Probabilistic Partnership Index (PPI) in Social Network Analysis using Kretschmer Approach

Nisa Sharafina CSF[1], W. Maharani[2] , Adiwijaya[3], N. Taniarza[4]

School of Computing
Telkom University
Bandung 40257, Indonesia
[1]sharafina.nisa@gmail.com , [2]wmaharani@gmail.com , [3]adiwijaya@telkomuniversity.ac.id , [4]natyataniarza@gmail.com

***Abstract*— Nowadays, Twitter has become an effective media to communicate as the increasing number of its user. The interaction or relation formed in Twitter could be represented into a graph and calculated using centrality measurement method. Centrality measurement can be used as parameter to determine the popularity or leverage level of an actor towards other actor. The value of centrality measurement is a weighted graph. To maximalize the result, every relation in a graph will be added value from Probabilistic Partnership Index (PPI) method calculation. Furtherly, the analysis and implementation with degree centrality are executed with Kretschmer method using the value from PPI measurement. From the conducted experimental process value, we observed that PPI and Kretschmer can be used as one of the centrality method to determine the leverage level and popular actor of an environment in Twitter.**

Keywords*— *Probabilistic Partnership Index (PPI); social network; degree centrality; Kretschmer;

I. Introduction

Recent years, the use of social media in the internet has increased incisively as the increase number of the users. Twitter is one of most popular social media nowadays. Based on the survey held by a third party Twitter analysis company, Twopcharts.com, user of Twitter had exceeded 1 billion users (this data taken on January 19th 2013). Within this condition, there are many relations formed by users. The relation between actors in Twitter can be used to represent its interaction. Then, this interaction can be measured and tested to perform its analysis or furtherly known as social network analysis (SNA) [6] [11].

Centrality measurement is one of the SNA measurement methods which can help to measure and determine the importance of relations between nodes in the network [3]. Centrality measurement consists of three different centrality measures, which are degree centrality, betweness centrality, and closeness centrality. But, this paper will only focus on degree centrality. Degree centrality is a method for measuring the number of ties/links a node has [10]. For more specific, degree centrality measure which will be used is degree centrality method introduced by Kretschmer. Kretschmer's degree centrality method is new complex measure of degree centrality including weighted ties. To differentiate, the previous degree centrality method was using unweighted ties.

On bibliometrics or webometrics analysis, there are four methods, known as collaborative linkage indexes, which can be used to measure the strength of nodes. The collaborative linkage indexes are Jaccard Index, Salton-Ochiai Indexes, Probabilistic Afinity Index (PAI), and Probabilistic Partnership Index (PPI). But, the objective in this paper is to analyze the degree centrality of social network Twitter using Probabilistic Partnership Index (PPI) indicator. Based on previous experiment, PPI can be used for measuring scientific lingkages between two countries, Japan and France. From conducted experiment, the result showed that PPI is useful in examining individual networks within complex exchanges [12]. Hence, using PPI as well as Kretschmer's method can help the analysis of social network to determine the popularity level of an actor in a community.

II. METHODOLOGY

A. Social Network Analysis

Social Network Analysis (SNA) is defined as the mapping and measuring of relationships and flows between people, groups, organizations, and other connected information by Krebs [11]. While L.C. Freeman said that SNA is the techniques focusing on uncovered patterns of people's interaction [6]. Scott also has another definition for SNA that is a set of methods for the investigation of relational aspects of social structures [15]. Of those opinions, it can be concluded that SNA is a method that can be used to analyze the social interactions within a group of people by looking at the behavior of those people while interact each other. The behavior between individuals can be seen not only in real world but also from the social network, like Twitter. It can be determined its structures and also patterns of the interactions.

