Identifying Social Media Influencers using Graph Analytics

Pankti Joshi
Master's in Computer Science
Lakehead University
Thunder Bay, Canada
pjoshi@lakeheadu.ca

Dr. Sabah Mohammed
Faculty of Computer Science
Lakehead University
Thunder Bay, Canada
mohammed@lakeheadu.ca

***Abstract*—Social network analysis has been an essential topic with broad content sharing from social media. Defining the directed links in social media determine the flow of information and indicates the user's influence. Due to the enormous data and unstructured nature of sharing information, there are several challenges caused while handling data. Graph Analytics proves to be an essential tool for addressing problems such as building networks from unstructured data, inferring information from the system, and analyzing the community structure of a network. The proposed approach aims to determine the influencers on Twitter data, based on the follower's count as well as the retweet count. Several graph-based algorithms are implemented on the data collected to find the influencer as well as communities in the network.**

***Index Terms*—Social Media, Twitter, Influencers, Graph Analytics, Graph Database**

I. Introduction

Nowadays, people are relying more on the visual representation of the data. It has also been observed that social media influencers have a great influence on society. It has become essential to identify the influencers in the network for the business organization good as well as the people's point of view. The impact of social media has increased, affecting people's opinions and thoughts. Nowadays, people rely too much on the content shared on the social network. Thus, to differentiate if the content shared comes from a reliable source, it is crucial to identify the social media influencer in the social network.

In the proposed paper, Twitter data is into consideration from which a graph is constructed using Neo4j. The twitter data is about the first GOP debate on Twitter and has been modified a little considering more information and focus on identifying the influencers in the data. For the social network analysis, the nodes represent the influencers, and they are connected by follows relationships. Various graphical algorithms have been implemented using Neo4j, which would help to identify the influencers in the network. The number of followers and the retweet count are primary attributes taken into considerations for deciding the influencers in the system.

The rest of the paper is organized as follows: Section II provides the problem definition. Section III offers Related Research Work. The information about the Dataset is stated in Section IV. The Research Methodology and Prototype details are encrypted in Section IV and V, respectively. Section VII states Experimental Analysis. Finally, the Results and Conclusion is mentioned thereafter.

II. Problem Definition

The goal was to define a robust model to identify the social media influencers from the real-time data. Due to the unstructured nature of data in the social network, it is challenging to infer potent information from the system and to identify the communities. The proposed method demonstrates the use of graph analytics to identify social media influencers. This would reduce a lot of inference time to locate the specifications of the influencers while doing the social network analysis. Graph-based analytics would enable a better comparison of the data and statistics. Thus, the proposed problem defines methods for computational social network analysis that defines three significant fields, creating the social network, analyze the structure and dynamics of a community, identify the most reasonable influencer in the system, and build inferences from the social networks.

III. Related Research Work

The authors, Edward Dong-Jin KIM and Brian Jia-Lee KENG, in their patent US20150120713, "Systems and Methods for Determining Influencers in a Social Data Network," proposes a system to identify the influencers in the network[9]. According to their research, the proposed model is fed in with the information to obtain a topic, based on the matter the users are determined from the network data. Each user represents a node in the model and then establish the relationship between the users. A topic network is created based on the users and their contacts. The first ranking process is used to rank the users in the system. The second-ranking of the users is adjusted based on the removal of outlier users. Identify the communities based on the second-ranking of the users. Identify the standard features in the communities identified. Thus, the output can be a successful detection of the communities, along with the differentiating community features. This proposed computing system proves to be essential in case of obtaining the communities for a particular topic. The idea combines hardware and software to identify an author of each data object. The patent idea revolves around authors

of the textual documents and identifying similar communities. This idea of the computing system is much different from the model proposed in this paper, as it is more focused on the social media influencers and community with the use of software applications called Neo4j. The significant advantage of the idea proposed in this paper is easy portability and easy implementation. Here the twitter-based influencers are identified, and communities are recognized. This model is efficient for small chats and discussions that were done on social networks.

