# Identification of Influential Nodes from Social Networks based on Enhanced Degree Centrality Measure

Amedapu Srinivas
Research Scholar
Dept. of Computer Science and Engineering
National Institute of Technology,
Tiruchirappalli,Tamil Nadu, INDIA - 620015
406111051@nitt.edu

R. Leela Velusamy
Associate Professor
Dept. of Computer Science and Engineering
National Institute of Technology,
Tiruchirappalli,Tamil Nadu, INDIA - 620015
leela@nitt.edu

*Abstract*— **A social network is a set of relationships and interactions among social entities such as individuals, organizations, and groups. The social network analysis is one of the major topics in the ongoing research field. The major problem regarding the social network is finding the most influential objects or persons. Identification of most influential nodes in a social network is a tedious task as large numbers of new users join the network every day. The most commonly used method is to consider the social network as a graph and find the most influential nodes by analyzing it. The degree centrality method is node based and has the advantage of easy identification of most influential nodes. In this paper a method called "Enhanced Degree Centrality Measure" is proposed which integrates clustering co-efficient value along with node based degree centrality. The enhanced degree centrality measure is applied to three different datasets which are obtained from the Facebook to analyze the performance. The response obtained is compared with existing methods such as degree centrality method and SPIN algorithm. By comparison, it is found that highest number of active nodes identified by the proposed method is 64 when compared with SPIN algorithm which identifies only 55.**

***Keywords— Social networs; influential node; degree centrality; clustering co-efficient***

## I. INTRODUCTION

The recent explosion of activity on the internet has given rise to huge amount of social network data which is usefully viewed as a collection of entities and associations among them. Social network analysis is concerned with uncovering patterns in the connections between entities. Recently, considerable attention has been devoted to social network analysis since the rise of the Internet and the World Wide Web which have enabled the collection of large amount of real data from social networks. A social network is the network of relationships and interactions among social entities such as individuals, organizations, and groups. Examples include blog networks, collaboration networks, email networks etc. In these networks, varieties of topics are discussed by different social network researchers.

Social network plays an important role for the spread of information since a piece of information can propagate from one node to another through a link on the network in the form of "word-of-mouth" communication. Thus, it is an important research issue to identify influential nodes for information diffusion on a social network in terms of sociology and "viral marketing" [1]. The process of propagation of information consists of (i) Selecting a set of most influential active nodes and feeding the information to these nodes, (ii) These active nodes in-turn propagate the information to the other nodes which may be active or inactive. The inactive nodes become active upon receiving the information through the links connecting the active nodes. These secondary nodes further transmit the information to the other nodes connected to them. This process of propagation continues until no more nodes could be turned active. Thus, the essential problem in diffusion is how to select a set of influential nodes as the active starting nodes [2].

There are several methods to identify influential nodes in social networks which include degree centrality, closeness centrality, betweenness centrality, the k-shell decomposition centrality, greedy algorithm etc. All these methods have their own limitations when applied to very large networks. For example, high degree nodes will be treated as influential nodes which all may belong to one large community [3]. Thus one of the tough jobs in social network is to identity the influential nodes.

In this paper a method called "Enhanced Degree Centrality Measure" is presented to identify the most influential nodes in a social network, which is developed based on degree centrality measure and cluster co-efficient value. The degree centrality considered by the proposed approach is per node based, that will help to identify the influence of each node more easily. The proposed approach incorporates the clustering co-efficient value with the node based degree centrality value to enhance identification of influential nodes. The enhanced degree centrality measure is used with three different datasets obtained from Facebook to analyse the performance.

978-1-4799-8047-5/15/$31.00 ©2015 IEEE

The important aspects in this paper to identify influential nodes are:

- Design of an enhanced degree centrality measure to find the influential nodes in a social network.
- Implementation and evaluation of the designed method with various benchmark tests and perform a comparative analysis.

The rest of the paper is organized as follows: the second section gives ideas about recent methods and researches regarding the identification of influential nodes. The third section plots the motivation behind proposing the method. The fourth section explains the design of the proposed method. In the fifth section the proposed method is implemented and tested and analysis of the test result is done. Sixth section concludes the paper.