B. Probabilistic Partnership Index (PPI) Measurement

Probabilistic Partnership Index (PPI) measures the scientific linkages which employ Monte-Carlo method for evaluating the expected values and standard deviations of the number of nodes of all patterns of cooperation [12]. PPI is formulated as follows:

$$PPI = (n_{ij} - \hat{E}_a[n_{ij}])/\hat{\sigma} \quad (1)$$

Where n_{ij} is the number of relations between node i and node j, $\hat{E}_a[n_{ij}]$ is the expected values, and $\hat{\sigma}$ is the standard deviation of the number of links between node i and node j which is estimated using Monte-Carlo. Monte-Carlo method is useful to solve complex problem which can't be solved analytically [12].

From this calculation, PPI=0 indicates that number of links between actors equals with the expected values. While PPI>0 indicates that number of links between actors is greater than the expected values and vice versa for PPI<0. Although the range result of PPI will ranging from $-\infty$ to $+\infty$, but the result will be normalized from 0 to 1.

C. Kretschmer Method

The measurement of degree centrality will use Kretschmer method. This method is the new complex measure of degree centrality based on the original measure of degree centrality related to Wasserman and Faust [9]. According to Coulon [19], the calculation of degree centrality on SNA is involving un-weighted ties because it is easier to calculate. But, Hildrun Kretschmer and Theo Kretschmer [16] introduced new complex method for degree centrality calculation which is involving weighted ties, known as Kretschmer method. By involving ties in the calculation, the value obtained will be more specific and accurate. Kretschmer method can be implemented to analyze co-authorship, citations, or web networks relations.

This method also creates new definition for centrality measures, which are calculation of the entropy, degree centrality (DC), complex degree centrality (CDC), and complex group degree centrality (CGDC). Because this paper only consider on the node level calculation, then the CGDC will not be calculated.

1) Total Relation (TR)

The total relation formed between user x and other users can be symbolized in TRx. This value indicates the total relation, which is the sum of interaction following, mention, and reply, formed between user x and other users. In other words, the calculation of total relation of a node is based on the total sum of U_{XAi}, the value of relations formed by node x to others node (Ai). TRx is formulated in:

$$TRx = U_{XA1} + U_{XA2} + U_{XA3} + U_{XA4} + \ldots + U_{XAi} \quad (2)$$

2) Entropy (Hx)

The entropy calculation is the probabilistic result of a node to flow and leverage information between nodes. The higher the entropy result of a node, then the probability of this node to get more information and share to other nodes is getting higher too.

The calculation of entropy H(Ki) as follows [16]:

For the series of numbers Ki(i=1,2,3,…,z), Ki≠0

$$h_i = K_i / \sum_{i=1}^{z} K_i \quad (3)$$

Then node entropy value is:

$$H(K_i) = -\sum_{i=1}^{z} h_i \,.\, \log_2 hi \quad (4)$$

Stipulation: if $\sum_{i=1}^{z} K_i = 0$, then $2^{H(K_i)} = 0$ (5)

3) Degree Centrality (DCx)

Degree centrality measurement is the calculation of total relations of an actor towards other actors, or in other words is the calculation of total relations node A, where Ki equals to A to Bi as follows:

$$K_i = U_{ABi} \quad (6)$$

So, the degree centrality value of a node equals to:

$$DC_A = 2^{H(Ki)} \quad (7)$$

4) Complex Degree Centrality (CDCx)

The complex degree centrality value is the centrality value of a node which is more complex due to the involvement of ties weight. The formula is:

$$CDC_A = (DC_A * TR_A)^{1/2} \quad (8)$$

where DC_A is the number of degree in node A and TR_A is the total sum of relation in node A. The CDC value of node A equals to the geometric mean of the number of nodes to which the node is connected and the total strength of the ties [16].

III. System Design and implementation

A. System Design

The system design for the implementation of those method is:

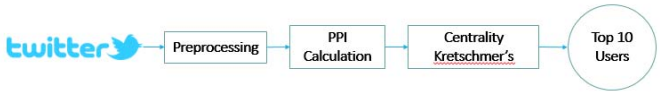

Fig. 1. System design

Based on this picture, the system design is generally divided into five steps as follows:

1) Data input

The data that will be used as the input of the system is data from Twitter and retrieved using NodeXL. The retrieved data is still in XML format. As SNA is represented into a graph, then the node for this graph is username in Twitter and the edge is the relation (following, mention, and reply) performed by user.