The researchers Lutu, and Patrica in their paper, "Using Twitter Mentions and a Graph Database to Analyse Social Network Centrality," the author Patricia uses the Centrality algorithm to calculate the usefulness of mentions on twitter to create a potential social network[1]. The follows relationship is taken into account to compute influential users. The graph-based centrality measures are taken into considerations like degree centrality, closeness centrality, betweenness centrality, and page rank algorithm. There are two models proposed - one just based on the follows relationship and another with mentioned relationships. In the proposed model by Lutu and Patrica, there are two relationships defined mentioned and follows in the graph network. The author successfully concluded that graphs based on the mentioned relationship provide more information as compared to figures based on follows relationship. This has been experimented based on weighted centrality measures. The concept proposed in the paper is different from the idea proposed by Lutu and Patrica because the primary purpose in the submitted paper focuses on finding the correct influencers in the social network, rather than debating on the kind of relationship performing better for social media network analysis.

Nattapon, Thanaphoom, Peerapon in their research work, "Social Network Analysis of Calling Data Records for Identifying Influencers and Communities," describes the methods of cleaning data and find the influencers in the social network in the field of telecommunication[2]. In the study, the actors are identified based on betweenness centrality, closeness centrality, modularity, in-degree, and out-degree. Significant steps performed include Data cleaning, Graph Visualization, influencers detection, and then community detection. The main idea proposed by the author is comparing the resultant influencers before cleaning and after cleaning the data using the web application created. The plan proposed by authors in this paper support numbers and provides the influencers based on the numerical data. In the paper intended model, majorly, any social network data can be analyzed - mathematical or categorical, and provide the influencers in the network. There is no idea to develop web applications for data considered yet for the proposed model.

The author Zeynep Zengin Alp, Şule Gündüz Öğüdücü, in their research work, "Identifying topical influencers on twitter based on user behavior and network topology," focuses more on finding the authenticity of the influencers based on the attributes like topical focus rate, activeness, genuineness, and speed of getting a reaction on specific topics. The authors have implemented their research called "Personalized PageRank" on the Turkish tweets combining the data obtained from user actions in Twitter as well as data from the network topology[11]. The algorithms are implemented in the distributed computing environment for making high-cost processing graphs easier. The proposed model analyses data using several algorithms to identify the influencers in the social media network. More focus and calculations have been carried on detecting the communities rather than focusing on the authenticity of the data and influencers.

Sounthar and Dr. B.Yinayaga, in their work, "Exploring gender-based influencers using Social Network Analysis," analyze the likes in Facebook cover photos to identify the influencers and their gender[3]. This analysis is done using the clustering coefficient, degree analysis, and triadic census. The results state that there are a large number of male influencers in the social network. The primary aim of the paper is to understand the importance of social network analysis and social interactions. The proposed concept in this paper majorly focuses on the social network analysis for the data from social media (Twitter data). Influencers are identified based on the "FOLLOWS" relationship, rather than review likes on Facebook.

IV. Dataset

A. Original Dataset

For the proposed model, the dataset from Kaggle "First GOP Debate Twitter Sentiment," is used. This dataset was published to analyze the tweets on the first 2016 GOP Presidential debate[8]. The dimension of the provided is 21 columns and 13871 rows. This data initially came from the Crowdflower's Data for Everyone Library. This dataset offers various attributes namely, id, candidate, candidate confidence, relevant yn, relevant yn confidence, sentiment, sentiment confidence, subject matter, subject matter confidence, candidate gold, name, relevant yn gold, retweet count, sentiment gold, subject matter gold, text, tweet coord, tweet created, tweet id, tweet location, user timezone. The twitter users in the dataset are commenting about the US political leaders based on the GOP debate 2016. The dataset was originally downloaded in the form of .csv file format.

B. Dataset Preprocessing

Before applying the graph analytics, several discrepancies in the data were removed. Data cleaning was the initial step in the data preprocessing. The data cleaning step comprises removing the null values, considering only the columns which are necessary and discarding the additional columns, and changing the format of the numbers to a unique arrangement of the column. The fields such as Followers and Following were added. This contains the id numbers of all the followers and the following in an array format. Lastly, the dataset was converted into JSON format for further computations. The necessary information majorly considered after data preprocessing includes id, text, followers, following, candidate, retweet count, and subject matter.

V. Research Methodology

This section provides a brief idea about few essential terminologies related to the proposed model.

