high level of connectivity and coherence in the data or network represented by the graph. Information, influence, or interactions can easily flow between any pair of vertices within the connected component, allowing for efficient communication and exchange of information throughout the entire system.

d) **Density analysis**

Graph Density: 0.2882109505475274

Given the no. of vertices and the no. of edges, the relatively high graph. density suggests that the vertices in the graph are well-connected, indicating a relatively dense network of relations or interactions between the entities represented by the vertices.

e) **Network type**

- Graph is Undirected
- Graph is connected.
- Network is not homogenous.

f) **Clustering Coefficients**

- This graph indicates that the clustering coefficients is 0.9718. This result indicates a high level of clustering or connectivity. within the graph. (because we are assuming that all group members are connected so it makes sense that each node that connected with all group members around that's why the clustering coefficient is high)
- Triangles: 115278616

  The large number of triangles in our graph indicates a high level of clustering or interconnectedness among the nodes. It suggests that nodes in the graph tend to form local clusters or communities, where multiple nodes are connected to each other in closed loops. This clustering of nodes can signify common characteristics, shared interests, or similar relationships among the entities represented by the nodes.

  Average number of triangles: 5763.9308

  The average number of triangles in the graph is 5763.9308, which indicates the typical level of clustering interconnectedness among the nodes in the graph. On average, each node is involved in approximately 5763 triangles, suggesting a significant level of local clustering.

- Maximum number of triangles: 15868
  The maximum number of triangles found in the graph is 15868. This value represents the highest level of clustering observed for any individual vertex in the graph. The vertex with the ID 321 has the maximum number of triangles, indicating that it is a highly connected node that participates in many triangular relationships with other nodes in the graph.
- Degree with the maximum number of vertices: 15868
- Number of vertices having the most abundant degree: 5
  The degree with the maximum number of vertices in the graph is 15868. This indicates that there is a specific degree value that is most common among the vertices in the graph, and it occurs in a significant number of vertices. The degree of a vertex represents the number of edges connected to that vertex.
  Furthermore, the number of vertices having the most abundant degree is 5. This means that there are multiple vertices in the graph that have the same highest degree value of 15868. In this case, there are 5 vertices that share this degree value, indicating a group of highly connected nodes.
- Transitivity of the graph
  Transitivity of the graph: (0.8956, 0.0027) The first value, 0.8956, represents the transitivity coefficient. Transitivity measures the tendency of nodes in a graph to form triangles or clusters of interconnected nodes. It is a measure of how interconnected and clustered the graph is. A higher transitivity coefficient indicates a greater tendency for nodes to form clusters.
  The second value, 0.0027, is the p-value associated with the transitivity coefficient. The p-value helps determine the statistical significance of the observed transitivity coefficient. In this case, the small p-value suggests that the observed transitivity in the graph is unlikely to have occurred by chance.
  In summary, the transitivity coefficient of 0.8956 indicates a high level of clustering and interconnectedness in the graph. This means that nodes in the graph tend to form triangles or clusters, with many nodes being connected to each other. The small p-value further supports the statistical significance of this observed transitivity.

g) **Centrality analysis**

Largest Degree Centrality: 15868

The largest Degree Centrality value of 15868 indicates that there is a node in the graph with the highest number of connections to other nodes. This node is highly central and influential in terms of the overall network connectivity. It serves as a key hub within the graph, connecting many other nodes. And in our graph this node is most likely to be presented in all 3 groups.

Largest Betweenness: 1853798 The largest Betweenness value of 1853798 indicates that there is a node in the graph that plays a critical role in connecting different parts of the network. Nodes with high Betweenness act as bridges or intermediaries, facilitating communication and information flow between different nodes or clusters in the graph.

Largest Closeness: 0.828802

The largest Closeness value of 0.828802 indicates that there is a node in the graph that is relatively closer to all other nodes in terms of geodesic distance. This node can be considered as having high accessibility and efficiency in terms of communication and information dissemination within the network.

VI. Conclusion

From the network conclusion, we concluded that there's subgroups within our network. Regarding our four communities this reveals shared interests, affiliations, or ideologies among the 4 group members, The above conclusion could enable researchers to study the formation, evolution, and behavior of these communities.

9313 users from a total of 140,000 users, their accounts are public. and we reached their friends. In social sciences these data and graphs could be used for demographic studies and behavior studies for these people a social behavior.

References

[1] K. F. Stanley Wasserman, *Social Network Analysis: Methods and Applications*, 1994.

[2] J. Scott, "What is social network analysis?" 2012.