Social Network Analysis of Twitter to Identify Issuer of Topic using PageRank

Sigit Priyanta[1], I Nyoman Prayana Trisna[2]
Department of Computer Science and Electronics
Universitas Gadjah Mada

***Abstract*—Twitter as widest micro-blogging and social media proves a billion of tweets from many users. Each tweet carry its own topic, and the tweet itself is can be retweeted by other user. Social network analysis is needed to reach the original issuer of a topic. Representing topic-specific Twitter network can be done to get the main issuer of the topic with graph based ranking algorithm. One of the algorithm is PageRank, which rank each node based on number of in-degree of that node, and inversely proportional to out-degree of the other nodes that point to that node. In proposed methodology, network graph is built from Twitter where user acts as node and tweet-retweet relation as directed edge. User who retweet the tweet points to original user who tweet. From the formed graph, each node's PageRank is calculated as well as other node properties like centrality, degree, and followers and average time retweeted. The result shows that PageRank score of node is directly proportional to closeness centrality and in-degree of node. However, the ranking with PageRank, closeness centrality, and in-degree ranking yield different ranking result.**

***Keywords*—Twitter ranking; social network analysis; graph-based algorithm; PageRank; graph centrality**

I. Introduction

The growing number of internet users is followed by the growing number of social media users as a virtual world network that connects users through various social media platforms such as Twitter. Twitter is the widest micro-blogging site, as well as the vast social network and can be defined as first-hand amateur online news source. Twitter contains billions of users with their particular "tweets" globally, with each tweet has its own topics which can be retweeted by other user. The vast data produced from this platform engage the research about social network analysis.

The focus on social network analysis is how to measure relations and flows between person, organization, or community. These objects are defined as nodes in graph, meanwhile the relationships or flows between two objects are represented in edges [1]. Social media analysis allows one to get a figure of the position of the node in a social network, which is described as a social graph [2]. From Twitter, a vast user-network graph that accommodates the users as the nodes and the followed-following status of two users as the edges is can be built [3]. On the other hand, Twitter network can create topic-specific graph that encompass users who tweets the topic as the nodes and the retweeted-status as the edges.

In social structure, power is defined as fundamental property. Despite the uncertain of what power of social structure in social network is, it can be described in three aspects: degree —how many nodes ties with a node, closeness —length of paths from a nodes to others, and betwenness —lying between pair of nodes [1]. Degree of node in social network can be prescribed as node which has most influence in the network. In topic-specific graph, degree of nodes can be correlated as the issuer of a topic. The degree of each nodes can be computed directly with valency, or through a graph-based ranking.

One of the popular graph-based ranking is PageRank which rank a nodes of graph based the in-degree and out-degree of the nodes. PageRank determines the importance of a node within a graph, by computing the information on graph globally and recursively [4]. The original purpose of PageRank is to rank all web pages based on the interconnection around that page, aside from each content of the pages. Until present day, this algorithm is still used in Google Search to get a relevant result to the query [5]. PageRank also can be used in text mining problem [6],[7] as long as it could be represented as graph.

A method to obtain the issuer of determined topic from Twitter with PageRank is proposed for this research. The proposed method is expected to be able to build a social network graph and determine the issuer of the topic.

II. Related Works

Prior research has explored graph representation of Twitter. In the research done by Myers et al. [8] provides topological feature of Twitter graph based following-followed status by the users. This research conduct analysis about degree of the user as node. Although the ranking model of user is not examined, this research shows that degree of each user can be used to define behaviour of Twitter user. In another research [9], Bild et al. conduct analysis based on the tweet-retweeted relation of users and represent the relationship with graph. This research does not rank each user specifically, but it shows that Twitter representation in graph is not only using following-followed status, but also retweet-retweeted relation.

PageRank is common algorithm to solve graph-based ranking. PageRank is used in text mining problem and is called TextRank [6]. PageRank can be used for term extraction [10] and sentences extraction [11]. For keywords extraction, words are represented as undirected graph where each word defined in node and co-occurrences of two words in several window context work as edge . Meanwhile in sentence extraction, each sentence is represented in node and similarity between two sentences is defined as the edges. The result of each case is based on each rank of corresponding task. This study shows that PageRank can be modified and used in ranking problem, as long as the problem can be defined in graph.