2) Pre-processing

The pre-processing process is creating nxn matrix or adjacency matrix to form a graph which can be done by using NodeXL too. This matrix will be used as the input of the Kretschmer's method.

3) Weighting ties

After creating matrix, every relation created by two nodes will be added weight 1 if there is a relation. Then the sum of the weight will be calculated using Probabilistic Partnership Index (PPI) indicator.

4) Kretschmer's centrality measurement

As the value of PPI calculation obtained, then it will be used as the input value or the value of total relation for the Kretschmer calculation.

5) Graph visualization and top 10 users.

B. Calculation Process

The calculation process is the measure of PPI as the calculation of total relation that will be used to calculate the degree centrality Kretschmer. Before calculate the total relation, firstly determine the graph that show the relation of the user. For example:

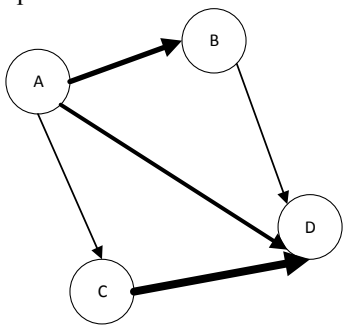

Fig. 2. Sample graph

Figure 2 shows the directed-weighted graph illustration. Every relations formed by two nodes have its own weight and direction. Then, this graph will be represented into asymmetric weighted adjacency matrix. Here is the matrix representation from above graph:

TABLE I. ASIMETRIC -WEIGHTED ADJACENCY MATRIX

Vertex	A	B	C	D
A	-	30	5	22
B	0	-	0	3
C	0	0	-	35
D	0	0	0	-

Based on table 1 above, every cell represent the value of relation between two nodes, that is the n_{ij} value. After the n_{ij} value obtained, then the expected value and standard deviations for each node can be calculated. The expected value and standard deviations obtained in this calculation is attained by using Monte Carlo algorithm and pseudo random number generator. The pseudo random number generator is performed at 100 times randomization for each node. Next is the Probabilistic Partnership Index calculation. For demonstration, there will be calculation of PPI value between relation A and B as follows:

n_{AB} = total relations, 30

μ = expected value, $= \frac{Total\ relations}{Sum\ of\ nodes} = \frac{97}{16} = 6.0625$

N = random number generator, 100 trials.

Expected value = for example $\hat{E}_a[n_{AB}] = 29.33$

Standard deviation =

$$\hat{O} = \sqrt{(\hat{n}AB + \mu)^2 . \frac{\hat{n}AB}{N} . \hat{n}AB}$$

$$= \sqrt{(1 + 6.0625)^2 . \frac{1}{100} . 1} \approx 0.70625$$

So, the PPI value is:

$PPI = (n_{AB} - \hat{E}_a[n_{AB}])/\hat{O}$

$= (30\text{-}29.33)/0.70625 \approx 0.95$

From this calculation, the PPI value obtained for relation A and B is 0.95. This value represent the scientific cooperation between node A and node B. So, this way of calculation is also applied to all nodes to obtain the PPI values for each node.

Once PPI has been calculated, the next process is to find the degree centrality value by using Kretschmer approach.

