

User Position Measures in Social Networks

Katarzyna Musiał
Wrocław University of
Technology, Institute of
Computer Science
Wyb. Wyspińskiego 27
Wrocław, Poland
katarzyna.musial
@pwr.wroc.pl

Przemysław Kazienko
Wrocław University of
Technology, Institute of
Computer Science
Wyb. Wyspińskiego 27
Wrocław, Poland
kazienko@pwr.wroc.pl

Piotr Bródka
Wrocław University of
Technology, Institute of
Computer Science
Wyb. Wyspińskiego 27
Wrocław, Poland
piotr.brodka
@pwr.wroc.pl

ABSTRACT

Network analysis offers many centrality measures that are successfully utilized in the process of investigating the social network profile. The most important and representative measures are presented in the paper. It includes indegree centrality, proximity prestige, rank prestige, node position, outdegree centrality, eccentricity, closeness centrality, and betweenness centrality. Both feature analysis and experimental comparative studies revealed the general profile of selected measures.

Categories and Subject Descriptors

E.1 [Data Structures]: Graphs and Networks; H.4 [Information Systems Applications]: Miscellaneous; J.4 [Social and Behavioral Sciences]: Sociology; K.4.0 [Computers and Society]: General

General Terms

Human Factors, Management

Keywords

centrality measure, measure comparison, social network, social network analysis

1. INTRODUCTION

With the growth of the popularity of the social networks available in the World Wide Web, the greater and greater number of methods have been developed to investigate and analyze features of these networks. One of the crucial issues in social network analysis is the problem of extracting the most important (central) members, i.e. evaluate user position with respect to other community members. There are several methods used for this purpose, such as indegree centrality [37], [1], proximity prestige [37], rank prestige [37], node position [22], [25], outdegree centrality [31], [35], eccentricity, closeness centrality [2], [33], or betweenness centrality [13], [14], [16]. For each of them the appropriate algorithms were proposed [8], [11].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

/The 3rd SNA-KDD Workshop '09/ (SNA-KDD'09), June 28, 2009, Paris, France. Copyright 2009 ACM 978-1-59593-848-0...\$5.00.

However, there is lack of comparative studies on both efficiency and general usability individual measures.

2. USER POSITION

Measures (also called metrics) are used in social network analysis to describe the actors' or ties' features, characteristic within social network as well as to indicate personal importance of individuals in social network. In this section the methods that are used to evaluate the position of a user in the network are presented. Only the measures that can be used in directed graphs are considered. Directional relationships give two types of position measures, i.e. prestige and centrality. Both of them consider how prominent a user is within a community. Prominent means that a given actor is particularly visible to the other members of the network [26], [20], [18]. Note that centrality is based on the "choices" made by a user whereas prestige depends on "choices" that a given user received from others [37]. Depending on the question to be answered, different indices can be used as not every index is suitable to every application. On the other hand, a given network can be meaningfully analyzed with different indices [8].

2.1 Prestige

Considering the prestige of an actor the number of receiving choices from other users is considered. In other words, a member can be seen as prestigious when he/she possesses many ties directed to this user. Thus, prestige is more refined concept than centrality and can be used only in directed graphs [37]. The idea of prestige was first introduced by Moreno, who called it status [29].

2.1.1 Prestige Based on Node Degree

Indegree centrality, called also degree prestige, is based on the indegree number so it takes into account the number of members that are adjacent to a particular member of the community [37], [13]. In other words, more prominent people are those who received more nominations from members of the community [1]:

$$IDC(x) = \frac{i(x)}{m-1} \quad (1)$$

where:

$i(x)$ — is the number of members from the first level neighborhood that are adjacent to x ;

m — the total number of members in the social network.

This measure is a local one as it takes into account only the direct votes.

2.1.2 Proximity Prestige

Proximity prestige $PP(x)$ reflects how close all other members within the social community are to member x [37]. This measure depends on the geodesic distances, e.g. the length of the shortest paths $d(x, y_i)$ from all members y_i to the considered member x :

$$PP(x) = \frac{\frac{p(x)}{m-1}}{\frac{1}{p(x)} \sum_{i=1}^{p(x)} d(x, y_i)} = \frac{p(x)^2}{(m-1) \cdot \sum_{i=1}^{p(x)} d(x, y_i)} \quad (2)$$

where:

$p(x)$ — the number of all members y_i in the network who can reach member x , i.e. there exists a path from these members y_i to the given member x ;
 m — the number of nodes in a network.

