Social network analysis using k-Path centrality method

Natya Taniarza¹, Adiwijaya² and Warih Maharani³
School of Computing, Telkom University, Bandung, 40257, Indonesia
¹natyataniarza@gmail.com, ²adiwijaya@telkomuniversity.ac.id, ³wmaharani@gmail.com

Abstract. k-Path centrality is deemed as one of the effective methods to be applied in centrality measurement in which the influential node is estimated as the node that is being passed by information path frequently. Regarding this, k-Path centrality has been employed in the analysis of this paper specifically by adapting random-algorithm approach in order to: (1) determine the influential user’s ranking in a social media Twitter; and (2) ascertain the influence of parameter α in the numeration of k-Path centrality. According to the analysis, the findings showed that the method of k-Path centrality with random-algorithm approach can be used to determine user’s ranking which influences in the dissemination of information in Twitter. Furthermore, the findings also showed that parameter α influenced the duration and the ranking results: the less the α value, the longer the duration, yet the ranking results were more stable.

1. Introduction
Social networks can be defined as finite sets of users and relationships between other users [1]. Users and relationships can be measured and mapped in graphs by interaction pattern analysis or defined as Social Network Analysis (SNA) where the user is represented as a node and the relation is represented as edge [2]. One of the most commonly used SNA applications is centrality measurement [3].

Centrality measurement is used to measure the ability of a node in flowing information. In general, the centrality measurement consists of 3 points of view: Degree centrality, Betweenness centrality and Closeness centrality [4]. By using the Betweenness centrality method, node most often located on the shortest path is the most influential node in the dissemination of information. While shortest paths are not always an optional path of information [5], enabling influential nodes in the flow of information is a node that is not on the shortest path.

The problem can be solved by using k-Path centrality method with random algorithm approach, where in determining path or path of a node does not use shortest path but by using random algorithm, so it will look for possible combinations of path of information or path. Therefore, in this paper use k-Path centrality method to know the ranking of users who have influence on Twitter network.

The remainder of this paper is structured as follows: section 2 briefly presents the methodology, section 3 is system design and implementation, section 4 reports the experiment and results from the case study conducted, and section 5 provides conclusion and suggestion for further research.
2. Methodology

2.1. Social Network Analysis
Social Network Analysis (SNA) is a mathematical method that allows to map and measure interaction paths formed between nodes on a network. SNA aims to understand how groups of individuals relate and how the consequences, and understand how they behave [3]. Scott [6] defines SNA as a set of methods used to investigate relationships in the data structure. So with SNA we can analyze the interactions within a group by looking at the social actors (nodes) with the relationships between them. Otte and Rousseau [7] shows that SNA can be used for implicit information retrieval purposes, including interaction and friendship relationships between users.

2.2. Centrality Measurement
The measurement of centrality is one example of applying SNA. Used to measure the ability of each node in streaming information. To date, many development methods can be used to measure centrality. Each method has its own advantages and disadvantages. Selection of methods can be tailored to the needs. Of the various methods, the most widely used according to Otte and Rousseau [7] are Degree centrality, Betweenness centrality and Closeness centrality.

2.3. k-Path Centrality
k-Path centrality ($C_k$) is the development of Betweenness centrality, where the same general concept is to emphasize the relationship of a node to another node. But on the calculation of this centrality without specifying the shortest paths as Betweenness does, according to Tharaka [8] shortest paths are not always the path through which information passes, in other words it allows nodes that have a high centrality value instead of nodes that are in the path of shortest paths. Tharaka also revealed that nodes that have a high $C_k$ then have a high $C_B$ as well [5,8]. Centrality calculation formula by using k-Path centrality method on a node $v$ for weightless graph that is [5,8]:

$$C_k(v) = \sum_{s \neq v} \sum_{1 \leq \ell \leq k} \frac{p_d(s;\ell)}{p(s;\ell)}$$

(1)

Where $p_d(s;\ell)$ is the number of paths from node $s$ to another node along $k$ that passes through the node $v$ as much as $\ell$ and $p(s;\ell)$ is the number of paths from node $s$ to another node along $k$ as much as $\ell$.

2.4. Random Algorithm Approach
The random algorithm approach is used to estimate the k-Path centrality value of each node with the concept of retrieving the nodes randomly using a uniform distribution. Because using the uniform distribution then the chances of each node to be drawn together. Calculation of k-Path centrality value in random algorithm approach [5,8] becomes:

$$C_k(v) = B(v) \frac{k n}{\ell}$$

(2)

$$\ell = 2 k^2 n^{1-2\alpha} \ln n$$

(3)

$$k = \ln (n + m)$$

(4)

Where $B(v)$ is the weight value of node $v$, the weight here means the number of nodes $v$ taken. $\ell$ is number of paths generated, $k$ is positive integer number which is the number of nodes in a single path. $n$ is number of nodes while $m$ is number of edges and $\alpha$ is real number ($0, \frac{1}{2}$).