Based on this calculation, the obtained result of centrality value node A is equal to 2.4597. This result represent the leverage level of node A towards another node in Twitter environment. To find out the centrality values of others node and its leverage level to its environment, the table IV below is showing the degree centrality result based on algorithmic calculation:

TABLE II. DEGREE CENTRALITY KRETSCHMER RESULT

No.	Total Relation (TR)	Entropy (Hk)	Degree Centrality (DC)	Complex Degree Centrality (CDC)	Id
1	2.204734	1.45991191	2.7509157	2.462729646	A
2	0.24916	-0.9715635	0.4142248	0.339314134	B
3	1.337115	1.025384	-0.0894234	0.055945123	C
4	-0.50225	1.27429056	2.4187984	6.74902E-17	D

C. Testing Result

To perform testing process, firstly the range value of the interaction should be determine. Based on the previous study on SNA, the range value of interaction which will be applied in this paper is:

TABLE III. SCENARIO OF WEIGHT CHANGING ON INTERACTION

Testing	Scenario 1			Scenario 2			Scenario 3		
	Followed/following	Mention	Reply	Followed/following	Mention	Reply	Followed/following	Mention	Reply
1	2	1.5	0.75	2	0.5	1	1.5	1.5	1
2	2	1.5	0.8	2	0.6	1	1.6	1.5	1
3	2	1.5	0.85	2	0.7	1	1.7	1.5	1
4	2	1.5	0.9	2	0.8	1	1.8	1.5	1
5	2	1.5	0.95	2	0.9	1	1.9	1.5	1
6	2	1.5	1	2	1	1	2	1.5	1
7	2	1.5	1.05	2	1.1	1	2.1	1.5	1
8	2	1.5	1.1	2	1.2	1	2.2	1.5	1
9	2	1.5	1.15	2	1.3	1	2.3	1.5	1
10	2	1.5	1.2	2	1.4	1	2.4	1.5	1
11	2	1.5	1.25	2	1.5	1	2.5	1.5	1

By differentiate the weight for each interaction, therefore the effect that occurs on the value of CDC for each user can be discovered.

1) First Scenario: Changing Value in Reply

Once the testing process is done, the CDC value for each user is obtained. The CDC value for each user can be the same, but mostly different due to the different value for each PPI result. CDC value for each user is:

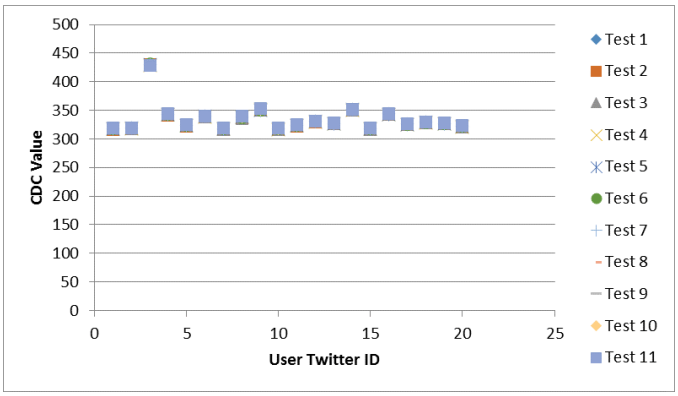

Fig. 3. Graph showing change value of CDC in the changing weight of reply.

The value of y-axis represent value of CDC for each user. Meanwhile the x-axis represent user id. The sum value of relations obtained from PPI calculation. As mentioned above, in the calculation process of PPI is applied pseudo random number generator which affected to expected value and standard deviation result. That is why the PPI result is different for each user.

The overall result for this scenario is user @nisasharafina got the highest value of CDC. It means that @nisasharafina is the central node with the highest CDC value to spread information and influence other users.