A. Social Network Properties

The term social network refers to the set of socially relevant nodes that are attached by one or more relations. The relation can be anything like friendship, fellowship, employment, or anything else that supports social network communications. Nodes or vertices represent the main actors in the network, links, or edges represent the join between the actors that is with what relationship is an actor connected to the other member within the system[6]. The size of the network refers to how many numbers of nodes it contains. In the mathematics side, the social network can be represented as a Graph, G = (V, E) by a set of nodes n, V with

$$|V| = n$$

and a set of edges m where,

$$E = (u, v)|u, v \in V$$

In this dissertation, the notation (i, j) is also used in the case of (u, v), to name the link from node i (respectively vertex u) to node j (respectively vertex v). For a directed graph, the flow of the information is clearly specified between the nodes. In a directed graph, the vertex has two types of degrees. The number of edges point in v gives the in-degree of a node v, while the number of edges points out v defines the out-degree of the node v.

B. Social Media Influencer

With the increased emergence of social networking sites, there is a real need to find the actors and there mutual connections with other people. Social media influencer refers to a person who can impact the decisions in society with their knowledge and expertise. With the use of social media tools like Facebook, Twitter, Instagram, and LinkedIn, people have influenced interactions among members who hold vital information for the company. Nowadays, digital influencers are a part of marketing strategies to influence the public and thus market their products and therefore engage the communities to retrieve information from the communities and know till what extend are the people satisfied by their products and services. It is essential to identify the influencers as their primary role includes word-of-mouth marketing of the products in the market. This could impact public opinion both positively as well as negatively. This can improve brand awareness, enhance the adoption of innovation. The well-connected audience can influence the views of the audience to brand campaigns and products.

C. Centrality in Social Networks

One of the significant features kept in mind to identify the social media influencer is the centrality measure. Centrality in graph theory indicates the most important vertices within a graph. The node which has the center position is node-centrality. Centrality measures are carried out to find the influencer in the network. This can be done by figuring out the most popular vertex. The high degree of popularity means higher chances of communication within the communities. However, this theory cannot always be correct where demand goes with the influencers in the network. However, there are several measures created for node centralities such as degree, closeness, betweenness, and eigenvector.

The Degree Centrality measures the number of contacts a node might have in the network. The more the number of connections or edges a node has, the higher is the centrality score of the vertex. The Closeness Centrality is a method to detect the nodes that can spread the information efficiently through the graph. It measures the average distance to all other nodes. Nodes with shortest distances to all other nodes get the higher closeness score. The Betweenness Centrality computes the number of times a node act as a bridge along the shortest path between two other nodes. By this theory, the vertices have a higher probability of occurring on a randomized chosen shortest path between two randomly chosen nodes that have a high betweenness. Eigenvector Centrality measure refers to the idea that the node is vital if its neighborhood is also famous. That means that the centrality score depends on the neighborhood nodes and their centrality score. The algorithm considers the direct as well as the indirect links with the nodes in the network.

D. Tools and Technology

Neo4j is a native open-source graph database management system. It is a NoSQL database and provides an ACID-compliant transactional backend for all the applications. This tool is enabled by drivers for popular programming languages such as python, java, javascript, and many more. The declarative query language is known as Cypher. It has a flexible graphical property that can adapt over time. It can accommodate large graphs with billions of nodes. The various graph-based algorithms can be computed using the Graph Algorithms Playground. Thus, this is an efficient tool for the proposed model and can efficiently calculate the influencers in the network.

E. Page Rank Algorithm

Page Rank Algorithm is a variant of the Eigenvector centrality. It is an algorithm that measures the directional influence of the nodes. It considers the impact of the neighbors as well as their neighbors. For instance, with few influential friends, scores can be higher than those with many less prominent friends.

Page rank is computed by computing the node rank based on degree centrality or by randomly traversing the graph or counting frequency of hitting node during the walks. It counts the number of quality links to a page, which would determine how valuable a page is. The higher volume of influential pages that are linked to a page of utter importance is the assumption used in this algorithm.

F. Community Detection Algorithms

The most popular algorithms considered for community detection include the Louvain algorithm and the Label Propagation algorithm. The Louvain algorithm is a hierarchical clustering algorithm. It is used for detecting the communities in a network. This increases the modularity score for each community on the basis of the quality of the nodes assigned to the network is good. This is on the basis of the density of the nodes connected in a community, as compared to how those nodes would be connected in a random network. The Louvain algorithm works by recursively merging the communities in a single node and executes modularity on the condensed graphs. The Label Propagation Algorithm (LPA) is used for finding the communities in a graph as well. But this is a fast algorithm as compared to the Louvain algorithm. It does not require prior information about the communities and can detect the communities using the network structure alone. One can assign initial labels to a narrow range of generated solutions. Hence, you can pick some initial communities beforehand and can use this as a semi-supervised method of locating the communities.