## II. Literature Review

Literature presents different topics for social network analysis. Among various topics, influential node identification, community analysis and spam detection have received a major role in social network analysis. Here, a few existing methods are presented on these topics. Narayanam, R. and Narahari, Y [4] have proposed an algorithm called ShaPley value-based Influential Nodes (SPINs) for solving the top-k nodes problem and the λ -coverage problem. They compared the performance of the SPIN algorithm with Maximum Degree Heuristic (MDH), High Clustering Heuristic (HCH), and Greedy Algorithms. With their experimentation on four synthetically generated random graphs and six real-world data sets (Celegans, Jazz, NIPS authorship data set, Netscience data set, High-Energy Physics data set, and Political Books data set), they proved that the SPIN approach is more efficient in execution time and influential nodes identification.

Masahiro Kimura et al. [5] have used bond percolation and graph theory and proposed good approximate solution and compared the results with conventional method. They have proved that their method is effective in theoretically and experimentally. Jingyu Zhou et al. [6] have designed Greedy Algorithm based on Users' Preferences (GAUP) which is a two stage mining algorithm to identify most influential nodes based on user preferences. In GAUP first evaluates uses preferences with a model called Singular Value Decomposition (SVD) and in second stage evaluates top-K nodes. With test data sets they proved GAUP performs better than the state-of-the-art greedy algorithm, SVD-based collaborative filtering and Hyperlink-Induced Topic Search algorithm (HITS).

Kimura, M et al. [7] have presented cascade model to find most influential nodes for information diffusion on social networks. In their SR-community structure algorithm they explored features which play important role in identifying influential nodes. With their experiments using large social networks data, they proved SR-community structure can be more strongly correlated with greedy solution than the Newman and Leicht community structure algorithm.

Lilian Weng et al. [8] have presented a method to generate predictive knowledge from community structure about what information will spread widely. They analyzed in computational social science, social media analytics, and marketing applications. They showed that a few viral memes spread across many communities, like diseases. They demonstrated that the future popularity of a meme can be predicted by quantifying its early spreading pattern in terms of community concentration. The more communities a meme permeates, the more viral it is.

Marco Vanetti et al. [9] have proposed a system allowing OSN users to have a direct control on the messages posted on their walls. This was achieved through a flexible rule-based system that allows users to customize the filtering criteria to be applied to their walls, and a Machine Learning-based soft classifier automatically labelling messages in support of content-based filtering.

## III. Motivation Behind The Approach

Social network analysis is concerned with uncovering patterns in the connections between entities. Recently, considerable attention has been paid to social network analysis since the rise of the Internet and the World Wide Web has enabled researchers to collect the data from real large social networks. A social network is the network of relationships and interactions among social entities such as individuals, organizations and groups. The social network analysis is one of the major topics in the ongoing research field. The major topic regarding the social network is finding the most influential object or person. The most commonly used method is to consider the social network as a graph and find the most influential node by analysing the graph. Recently, Ramasuri Narayanam and Yadati Narahari [4] have proposed a method for finding the influential nodes from the social network by considering the network as a weighted undirected graph. The method uses a parameter called "ShaPley value" for identifying the active nodes. In this paper a method is proposed called "Enhanced Degree Centrality Measure" to identify the influential nodes from a social network data based on degree centrality of each node. The proposed method enhances the basic degree centrality formula by incorporating the clustering co-efficient. In graph theory, a clustering co-efficient is a measure of the degree to which nodes in a graph tend to cluster together. Evidence suggests that in most real-world networks, and in particular social networks, nodes tend to create tightly knit groups characterised by a relatively high density of ties; this likelihood tends to be greater than the average probability of a tie randomly established between two nodes. The exploration of nodes will be higher, when incorporating the clustering co-efficient with the degree centrality value. The enhancement is aimed to improve the accuracy in finding the most influential nodes from the social networks.

## IV. Proposed Method For Influential Node Calculation

This section explains the design and development of the proposed algorithm for influential nodes identification. Here, two popular heuristics, degree centrality and clustering co-efficient, are effectively combined to design an algorithm for influential nodes selection.

### *A. Basic Features*

The proposed method is based on two major terms adopted from the graph theory, viz., degree centrality and clustering co-efficient.