3. System Design and Implementation
In this section, the system design of the current research will be elaborated. The steps of the process are as shown in Figure 1.

3.1. Twitter Dataset
Twitter data is a data set derived from social media Twitter taken using NodeXL software. The data obtained is a table of relations consisting of user names and relationships between users. Data set used in this experiment comes from social media Twitter taken from Twitter @natyataniarzaa users. Data in the form of friendship relation between user and other user. The relations used are follower and followed / following relations. From the retrieval of data obtained by the user as much as 377 and 17006 relation.

3.2. Preprocessing
Preprocessing is done to get the matrix ready to be input to the system, this process is done in NodeXL software. The detail steps in preprocessing are as shown in Figure 2. The first is the process of merge duplicate on the relation table, this is done because on Twitter data allows the existence of the same data or duplicate. For example when user A performs the following to user B with user A being a user follower B, this data is what will be merged. Results table of relationships formed on Twitter user @natyataniarzaa can be seen in Figure 3.

The non-duplicate relation table is then represented in the form of \( n \times n \) matrix, where \( n \) is the number of users retrieved. The example of \( n \times n \) matrix results that formed on Twitter @natyataniarzaa user can be seen in Figure 4.
3.3. \textit{k-Path Centrality Measurement}

To facilitate the calculation of \(k\)-Path centrality is done in 3 stages, namely initialization, path determination and accumulation (see Figure 5).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure5.png}
\caption{\textit{k-Path Centrality Measurement Process.}}
\end{figure}

In the first process is the initialization of the required parameters values when the path determination and accumulation. The second process is the process of determining the path, in this process produced a combination of simple paths taken at random. The path determination process aims to give the weight of each node that is picked up or missed. Each retrieval of values and nodes is taken randomly with a uniform distribution approach, where each value or node has the same chance of being retrieved. The last process is the process of accumulation, this process aims to get the value of the centrality of each node based on the weight obtained from the previous process included in equation 1.

4. \textbf{Experimental Result}

The result of the experiment is intended to implement and analyze the effect of parameter \(\alpha\) on processing time and ranking result. This experiment aims to know the rank of the influential user in spreading information on Twitter. The detail of how the procedure is conducted will be detailed in the following subsections.

4.1. \textbf{Experimental Setup}

Based on the objectives to be achieved, tested the value of parameter \(\alpha\). In the process of testing is done changes the value of \(\alpha\) as much as 49 times that starts from 0.01 and increased 0.01 each test to 0.49.

The dataset used in this experiment comes from Twitter social media taken from Twitter \@natyataniarzaa users with the help of NodeXL software. Data in the form of friendship relation between user and other user. The relations used are follower and followed/following relations. From the retrieval of data obtained by the user as much as 377 and 17006 relations.

4.2. \textit{Analysis Parameter \(\alpha\) on The Processing Time}

Testing conducted 49 times on Twitter data, where each test is done as much as 30 times then generated the average time in seconds. In order to more easily understand the effect of \(\alpha\) then given the visualization in graphical form as follows:
Based on Figure 6 can be seen that the graph is a non-linear graphic where the smaller the value of $\alpha$ used then the time required to obtain the value of centrality longer. This can be due to the smaller value of $\alpha$ used, the more combinations of paths are generated. To find the combination of these paths depends on the duration of the calculation of centrality.

4.3. Analysis Parameter $\alpha$ on Ranking Result

In this experiment is done 49 times where each test is done as much as 10 times. In this paper is shown 3 tables of test results that can represent the overall test.

Based on Table 1 it can be seen that the 'hmifitt' and 'johan_jojo' nodes are exchanged positions when the 8th and 10th tests can be caused by two things, first because the difference in the centrality value of the two nodes is small and the second the number of each node picked up may change as a result of random path determination to allow the number of 'johan_jojo' nodes to be retrieved in the 8th and 10th tests more than the number of captured 'hmifitt' nodes.