Study of ranking for Twitter user is done in prior research [12]. This research propose a ranking method where it uses three aspects to rank influence level of user in specific hashtag: followers, retweets, and favorites. This method is called TRank and yet using graph-based ranking method. For a graph-based ranking, Kwat et al. [3] perform an analysis of Twitter users based on follower-following topology and follower-tweet relation, as well as analysis of trending topics. In analysis of Twitter users, this study proposes ranking of users in two approaches: by PageRanking each user based on follower-following status, and amount of retweet in certain tweets to determine the rank of users. In another research [13], PageRank is used to get influential Twitter user on specific topic. The proposed method is called TwitterRank. The similarity between these researches are that they crawl the user first, and then collect all the tweet from each corresponding user, as well as each follower and following of each user.

III. Methodology

This research aims to build unweighted directed graph based on specific topic or term on Twitter. Unlike the previous research [13], [12] where the used data is based on the user, this research uses tweets as the base data for this research. In brief, the overall research works like the Fig. 1.

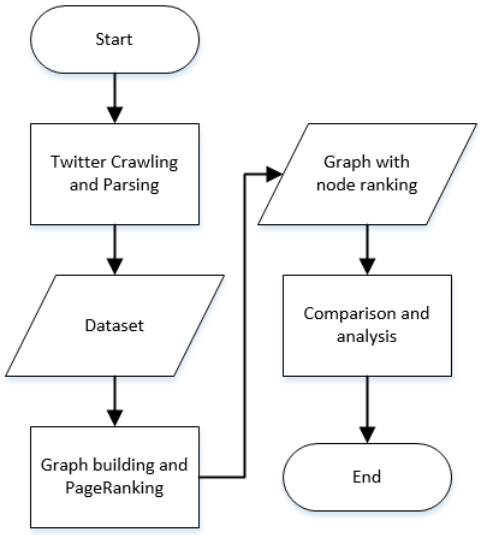

Fig. 1. Methodology of research.

A. Dataset

Data is obtained from Twitter with Twitter API. Due to limitation of request, only 100 tweets per one request is collected, and only 180 requests is executed per 15 minutes[1]. For this research the conducted search term is "Jokowi Capres". Unlike other research [3][13], the crawled tweets are based on the related tweet with the search term, not the users.

For each tweet, the tweet itself is scraped with the timestamp and user who tweets. If the tweet is retweeted from other user, the tweet is parsed to get the original user who tweet. Some examples for dataset is provided in Table I

[1]https://developer.twitter.com/en/docs/basics/rate-limiting.htm

TABLE I. Sample of dataset

tweet	"@VIVAcoid Dukungan kepala daerah itu ke pak jokowi itu sdh benar karna sampai saat ini dialah kepala negara yg"
timestamp	Wed Sep 19 05:53:15 +0000 2018
username	"asril_zn"
retweet_from	None
tweet	"RT @Jopiesays: Coba kita berandai-andai jika apa yg sudah direncanakan Pak @jokowi bisa berjalan sesuai RENCANA"
timestamp	Wed Sep 19 05:50:11 +0000 2018
username	"Sarah_Pndj"
retweet_from	"Jopiesays"

B. Tweets as Graph

The topic-specific tweets are represented into graph. The user who tweet pictured as nodes. If Table I represented in graph it will be shown as in Fig. 2. If the tweet is a retweeted tweet, the nodes points to other node, where the other node is the user who originally tweet. The formed graph is directed graph which point direct from retweeted user to original user.

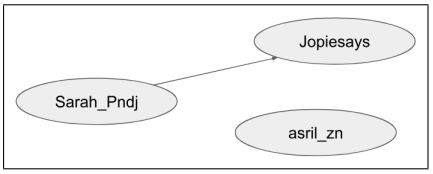

Fig. 2. Sample graph formed from Table I.

As graph representation, in-degree of each node is how many other users retweeted the tweet from the user, meanwhile out-degree of each node is how many the user retweet a tweet from other user.

C. Graph Building and PageRanking

Directed graph is built by following the example from Fig. 2. Supposedly the graph G has set of nodes N and set of edges E where E is subset of $N \times N$. For each node N_i, $In(N_i)$ is all nodes that point to N_i, meanwhile $Out(N_i)$ is all nodes pointed by N_i, so that $|In(N_i)|$ is in-degree of node N and $|Out(N_i)|$ is out-degree of node N. Equation 1 is used to get PageRank score of each node, where d is the damping factor which can be set between 0 and 1, and usually set into 0.85 [4].

$$S(N_i) = (1-d) + d * \sum_{j \in In(N_i))} \frac{1}{|Out(N_j)|} S(N_j) \quad (1)$$

This ranking method is done repeatedly, where the initiation score for each node is set as 1. After that node is ranked based on the PageRank score from Equation 1. The first rank of the ranking is considered as the issuer of the topic.