VI. Prototype

The proposed model evaluates the influencer's score on the preprocessed data. The dataset considered can be fed in the Neo4j database. This data can be read using the cypher query language. There is a total of 13871 nodes formed while executing the query. Due to the time and space complexity, only 300 rows can be shown at a time in the Neo4j browser. All the instances are executed using the Neo4j browser interface. Once the data is successfully loaded in the browser, one can download the Graph Playground Algorithms to perform the relationships between the nodes and find the most influential person in the network.

Fig 1. provided depicts the graph constructed for an instance of 20 influencers from the dataset. The nodes represent the name of the influencers, and the links defined is based on the "FOLLOWS" relationship[7]. The graph construction is based on the "FOLLOWS" relationship. The dataset contains a column with an attribute named "candidate" and "sentiment." The most influential person is obtained using the "FOLLOWS" relationship. There are multiple algorithms executed on the same dataset so that one can infer the output from various algorithms and can compare the results obtained in an efficient manner. There are in total 249660 properties developed for the dataset. The properties include the relationship developed between the nodes in the graph.

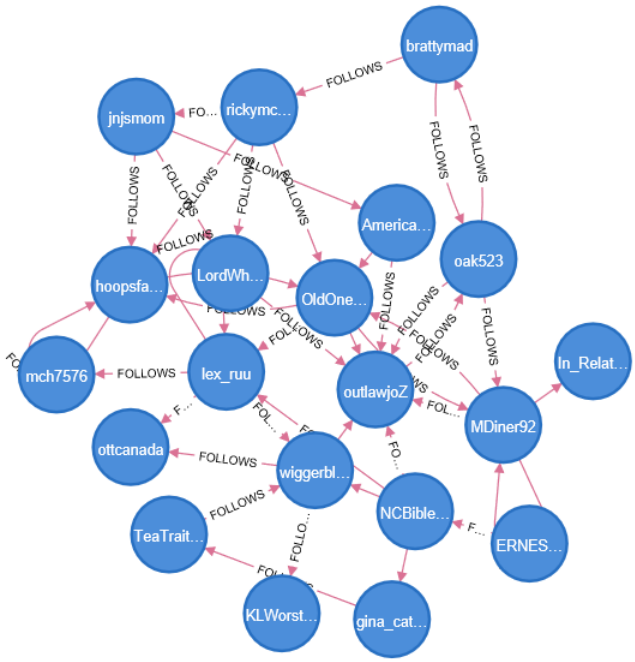

Fig. 1. 20 Influencers in Dataset

TABLE I
Degree

ID	Score
110	984
284	975
109	975
73	969
19	969

TABLE II
Page Rank Algorithm

ID	Score
172	22.17
284	22.15
110	22.12
16	22
59	21.9

TABLE III
Betweenness

ID	Score
110	361239.052
284	358073.4182
172	357730.6685
237	357253.9566
16	356694.8526

TABLE IV
Closeness

ID	Score
172	0.536059365
284	0.535976505
110	0.535955794
59	0.535583272
256	0.535459213

TABLE V
TRIANGLES AND CLUSTERING COEFFICIENT

ID	Triangles	Coefficient
1	11588	0.007546098
2	9095	0.005990832
4	8446	0.0057802
8	7935	0.004972456
3	7577	0.005107506

TABLE VI
LABEL PROPAGATION AND LOUVAIN MODULARITY

ID	Label Propagation Community	Louvain Modularity Community
11	2	210
12	2	210
13	2	210
14	2	210
15	2	210

The algorithms that have been implemented to identify the influencers in the social media network include Page Rank algorithm, Triangles, Clustering Coefficients, and Closeness Centrality. The algorithms performed to determine the communities include the Louvain algorithm and the Label Propagation Algorithm. The experimental analysis of the proposed model is discussed in the section below.

VII. EXPERIMENTAL ANALYSIS

This section provides a brief output about the influencers based on the various algorithms implemented. The tables provided below shows the ID of the top five influencers according to the algorithms implemented along with the score. Here, the Degree, Page Rank, Betweenness, and Closeness algorithms represent the Centrality algorithm. The Louvain algorithm, the Triangles, and, Clustering Coefficient algorithm, and Label Propagation algorithm are used as the Community Detection Algorithms. Table 1, Table 2, Table 3, Table 4, is for Degree Centrality, Page Rank Algorithm, Betweenness, Closeness. Table 5 and Table 6 are for community detection algorithms like triangles and clustering coefficient, label propagation, and Louvain modularity algorithm.