| rank | 1st test | 2nd test | 3rd test | 4th test | 5th test | 6th test | 7th test | 8th test | 9th test | 10th test |
|------|----------|----------|----------|----------|----------|----------|----------|----------|----------|-----------|
| 1    | 1        | 1        | 1        | 1        | 1        | 1        | 1        | 1        | 1        | 1         |
| 2    | 2        | 2        | 2        | 2        | 2        | 2        | 2        | 2        | 2        | 2         |
| 3    | 3        | 3        | 3        | 3        | 3        | 3        | 3        | 3        | 3        | 3         |
| 4    | 4        | 4        | 4        | 4        | 4        | 4        | 4        | 4        | 4        | 4         |
| 5    | 5        | 5        | 5        | 5        | 5        | 5        | 5        | 5        | 5        | 5         |
| 6    | 6        | 6        | 6        | 6        | 6        | 6        | 6        | 6        | 6        | 6         |
| 7    | 7        | 7        | 7        | 7        | 7        | 7        | 7        | 7        | 7        | 7         |
| 8    | 8        | 8        | 8        | 8        | 8        | 8        | 8        | 8        | 8        | 8         |
| 9    | 9        | 9        | 9        | 9        | 9        | 9        | 9        | 9        | 9        | 9         |
| 10   | 10       | 10       | 10       | 10       | 10       | 10       | 10       | 10       | 10       | 10        |

Based on Table 2 it can be seen that the 'edwinugroho' and 'permib' nodes can fit into the top 10, this can be due to the centrality of a node varying each test. The value of this centrality can be variable because in determining the combination path is done randomly, thus allowing the path in first test will be different from the second test.

| rank | 1st test | 2nd test | 3rd test | 4th test | 5th test | 6th test | 7th test | 8th test | 9th test | 10th test |
|------|----------|----------|----------|----------|----------|----------|----------|----------|----------|-----------|
| 1    | 1        | 1        | 1        | 1        | 1        | 1        | 1        | 1        | 1        | 1         |
| 2    | 2        | 2        | 2        | 2        | 2        | 2        | 2        | 2        | 2        | 2         |
| 3    | 3        | 3        | 3        | 3        | 3        | 3        | 3        | 3        | 3        | 3         |
| 4    | 4        | 4        | 4        | 4        | 4        | 4        | 4        | 4        | 4        | 4         |
| 5    | 5        | 5        | 5        | 5        | 5        | 5        | 5        | 5        | 5        | 5         |
| 6    | 6        | 6        | 6        | 6        | 6        | 6        | 6        | 6        | 6        | 6         |
| 7    | 7        | 7        | 7        | 7        | 7        | 7        | 7        | 7        | 7        | 7         |
| 8    | 8        | 8        | 8        | 8        | 8        | 8        | 8        | 8        | 8        | 8         |
| 9    | 9        | 9        | 9        | 9        | 9        | 9        | 9        | 9        | 9        | 9         |
| 10   | 10       | 10       | 10       | 10       | 10       | 10       | 10       | 10       | 10       | 10        |

1. natyataniarzaa 2. infobdg 3. infoitt 4. detikcom 5. sidiik 6. bem_itt 7. Kuliner_Bandung 8. hmifitt 9. johan_jojo 10. Adevaoktoveri
Based on Table 3 it can be seen that only 'natyataniarzaa' and 'infobdg' nodes that do not change rank, this can happen for two reasons. First because the centrality value of the two nodes is far adrift from the other nodes, so that no node can defeat the centrality of the two nodes, and the second because all nodes have relation with the natyataniarza node so nat 'natyataniarzaa' is a strong link in the deployment information, this is what causes node 'natyataniarzaa' always in the top position.

Based on Table 3 it can be seen that only 'natyataniarzaa' and 'infobdg' nodes that do not change rank, this can happen for two reasons. First because the centrality value of the two nodes is far adrift from the other nodes, so that no node can defeat the centrality of the two nodes, and the second because all nodes have relation with the natyataniarza node so nat 'natyataniarzaa' is a strong link in the deployment information, this is what causes node 'natyataniarzaa' always in the top position.

**Table 3. Rank Result Where α = 0.4.**

| Rank | Node            | 1st test | 2nd test | 3rd test | 4th test | 5th test | 6th test | 7th test | 8th test | 9th test | 10th test |
|------|-----------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|-----------|
| 1    | natyataniarzaa  | 1        | 1        | 1        | 1        | 1        | 1        | 1        | 1        | 1        | 1         |
| 2    | infobdg         | 2        | 2        | 2        | 2        | 2        | 2        | 2        | 2        | 2        | 2         |
| 3    | infoitt         | 3        | 3        | 3        | 3        | 3        | 3        | 3        | 3        | 3        | 3         |
| 4    | detikcom        | 4        | 4        | 4        | 4        | 4        | 4        | 4        | 4        | 4        | 4         |
| 5    | sidiik          | 5        | 5        | 5        | 5        | 5        | 5        | 5        | 5        | 5        | 5         |
| 6    | bem_itt         | 6        | 6        | 6        | 6        | 6        | 6        | 6        | 6        | 6        | 6         |
| 7    | Kuliner_Bandung | 7        | 7        | 7        | 7        | 7        | 7        | 7        | 7        | 7        | 7         |
| 8    | hmifitt         | 8        | 8        | 8        | 8        | 8        | 8        | 8        | 8        | 8        | 8         |
| 9    | johan_jojo      | 9        | 9        | 9        | 9        | 9        | 9        | 9        | 9        | 9        | 9         |
| 10   | Adevaoktori     | 10       | 10       | 10       | 10       | 10       | 10       | 10       | 10       | 10       | 10        |

Based on Table 1, 2, and 3 it can be concluded that the value of α affects the ranking results, where the greater α used the more changes in the ranking results. Changes in ranking results because of the greater the value of α then the combination of paths taken less, so the combination of each test path will be much different than the combination of paths taken more.