D. Comparing and Analyzing PageRank

The result of PageRank scoring is analyzed with other centrality methods. Each centrality that is used in the comparison are:

- Closeness centrality
- Betweenness centrality

All of the other centrality methods is computed in formed graph as well.

Closeness centrality is defined as reciprocal sum of shortest distance from one node to other nodes. Supposed in the graph there is node N with set of other nodes M and $N \notin M$, the closeness centrality $C_c(N)$ is formulated as Equation 2.

$$C_c(N) = \frac{n_N - 1}{\sum_{i=1}^{n_N - 1} d(N, M_i)} \tag{2}$$

where $d(N, M_i)$ is distance between node N to M_i and n_N is number of nodes in the graph that can reach N including the node itself. [14]. Higher number of closeness centrality of one node means more central that node is. For directed graph, $d(N, M_i)$ is not zero if M_i is predecessor of N, which means M_i points to N directly or transitively.

The closeness centrality from Equation 2 is often normalized, by multiplying to the ratio of reachable nodes and all of the nodes [15]. The normalized formula for Equation 2 is written in Equation 3 where n denotes number of all nodes in the graph.

$$C_c(N) = \frac{n_N - 1}{n - 1} \frac{n_N - 1}{\sum_{i=1}^{n_N - 1} d(N, M_i)} \tag{3}$$

The betweenness centrality of the node is described as numbers of nodes N is passed between shortest path of other nodes L and M divided by all of numbers of shortest path between L and M. Formally, betweenness centrality is written as Equation 4.

$$C_b(N) = \sum_{L \neq M \neq N \in V} \frac{\sigma_{LM}(N)}{\sigma_{LM}} \tag{4}$$

where V is all nodes in graph, σ_{LM} is numbers of shortest paths of L to M and $\sigma_{LM}(N)$ is numbers of shortest paths of L to M that passed N [16].

PageRank result of the formed graph is also analyzed with number of followers and in-degree of each nodes. This research also compute how fast the tweet from one node is retweeted by other node by calculating average of time difference between original tweet and retweeted tweet. This calculation is called Average Retweet Time (ART). Not all nodes have ART, because not all nodes are retweeted by other nodes. Only node with in-degree more than zero has ART.

All of the numbers, including centrality from Equation 2 and 4 as well as ART are called node properties. The correlation between PageRank score with each node properties is calculated with Equation 5

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{5}$$

where r is the correlation coefficient, $x_i y_i$ is pair of two data indexed by i, n is the size of the data and $\bar{x}$ is mean of data x.

IV. Result

The graph is built with NetworkX library from Python, and is visualized with Gephi. The graph visualized in Fig. 3. Only non-isolated nodes is visualized due to the vast of the graph. From the Fig. 3, it can be seen that most of the nodes are not visible, but the nodes with high degree can be seen connected to many other nodes.

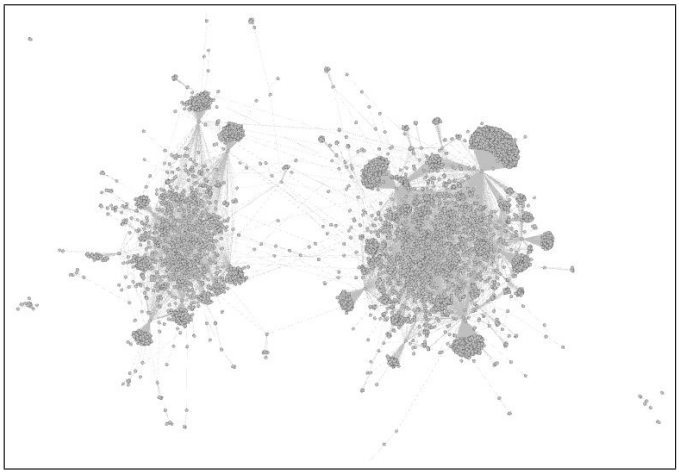

Fig. 3. Formed graph.

Using Equation 1, the PageRank of each node is computed. The top twenty result of each ranking is provided. Table II provide the result of node properties such as closeness centrality and betweenness centrality, and other result such as in-degree, number of followers and ART in Table III.