*1) Degree Centrality:* The first and conceptually simplest method is the degree centrality, which is defined as the number of links incident upon a node (i.e., the number of ties that a node has). The degree can be interpreted in terms of the immediate risk of a node for catching whatever is flowing through the network (such as a virus, or some information). In the case of a directed network (where the ties have direction), usually two separate measures of degree centrality are defined, namely in-degree and out-degree. In-degree is a count of the number of ties directed to the node and out-degree is the number of ties that the node directs to others. When the ties are associated to some positive aspects such as friendship or collaboration, in-degree is often interpreted as a form of popularity, and out-degree as gregariousness. Since, the proposed approach considers an undirected graph, degree is used in common. The degree basically means the ties of a node with other nodes.

| | v1 | v2 | v3 | v4 | v5 | v6 | v7 | v8 | v9 | v10 | v11 | v12 | Deg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| v1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| v2 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| v3 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| v4 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 5 |
| v5 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |
| v6 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 5 |
| v7 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 4 |
| v8 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 5 |
| v9 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 5 |
| v10 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 6 |
| v11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 4 |
| v12 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 4 |

Fig. 1. Sample data matrix

Consider the Fig. 1 sample data matrix, which will generate the graph as shown in Fig. 2.

The degree centrality of each node of the Fig. 2 graph can be calculated with the help of the following equations. The centrality can be represented using "deg".

$$\deg(G) = \sum_{j=1}^{|v|} \frac{C(v^*) - C(v_j)}{H(v_i)} \tag{1}$$

G is the sample graph, v* is the vertex with highest associations in the whole graph (i.e., v10 with 6), $v_j$ represents the connections of the selected node, and H is node level centrality, i.e. the centrality corresponding to a number of nodes associated with the selected node $v_{i.}$ The value of H can be calculated using equation (2)

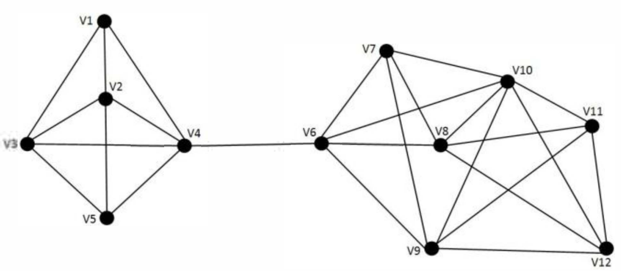

Fig .2. Sample graph (G)

$$H(v_i) = \sum_{j=1}^{|y|} C(y^*) - C(y_j) \tag{2}$$

Here, the node y* and $y_j$ are from the connected graph Y, where Y is a subset of graph G with the nodes which are directly connected to $v_i$. The node y* represents the node with highest number of connections in graph Y. The node $y_j$ is one of the adjacent node to node $v_i$. When the node "v4" is considered, the calculation on deg (G) will proceed as,

$$H(v4) = (v6 - v1) + (v6 - v2) + (v6 - v3) + \\ (v6 - v4) + (v6 - v5) + (v6 - v6)$$

$$H(v4) = 6$$

Thus centrality corresponding to node v4 will be,

$$\deg(v4) = (6-5)/6 = 0.1667$$

The whole nodes in the graph G will be selected for the purpose of calculating degree centrality as per the above procedure. The proposed method extracts the degree centrality with each node in order to understand the most influential nodes. According to the proposed approach the node with least degree centrality (as H value will be high for the most influential node and it is in the denominator) can be considered as the node with most number of connections and hence the most influential node.

*2) Clustering Co-efficienty:* One property of a graph is the clustering co-efficient. The clustering co-efficient is the measure of the extent to which one node is also friends of other nodes. The measure has become popular due to a 1998 paper in Nature by Watts and Strogatz [10]. This property is sometimes called the local clustering co-efficient. The clustering co-efficient (CC) basically draws the association of one node to another, which can be calculated using the equations (3) and (4).