According to the experimental analysis and referring from the centrality algorithms, the top three influencers in the GOP debate are with ID's 110, 284, and 172. While executing the Community Detection algorithm, it can be commented that ID 11, 12, 13, 14, and 15 can form a community with the most substantial correlation factors. Also, both the algorithms-Label Propagation as well as Louvain algorithm provides exactly the same results with the community score 2, and 210 respectively. It can also be stated that ID's 1,2,4,8, and 3 are highly dense nodes with highest number set of three nodes connected, where each node has a relationship to all other nodes.

VIII. RESULTS

From the experimental analysis and execution of the proposed model, the stated results are mentioned in this section. The inferred information from the output obtained is that "In the GOP debate held in August 2016, most of the people have a negative sentiment for Donald Trump, as per a tweet provided by Clever Otter (ID: 110) and the major subject matter is, "FOX News or Moderators" which needs to be taken care of as twitted by fc7822 (ID: 172). Both the candidate's ID 172 as well as 284 support No candidates, as mentioned in the candidate poll.

IX. CONCLUSION

It can be finally concluded that Graph Analytics is a major tool to identify the major influencers in the social network. This paper is based on identifying the influencers based on the "FOLLOWS" relationship. This proposed model can be implemented by using several other attributes too, such as the retweet count or the candidate mentioned as the key factors to identify the influencers. The proposed model focuses more on implementing various centrality algorithms and Community detection algorithms in the Neo4j Graph Algorithms Library to explore the Twitter Graph.

REFERENCES

[1] Lutu, Patricia E. Nalwoga. "Using Twitter Mentions and a Graph Database to Analyse Social Network Centrality." In 2019 6th International Conference on Soft Computing Machine Intelligence (ISCMI), pp. 155-159. IEEE, 2019.

[2] Werayawarangura, Nattapon, Thanaphoom Pungchaichan, and Peerapon Vateekul. "Social network analysis of calling data records for identifying influencers and communities." In 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 1-6. IEEE, 2016.

[3] Manickavasagam, Sounthar, and B. Vinayaga Sundaram. "Exploring gender based influencers using Social Network Analysis." In 2014 Sixth International Conference on Advanced Computing (ICoAC), pp. 224-228. IEEE, 2014.

[4] Hu, Yuheng, Shelly Farnham, and Kartik Talamadupula. "Predicting user engagement on twitter with real-world events." In Ninth International AAAI Conference on Web and Social Media. 2015.

[5] Werayawarangura, Nattapon, Thanaphoom Pungchaichan, and Peerapon Vateekul. "Social network analysis of calling data records for identifying influencers and communities." In 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 1-6. IEEE, 2016.

[6] Tridetti, Stéphane. "Social network analysis: detection of influencers in fashion topics on Twitter." (2016).

[7] Mark Needham, "Finding influencers and communities in the Graph Community," May 15, 2019. [Online]. Available: https://medium.com/neo4j/finding-influencers-and-communities-in-the-graph-community-e3d691296325. [Accessed February 20, 2020].

[8] Crowdflower's Data for Everyone library, "First GOP Debate Twitter Sentiment Analyze tweets on the first 2016 GOP Presidential Debate," 2017. [Online]. Available: https://www.kaggle.com/crowdflower/first-gop-debate-twitter-sentiment. [Accessed February 10, 2020].

[9] Kim, Edward Dong-jin, and Brian Jia-lee Keng. "Systems and Methods for Determining Influencers in a Social Data Network." U.S. Patent Application 14/522,471, filed April 30, 2015.

[10] Francalanci, Chiara, and Ajaz Hussain. "Influence-based Twitter browsing with NavigTweet." Information Systems 64 (2017): 119-131.

[11] Alp, Zeynep Zengin, and Şule Gündüz Öğüdücü. "Identifying topical influencers on twitter based on user behavior and network topology." Knowledge-Based Systems 141 (2018): 211-221.

[12] Lakshmi, NV Muthu, and T. Lakshmi Praveena. "A Review of Graph Based Algorithms in Social Media Data Analytics." JETIR, no. May (2018).

[13] Bakshy, Eytan, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. "Identifying influencers on twitter." In Fourth ACM International Conference on Web Seach and Data Mining (WSDM). 2011.