TABLE II. Top 20 user based on PageRank and each centrality

Rank	Username	PageRank	Closeness	Betweenness
#1	KaosPerjuangan	0.069	0.225	0
#2	PollingLagi	0.032	0.108	0
#3	asep_maoshul	0.031	0.115	0
#4	DeanaZuliana	0.019	0.047	0.0000994
#5	P3nj3l4j4h	0.017	0.051	0.000112
#6	kurawa	0.016	0.053	0
#7	SumardiAcehID	0.015	0.063	0
#8	nissa080789	0.015	0.010	0
#9	ASapardan	0.014	0.052	0.0000756
#10	purwo82092883	0.013	0.041	0.0000348
#11	RizmaWidiono	0.013	0.045	0.00011
#12	Dahnilanzar	0.011	0.043	0
#13	IreneViena	0.011	0.046	0
#14	MSAokepunya	0.011	0.040	0
#15	ruhutsitompul	0.011	0.039	0
#16	Mbah_lemper	0.010	0.032	0.0000719
#17	V_Stone_Kardol	0.010	0.039	0.0000163
#18	RustamIbrahim	0.010	0.031	0
#19	JajangRidwan19	0.010	0.044	0.0000598
#20	TheArieAir	0.010	0.043	0.0000821

The correlation coefficient between PageRank and other node properties is calculated with Equation 5. The correlation coefficient between each properties is showed in Table IV. Note that in Table IV there is no ART as the property correlation, since not all nodes have ART. If all the nodes are filtered so that only nodes with ART is used, the correlation of PageRank and each properties is showed in Table V.

V. Discussion

Both Tables IV and V show that PageRank score is proportional to the closeness centrality and in-degree of the node and has no betweenness centrality and number of followers, as well as ART. However, approach by PageRank yield different ranking than approach by closeness centrality or in-degree.

TABLE III. TOP 20 USER BASED ON PAGERANK AND EACH PROPERTIES

Rank	Username	PageRank	In-degree	Followers	ART (seconds)
#1	KaosPerjuangan	0.069	2488	2343	103756
#2	PollingLagi	0.032	1181	28466	48286
#3	asep_maoshul	0.031	1273	1380	132251
#4	DeanaZuliana	0.019	113	3217	32446
#5	P3nj3l4j4h	0.017	245	26529	22823
#6	kurawa	0.016	507	309578	28921
#7	SumardiAcehID	0.015	693	4532	134108
#8	nissa080789	0.015	114	1264	37608
#9	ASapardan	0.014	272	9267	31461
#10	purwo82092883	0.013	127	1952	20609
#11	RizmaWidiono	0.013	278	27589	38556
#12	Dahnilanzar	0.011	475	92323	23035
#13	IreneViena	0.011	482	48039	38797
#14	MSAokepunya	0.011	457	11499	27570
#15	ruhutsitompul	0.011	434	1960566	24043
#16	Mbah_lemper	0.010	355	4932	46369
#17	V_Stone_Kardol	0.010	23	956	17166
#18	RustamIbrahim	0.010	343	23184	6710
#19	JajangRidwan19	0.010	205	12155	17898
#20	TheArieAir	0.010	96	56364	22228

TABLE IV. CORRELATION COEFFICIENT BETWEEN PAGERANK AND OTHER NODE PROPERTIES

Correlation	PageRank	C_c	C_b	In-Degree	Followers
PageRank	1	0.946729	0.391378	0.945641	0.02292
C_c	0.946729	1	0.401989	0.89803	0.023223
C_b	0.391378	0.401989	1	0.167817	0.002784
In-Degree	0.945641	0.89803	0.167817	1	0.025589
Followers	0.02292	0.023223	0.002784	0.025589	1

Table VI show different result of ranking from those three approaches. From Table VI, the first rank user from each ranking approach is same, meanwhile the other rank are different.

The different rank between PageRank and closeness centrality is because closeness centrality only determine the rank based on the one node, for example N, and calculate how many other nodes that can reach node N directly or transitively, without estimate that other nodes not only can reach node N. PageRank considers that other nodes is not only pointing to N. Those other nodes that points to many nodes contributes lower PageRank score to node N than nodes that only points to N, meanwhile in closeness centrality it is considered same.