**4.4. Result**

From the previous analysis based on two angle of view that is time and result of ranking in taking value of α used, hence at this test value of α used is 0.2, where time used not too long and rank yielded stable enough. On this Twitter data is tested 30 times then averaged to get top 10 users who have the highest centrality value. Here is displayed the value of centrality, the number of follower relations and relations following each node:

**Table 4. Top 10 Rank Result Twitter User**

| Rank | Node      | Number of follower relation | Number of following relation | Centrality value |
|------|-----------|-----------------------------|-------------------------------|------------------|
| 1    | natyataniarzaa | 232                          | 320                           | 1788.55          |
| 2    | infobdg   | 252                          | 41                            | 293.49           |
| 3    | infoitt   | 232                          | 1                             | 231.65           |
| 4    | detikcom  | 151                          | 0                             | 191.17           |
| 5    | sidiik    | 176                          | 192                           | 173.41           |
| 6    | bem_itt   | 153                          | 1                             | 158.18           |
Based on Table 4 it can be seen that the number of relations following a node has no effect on the ranking result of the node. While the number of follower relationships is quite influential on the node ranking results, but that does not mean that the node has the lowest number of follower relation ranked the lowest as well. For example, the 'Kuliner_Bandung' node has the lowest follower relation among the other 9 nodes, but the 'Kuliner_Bandung' node is not in the lowest rank. This is because the position of node 'Kuliner_Bandung' is a strong link in the dissemination of information in the network. The strong connector here means it lies between a set of nodes. To compare the location of the 'Kuliner_Bandung' node can be seen in Figure 7.

![Figure 7. Follower Relations Formed on Node ‘Kuliner_Bandung’](image)

5. Conclusion
From the results of experiments conducted found that in the calculation of centrality is influenced by the value of $\alpha$, the smaller the value of $\alpha$ used then takes longer time calculation, but the ranking results tend to be more stable. The best $\alpha$ value parameter in this paper is 0.2, where the time required is not too long and the ranking result is quite stable. In addition to the number of followers, the ranking results are influenced by the user's position. The stronger the user as a liaison of a group of users the more influential the user is in the dissemination of information.

For further research can perform an accuracy analysis of k-Path centrality method, for example by kuisoner or survey. It can also implement the weighting effect on the k-Path centrality method for the weighted graph, by adding the mention and reply relations. Analyzing the influence of data characteristics, for example with the diversity of data amount or depth of graph used.

References
[1] R. Rabade, N. Mishra, S. Sharma. (2014). “Survey of Influential User Identification Techniques in Online Social Networks”. Rajiv Gandhi Proudyogiki Vishwavidyalaya: School of Information Technology.
[2] Hopkins, Adelaide. (2010). *Graph Theory, Social Networks and Counter Terrorism*. University of Massachusetts: Department of Mathematics.
[3] K. Hildrun and K. Theo. (2010). “A New Centrality Coefficient for Social network Analysis Applicable to Bibliometric and Webometric Data”. Humboldt-University Berlin: Department of Library and Information Science.
[4] E. Santos, E. D. Sotelino, Y. Cao, E. Brown, E. Santos. (2007). “Effective and Efficient Methodologies for Social Network Analysis”.

7
[5] A. Tharaka, T. Rahul, K. Nicolas, S. Ramanuju, I. Andrian. (2010). “K-Path Centrality: A New Centrality Measure in Social Network”. University of South Florida.

[6] Scoot, John. (2011). “Social network Analysis Theory and Application”.

[7] E. Otte and R. Rosseau. (2002). “Social Network Analysis: a Powerful Strategy, also for the Information Science”.

[8] AlohaKoo, Tharaka. (2010). “Path centrality: A new centrality measure in networks”. University of South Florida.

[9] Maharani, W., Adiwijaya and Gozali, A.A., 2014, October. Degree centrality and eigenvector centrality in twitter. In Telecommunication Systems Services and Applications (TSSA), 2014 8th International Conference on (pp. 1-5). IEEE.

[10] Rachman, Z.A. Maharani, W. and Adiwijaya, 2013, March. The analysis and implementation of degree centrality in weighted graph in Social Network Analysis. In Information and Communication Technology (ICoICT), 2013 International Conference of (pp. 72-76). IEEE.