$$CC(v) = \frac{No.of\ neighbor\ nodes\ connected\ by\ edges}{Max.no.of\ edges\ may\ exits\ among\ neighbors} \tag{3}$$

$$Max.no.of\ edges\ among\ n\ nodes = \frac{n*(n-1)}{2} \tag{4}$$

Consider node v2 for calculating the clustering co-efficient, the total numbers of connected neighbours to the v2 are 4, which are v1, v3, v4 and v5. Thus, the clustering co-efficient value of v2 can be calculated as,

$$CC(v2) = \frac{5}{\frac{4*(4-1)}{2}} = \frac{5}{6} = 0.833$$

### B. Enhanced Degree Centrality Measuere

In this method clustering co-efficient value is used to enhance the degree centrality value of nodes, i.e., the enhanced degree centrality is the dot product of degree centrality (1) and clustering co-efficient (3), which is show in the equation (5).

$$En\_\deg(node_i) = \frac{C(v^*) - C(node_i)}{H(node_i)} * CC(node_i) \qquad (5)$$

The next phase of the proposed approach is to sort the nodes according to their value based on (5). The average value of the enhanced degree centrality is considered as a threshold for filtering the nodes. The active nodes are selected from the sorted nodes based on whose degree centrality value is above the threshold value. Then the top quarter nodes of the active nodes are considered as the most influential nodes.

Considering Fig. 2, the degree centrality of node v2 can be calculated as,

$$\deg(v2) = \frac{C(v10) - C(v2)}{H(v2)} = \frac{6-4}{6} = 0.333$$

Clustering co-efficient of v2 can be calculated as,

$$CC(v2) = \frac{5}{\frac{4*(4-1)}{2}} = \frac{5}{6} = 0.833$$

Thus according to the proposed formula, the enhanced degree centrality of the node v2 becomes,

$$En\_\deg(v2) = \frac{C(v10) - C(v2)}{H(v2)} * CC(v2) = 0.333 * 0.833 = 0.277$$

And from the experiments, it is observed that the enhanced degree centrality equation is able to identify more number of effective nodes when compared to the conventional degree centrality calculation.

## V. Experimental Results

This section discuss about the experimental analysis of the proposed approach. The experimental setup includes Intel core i5 processor, 2 GB of RAM and 500 GB hard disk.

### A. Dataset Description

The dataset considered for analysing the proposed approach is a network data in the form of a graph. The data set considered is part of the social network data from Facebook [11]. This network describes the friendship relations between the users of Facebook. It was collected in April 2009 through data scraping from Facebook. The selected graph is undirected and un-weighted. The total size of the graph is 59,216,214 vertices with edge weights of 95,522,012. As per the dataset, the maximum degree of the graph is 4960 and the average is 3. Three sets of data are extracted from the graph to test the proposed approach and each set possess 1000 records. The response of the proposed approach with all the three sets is considered for the experimental analysis.

### B. Performance Evaluation

The proposed approach uses the three sets of Facebook network data for testing the enhanced degree centrality. The 1000 nodes dataset is again classified into five equal intervals of 200 nodes each for the detailed analysis of the proposed approach.