PageRank and in-degree yield different result, despite the high correlation between them. It is different because scoring in one node—supposed it is called N, because the properties of $In(N)$. The in-degree ranking only use $|In(N)|$ as consideration for N scoring, while PageRank also deal with $Out(N_i)$ where $N_i \in In(N)$. It is similar to difference between PageRank and closeness centrality.

Fig. 4 describes as an illustration how PageRank, closeness centrality and in-degree ranking generate different result. In a glimpse, it can be concluded that node A, B, and C have same

TABLE V. CORRELATION COEFFICIENT BETWEEN PAGERANK AND OTHER NODE PROPERTIES WITH ART

Correlation	PageRank	C_c	C_b	In-Degree	Followers	ART
PageRank	1	0.944	0.358	0.943	0.025	0.226
C_c	0.944	1	0.364	0.893	0.021	0.202
C_b	0.358	0.364	1	0.127	-0.013	0.024
In-Degree	0.943	0.893	0.127	1	0.035	0.255
Followers	0.025	0.021	-0.013	0.035	1	-0.020
ART	0.226	0.202	0.024	0.255	-0.020	1

TABLE VI. RANKING COMPARISON USING PAGERANK AND IN-DEGREE

No.	PageRank	Closeness Centrality	In Degree
1	KaosPerjuangan	KaosPerjuangan	KaosPerjuangan
2	PollingLagi	asep_maoshul	asep_maoshul
3	asep_maoshul	PollingLagi	PollingLagi
4	DeanaZuliana	SumardiAcehID	SumardiAcehID
5	P3nj3l4j4h	kurawa	kurawa
6	kurawa	ASapardan	IreneViena
7	SumardiAcehID	P3nj3l4j4h	Dahnilanzar
8	nissa080789	DeanaZuliana	MSAokepunya
9	ASapardan	IreneViena	narpatisuta
10	purwo82092883	RizmaWidiono	ruhutsitompul
11	RizmaWidiono	JajangRidwan19	Mbah_lemper
12	Dahnilanzar	JKFC23456789	AntoniRaja
13	IreneViena	TheArieAir	RustamIbrahim
14	MSAokepunya	Dahnilanzar	RajaPurwa
15	ruhutsitompul	purwo82092883	permadiaktivis
16	Mbah_lemper	narpatisuta	arch_v3nture
17	V_Stone_Kardol	MSAokepunya	RizmaWidiono
18	RustamIbrahim	ruhutsitompul	ASapardan
19	JajangRidwan19	V_Stone_Kardol	MCAOps
20	TheArieAir	RockyGaring	P3nj3l4j4h

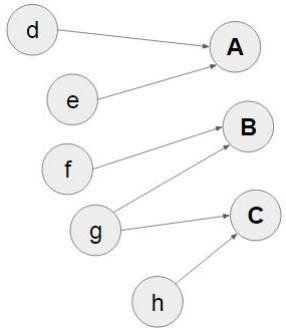

Fig. 4. Sample graph for comparison between PageRank, closeness centrality, and in-degree.

in-degree, and Equation 3 yield that node A, B, and C has same value of closeness centrality because each of those nodes are equally pointed by 2 other nodes. However from Equation 1, node A has higher PageRank score than B and C, where B and C has same score. This is because node g connects to both B and C so that g has higher out-degree. In Twitter, it is considered that node g is the user who retweet an issue from more than one issuer or user. User g is considered to split its focus of the issue to both users B and C. This make both of the user B and C are considered to have small tendency for the issue, which is shown from their corresponding PageRank. User is treated as more influential issuer if other users retweet to that user only.

Closeness centrality and in-degree scoring has high correlation by the Table IV. Even though yield the similar result and has similar difference with PageRank, ranking by closeness centrality and in-degree have different approach. Closeness centrality considers all other nodes that point directly and transitively, meanwhile in-degree only consider other nodes that point directly.

PageRank and betweenness centrality have low correlation coefficient based on Table IV that signifies that there are no correlation between those two. It is because PageRank focus on nodes that is pointed by another nodes directly or transitively, while betweenness centrality focus on nodes in between that deliver one node to other node. In Twitter,

node with betweenness centrality can be interpreted as user U who retweet a tweet from user V, and other user W retweet from user U. PageRank takes user V as most important user because the issue is began from that user, whereas betweenness centrality takes user U as most important user because user U is the one who spread the issue from V to W. In another case, where user U is retweeted by user V and W, PageRank scores user U as the highest rank. However, either user U, V, and W have the same score in betweenness centrality. It is because in directed graph, user U does not bridge between user V and W so that user U has no significant betweenness centrality score.