2) Second Scenario: Changing Value in Mention

The second scenario is the changing value in Mention. After the testing process finished, the CDC value has obtained. The CDC value for each user also can be the same or different. Here is the graph showing the CDC value for some users:

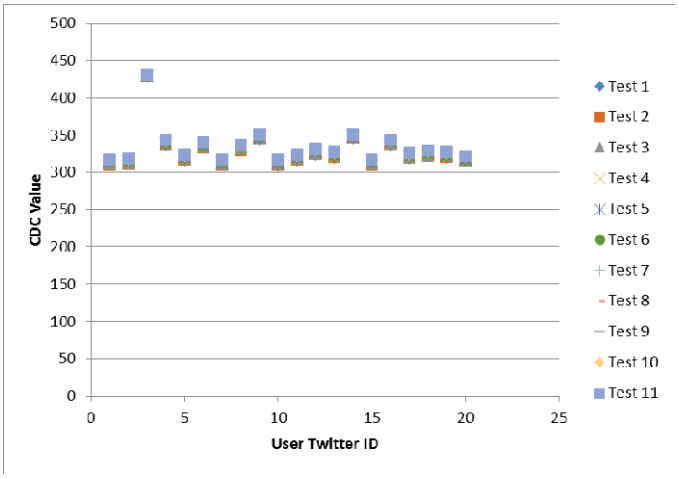

Fig. 4. Graph showing change value of CDC in the changing weight of mention.

Overall, the CDC value for each user increase from the first until the eleventh test as the increase number of the total relation which is obtained from PPI calculation. It means that if the weight of a node is increase, the CDC value will also increase. Hence, mention is one of the influential interaction in Twitter.

3) Third Scenario: Changing Value in Following

The third scenario is changing in the value of following interaction, but the value for mention and reply still the same. After testing process, the CDC value is obtained. Here is the result:

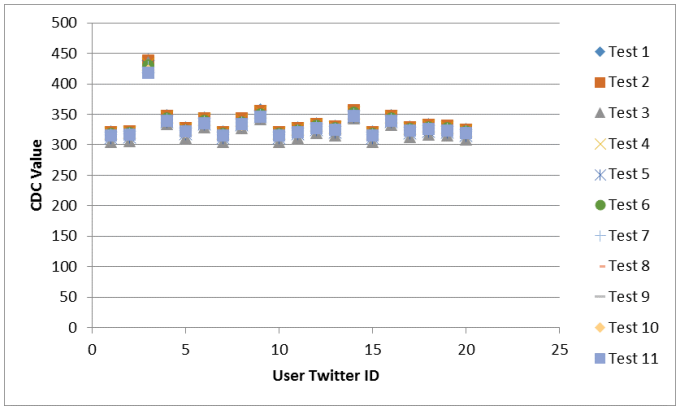

Fig. 5. Graph showing change value of CDC in the changing weight of following.

From the picture above, CDC values from the first until the last testing is showing that the value remain increase but seems on the same value. This result is different with the result obtained in the first and second scenario, where the result is always increase incisively. Obtaining stable values is due to the amount of interaction following on the data analysis which is bigger than the others. Moreover, almost all user are having following interaction in the data analysis. Next, the effect of PPI weighting which is applying Monte Carlo method and pseudo random number generator.

Hence, it can be concluded that the change value of the weight in following interaction give rise to the change value of the CDC.

IV. CONCLUSION

Based on the experimental evaluation process and analysis, the conclusion is the result of degree centrality Kretschmer testing without using PPI is that the total number of relations proportional to the CDC value. While the result obtained from combining Kretschmer with PPI is showing that the total number of relations proportional to the CDC value but in some cases of testing, the CDC value of a node increase or decrease because of using Monte Carlo and pseudo random number generator in PPI. From those three interaction in Twitter, following is the most influential interaction due to its domination in data and also incisive change in CDC value. The additional value in interaction, following, mention, and

reply, during testing process of Kretschmer without using PPI method is proportional to its CDC value of a node. As well as during testing process of Kretschmer with PPI, the additional value in interaction is also proportional to its CDC value of a node although the CDC value is not always increase, but mostly from the first to the eleventh testing stage the value is increasing.

Through this experiments, it shown that Probabilistic Partnership Index (PPI) can be used as an alternative weighting method to degree centrality calculation in social network based on the result given which shown more specific number. All in all, this experiment can be analyzed in the future based on, firstly, the dataset to be used in the next experiment should be data that can represent the following, mention, and reply with equally or nearly on the same percentage so that the data that have been generated can represent the actual conditions in Twitter. In closing, this PPI can be applied to study other social media such as Facebook, Plurk, Flickr, or Youtube.