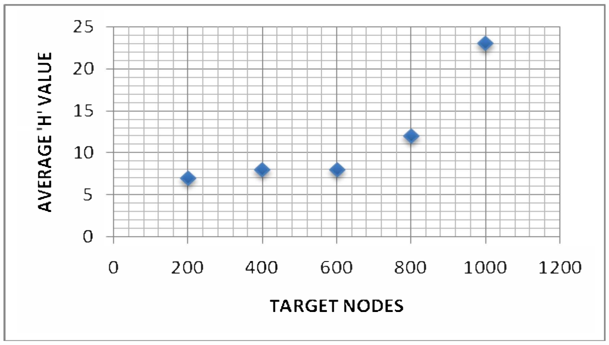

Fig. 3. Dataset1 H value

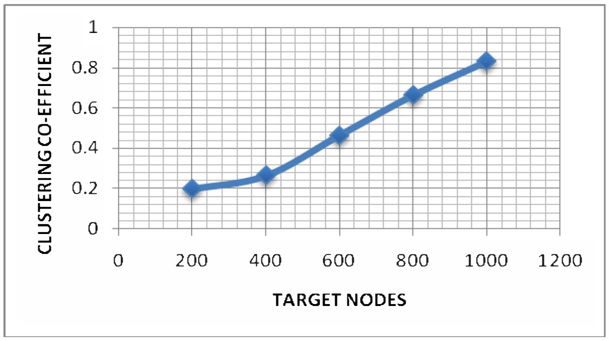

Fig. 4. Dataset1 clustering co-efficient value

Fig. 5 represents the performance of the proposed approach according to the first 1000 data nodes. In figures 3 and 4, the average H value and CC values obtained for the Dataset1 are presented. It is observed that the average H and clustering co-efficient values are increased as the number of nodes is increased, and these values will play a major role in the increase in the number of effective nodes. The result of the active nodes identified for the Dataset1 is plotted in Fig. 5 which represents both the degree centrality based calculation and the enhanced degree centrality based calculations. The results show that more number of active nodes is identified with enhanced degree centrality compared to the normal degree centrality.

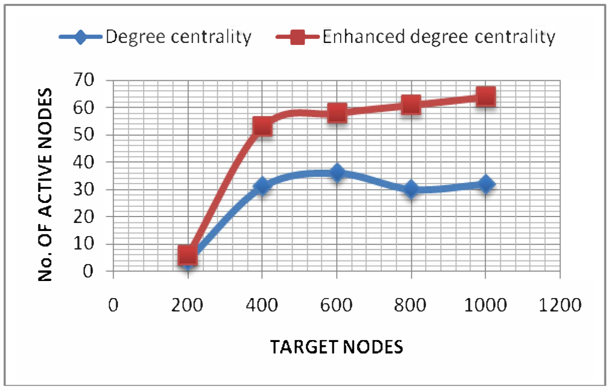

Fig. 5. Dataset1 active node count

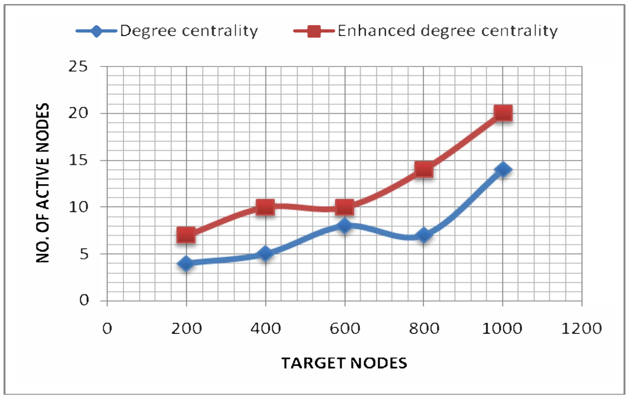

Fig. 6. Dataset2 active node count

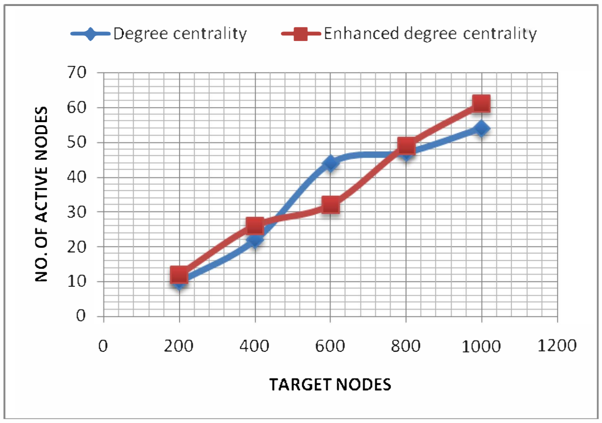

Fig. 7. Dataset3 active node count

Figures 6 and 7 represent the analysis of the number of active nodes for Dataset2 and Dataset3, respectively. Again, it can be seen that the enhanced degree centrality method resulted in more number of active nodes compared to the normal degree centrality method for both Dataset 2 and 3. The number of active nodes obtained for the Dataset1 with the enhanced degree centrality is 64, whereas with the normal degree centrality, the number of active nodes is 54.

### C. *Comparative Analysis*

The performance of the proposed approach is compared with the existing methods to evaluate its effectiveness. The method [4] finds the active nodes based on a parameter called the ShaPley value. On the other hand, the proposed method uses the degree centrality.