PageRank score and number of follower's correlation coefficient is nearly 0. It signify that there are no linear correlation between PageRank score and follower of users. However sample result in Table III shows that higher PageRank requires more followers. It conclude that number of followers of user does not determine how influential that user to specific topic, but influential users of topic usually have high number of follower.

For correlation between PageRank and ART, Table V show no linear correlation between PageRank score and ART, as well as other node properties. However, sample result from Table III shows that most of users with high score of PageRank have ART more than 4 hours.

VI. Conclusion

Twitter as the biggest micro-blogging site contains billions of information in a form of tweet and each tweet has its own topic. Social network analysis can be used to get the network of a specific topic and get the possible issuer of the topic.

This research has conducted method to get the issuer of topic in Twitter using PageRank and analyze with other centrality and properties. Total of 18000 tweet from Twitter are scrapped, with its corresponding user and origin user who tweet. Each user is represented in node, which is then built into directed graph. PageRank scoring is applied to the graph, which gives each node a score for ranking, as well as other centrality and properties.

The ranking result from PageRank is quite different with ranking that use closeness centrality and in-degree of nodes as the ranking key, even though they have high correlation coefficient that signify linear correlation. PageRank take that user is more influential issuer if other users retweet to that user only. In another comparison, PageRank yield different result with betweenness centrality, because PageRank focus on which node that is pointed by other nodes, not focusing on node that bridges other nodes. Meanwhile, the number of followers and average retweet time do not determine how influential a user can in specific topic, but highly influential user of topic is usually followed with high numbers of followers.

Even though this research could not evaluate which ranking method is better, this research shows the method to get the topic issuer from Twitter. In future study, it is suggested to increase the data into million as well to try other graph-based algorithm other than PageRank and its modification derivatives with more analysis with other properties and centrality methods.

Acknowledgment

The authors would like to thank to Directorate of Research and Faculty of Mathematics and Natural Science Universitas Gadjah Mada for supporting the research.

References

[1] M. Jamali and H. Abolhassani, "Different aspects of social network analysis," in *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)*, Dec 2006, pp. 66–72.

[2] R. Alhajj and J. Rokne, *Encyclopedia of Social Network Analysis and Mining*. Springer Publishing Company, Incorporated, 2014.

[3] H. Kwak, C. Lee, H. Park, and S. Moon, "What is twitter, a social network or a news media?" in *Proceedings of the 19th international conference on World wide web*. AcM, 2010, pp. 591–600.

[4] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[5] J. Le and S. Kumar, "Pagerank–the elite algorithm: A research analysis of google's pagerank algorithm on controversial search terms and bias in search," 2017.

[6] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004.

[7] S. Krek, C. Laskowski, and M. Robnik-Šikonja, "From translation equivalents to synonyms: Creation of a slovene thesaurus using word co-occurrence network analysis," in *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference*. Lexical Computing, 2017, pp. 93–109.

[8] S. A. Myers, A. Sharma, P. Gupta, and J. Lin, "Information network or social network?: the structure of the twitter follow graph," in *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 2014, pp. 493–498.

[9] D. R. Bild, Y. Liu, R. P. Dick, Z. M. Mao, and D. S. Wallach, "Aggregate characterization of user behavior in twitter and analysis of the retweet graph," *ACM Transactions on Internet Technology (TOIT)*, vol. 15, no. 1, p. 4, 2015.

[10] Z. Zhang, J. Petrak, and D. Maynard, "Adapted textrank for term extraction: A generic method of improving automatic term extraction algorithms," *Procedia Computer Science*, vol. 137, pp. 102–108, 2018.

[11] F. Barrios, F. López, L. Argerich, and R. Wachenchauzer, "Variations of the similarity function of textrank for automated summarization," *arXiv preprint arXiv:1602.03606*, 2016.

[12] M. Montangero and M. Furini, "Trank: Ranking twitter users according to specific topics," in *Consumer Communications and Networking Conference (CCNC), 2015 12th Annual IEEE*. IEEE, 2015, pp. 767–772.

[13] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 261–270.

[14] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social networks*, vol. 1, no. 3, pp. 215–239, 1978.

[15] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge university press, 1994, vol. 8.

[16] U. Brandes, "On variants of shortest-path betweenness centrality and their generic computation," *Social Networks*, vol. 30, no. 2, pp. 136–145, 2008.