Acknowledgments
The research is an initial research supported by Research Grant Hibah Bersaing DIKTI 2014.

References

[1] A. Sitaram, "Predicting the Future With Social Media," Social Computing Lab HP Labs, Palo Alto, California, USA, 2010.

[2] C. Giorgos, "Social Network Analysis (SNA) Including a Tutorial on Concepts and Methods," National University of Singapore, Singapore.

[3] Crnovrsanin, Tark; Carlos D., Correa. and Kwan-Liu Ma, "Social Network Discovery basd on Sensitivity Analysis," *Advances in Social Network Analysis and Mining,* pp. 107-112, 2009.

[4] E. Otte & Rosseau, "Social Network Analysis: a Powerful Strategy, also for the Information Sciences," *Information Science,* vol. 28, pp. 443-455, 2002.

[5] Java, Akshay; Xiaodan Song; Tim Finin, dkk, "Why We Twitter: Understanding Microblogging Usage and Communities," *9th WEBKDD and 1st SNA-KDD Workshop 2007,* 2007.

[6] L. C. Freeman, "The Study of Social Networks," 2002. [Online]. Available: http://www.insna.org/INSNA/na_inf.htm. [Accessed 20 November 2012].

[7] M. Newman, "A Measure of Betweenness Centrality Based on Random Walks," Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109-1120, 2003.

[8] M. A. Russel, Mining the Social Web 1st Ed, USA: O'Reilly, Inc, 2011.

[9] S. Wasserman and K. Faust, Social Network Analysis, Methods, and Application, Cambridge: Cambridge University Press, 1994.

[10] Stepanyan, Karen; Kerstin Borau; and Carsten Ullrich, "A Social Network Analysis Perspective on Student Interaction Within the Twitter Microblogging Environment," *10th IEEE International Conference on Advanced Learning Technologies,* pp. 70-72, 2010.

[11] V. Krebs, "How to do Social Network Analysis," 2006. [Online]. Available: http://www.orgnet.com/sna.html. [Accessed 20 November 2012].

[12] Yamashita, Yasuhiro and Yoshiko Okubo, "Patterns of Scientific Collaboration between Japan and France: Inter-sectoral Analysis using Probabilistic Partnership Index (PPI)," *Akademiai Kiado, Budapest and Springer Dordrecht, Scientometrics,* vol. 68, no. 2, pp. 203-324, 2006.

[13] E. Purnama, "Estimation of Rumor Sources in Social Network," School of Electrical and Electronic Engineering, Nanyang Technological University , Singapore, 2012

[14] A. Sulasikin, "Analisis Degree Centrality dalam Social Network Analysis menggunakan Probabilistic Affinity Index (PAI) Pada Graf Berarah-berbobot," Fakultas Informatika, Institut Teknologi Telkom, Bandung, 2012.

[15] Eunice E. Santos, Chair et al, "Effective and Efficient Methodologies for Social Network Analysis," Virginia, USA, 2007.

[16] Kretschmer, Hildrun & Theo Kretschmer, "A New Centrality Measure for Social Network Analysis Applicable to Bibliometric and Webometric Data," Department of Library and Information Science, 26-D-10117, Humboldt-University Berlin, 2010.

[17] Roger S. Pressman, Software Engineering: A Practitioner's Approach 5th Edition, Mc Graw Hill, 2001.

[18] Z. A. Rachman, W. Maharani and Adiwijaya., "The Analysis and Implementation of Degree Centrality in Weighted Graph in Social Network Analysis," *Proceeding of International Conference on Information & Communication Technology 2013,* 2013.

[19] F. Coulon, "The use of Social Network Analysis in Innovation Research: A Literature Review," Lund University, Sweden, 2005.