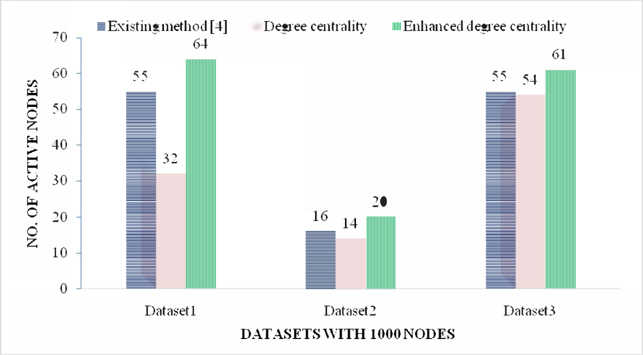

Fig. 8. Comparative Analysis

Fig. 8 shows the responses of the comparative analysis of the proposed approach and the method based on [4]. The methods are plotted based on three datasets with 1000 nodes each. Each of the datasets is different from one another in the number of connections between the nodes. The results show that the proposed approach is a better method over the existing methods based on the number of active nodes identified per dataset. The ShaPley value based method obtained better results over the normal centrality based method in all the datasets. The maximum number of active nodes count obtained for Dataset1 by the method [4] is 55 and by the proposed approach is 64. The comparative analysis shows that the proposed approach is capable of identifying more influential nodes than the existing method [4].

## VI. Conclusion

The influential nodes identification in a social network is a tedious task as an immense amount of users are added to the social network every day. Hence, in this paper, a method to identify the most influential nodes based on an enhanced degree centrality measure is developed. The basic parameters used by the proposed approach are degree centrality and clustering co-efficient. The proposed approach incorporates the clustering co-efficient value with the node based degree centrality value to enhance the count of the influential nodes. The enhanced degree centrality measure is used for three different datasets to analyse the performance. The responses obtained from the experimental analysis and comparison with the existing methods shows that the proposed method is more effective in identifying more number of influential nodes.

## References

[1] Masahiro Kimura, Kazumasa Yamakawa, Kazumi Saito, and Hiroshi Motoda, "Community Analysis of Influential Nodes for Information Diffusion on a Social Network", IEEE International Joint Conference on Neural Networks, pp.1358 - 1363, 2008.

[2] Feng Zou, Zhao Zhang, and Weili Wu, "Latency-Bounded Minimum Influential Node Selection in Social Networks", Wireless Algorithms,

Systems, and Applications, Lecture Notes in Computer Science, Volume 5682, pp 519-526, 2009.

[3] Xiaohang Zhang, Ji Zhu, Qi Wang, Han Zhao, "Identifying influential nodes in complex networks with community structure", Knowledge-Based Systems, vol. 42, pp. 74–84, 2013.

[4] Narayanam, R. , Narahari, Y.,"A ShaPley Value-Based Approach to Discover Influential Nodes in Social Networks", Automation Science and Engineering, IEEE Transactions on Vol. 8, no. 1, pp. 130-147, 2011.

[5] Masahiro Kimura, Kazumi Saito, Ryohei Nakano, Hiroshi Motoda, "Extracting influential nodes on a social network for information diffusion", Data Mining and Knowledge Discovery , Vol 20, No 1 , pp 70-97 , 2010.

[6] Jingyu Zhou, Yunlong Zhang, Jia Cheng, "Preference-based mining of top-K influential nodes in social networks", Future Generation Computer Systems, Vol. 31, pp 40–47, 2014.

[7] Kimura, M., Yamakawa, K. ; Saito, K. ; Motoda, H.," Community Analysis of Influential Nodes for Information Diffusion on a Social Network" IEEE International Joint Conference on Neural Networks, pp. 1358 - 1363, 2008.

[8] Lilian Weng, Filippo Menczer, Yong-Yeol Ahn, "Virality Prediction and Community Structure in Social Networks", Scientific Reports, No. 2522, 2013.

[9] Marco Vanetti, Elisabetta Binaghi, Elena Ferrari, Barbara Carminati, Moreno Carullo, "A System to Filter Unwanted Messages from OSN User Walls," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 2, pp. 285-297, 2013.

[10] Duncan J. Watts, and Stevan H. Strogatz, "Collective dynamics of „small-world networks'", Nature, vol 393, pp 440-442, 1998.

[11] dataset website: " http://konect.uni-koblenz.de/networks/facebook-sg"