2.1.3 Rank Prestige

Rank prestige depends not only on geodesic distance and the number of relationships, but also on the prestige of members connected with the member [21]. "It's not what you know, but whom you know" [37]. In this method, first the initial centrality of each member is established by utilization of one of the existing measures e.g. degree, betweenness or closeness centrality. These necessary preliminary node positions are used as the input for the core part of the method and highly influence the outcome, i.e. different initial measures result in different final values. Having initial values assigned, rank prestige for the given member is calculated as the sum of the initial centrality values of all other members that are connected to this node. A shortcoming of this method is that members who are not chosen by others have centrality equal zero. In consequence, these members contribute nothing to any member that is connected to them [6]. As an extension of this method, called alpha-centrality, Bonacich and Lloyd [6] propose an additional input status to ascribe each network member. This status is derived from member's general position in the company or family rather than from their relationships in the network. The numeric values of the status are added to the eigenvector-based prestige derived from member relationships. Note that unfortunately it is not always possible to establish such an additional status, especially for large social networks with thousands of members in the Internet. Another modification of the rank prestige takes into account not only the direct connections but also the indirect ones. Moreover, each indirect path is assigned with an appropriate weight [5].

The measure of prestige that was built upon the rank prestige measure that was utilized in the Web network, i.e. the network of web pages, is PageRank. Kumar et al. claim that the Web can be seen as a social network [28] and this enables similar node measures to be applied both to social networks and hypertext or web-based systems [7]. PageRank was introduced by Brin and Page to assess the value and importance of web pages [9], [4], [10]. The PageRank value of a web page takes into consideration PageRanks of all other pages that link to this particular one. Google uses this mechanism to rank the pages in their search engine.

2.1.4 Node Position

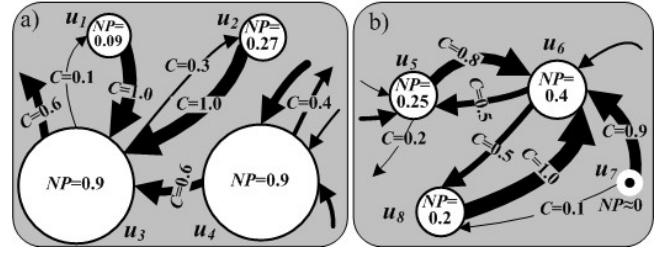


Figure 1: Two fragments of a social network. The size of the node circle corresponds to the value of its node position. The arrows reflect commitment values. $\varepsilon \approx 1$

Another measure, node position, was introduced and preliminary studied in [22] [23], [25]. Node position function $NP(x)$ of node x respects both the value of node positions of node's x connections as well as their contribution in activity in relation to x , in the following way:

$$NP(x) = (1-\varepsilon) + \varepsilon \cdot (NP(y_1) \cdot C(y_1 \rightarrow x) + \dots + NP(y_m) \cdot C(y_m \rightarrow x)) \quad (3)$$

where: ε — the constant coefficient from the range $[0, 1]$;
 y_1, \dots, y_m — acquaintances of x , i.e. nodes that are in the direct relation to x ; m — the number of x 's acquaintances;
 $C(y_1 \rightarrow x), \dots, C(y_m \rightarrow x)$ — the function that denotes the contribution in activity of y_1, \dots, y_m directed to x .

The value of ε denotes the openness of node position measure on external influences: how much x 's node positions are more static and independent (small ε) or more influenced by others (greater ε). In other words, the greater values of ε enable the neighborhood of node x to influence the x 's nodes position to a large extent.

In general, the greater node position one possesses the more valuable this member is for the entire community. It is often the case that we only need to extract the highly important persons, i.e. with the greatest node position. Such people are likely to have the biggest influence on others. As a result, we can focus our activities like advertising or target marketing solely on them and we would expect that they would entail their acquaintances. The node position of member x is inherited from the others but the level of inheritance depends on the activity of the members directed to this person, i.e. intensity of mutual communication. Thus, the node position depends both on the number and quality of relationships. A member can possess the high node position if some other people transfer their high NP to them. For example, the node position of member u_3 in Fig. 1a is 0.9 mostly come from u_3 's high commitment in the activities of member u_4 , $C(u_4 \rightarrow u_3) = 0.6$ and $C(u_4 \rightarrow u_3) \cdot NP(u_4)$ equals as much as 0.54. The contribution of two other members u_1 and u_2 in $NP(u_3)$ is only 0.36, even though their commitment values are the greatest possible $C(u_1 \rightarrow u_3) = C(u_2 \rightarrow u_3) = 1$. On the other hand, despite the very high $NP(u_3)$, the value of $NP(u_1)$ is only 0.09 due to very low u_1 's participation in u_3 's activity, $C(u_3 \rightarrow u_1) = 0.1$. Member u_3 is the only one who is active towards member u_1 . The node position of member u_6 is medium-sized: $NP(u_6) = 0.4$, although three other persons u_5 , u_7 , and u_8 pass most of their activities to u_6 : $C(u_5 \rightarrow u_6) = 0.8$, $C(u_7 \rightarrow u_6) = 0.9$, and $C(u_8 \rightarrow u_6) = 1$, Fig. 1b. It results from the low or very

low node position of u_6 's acquaintances: $NP(u_5) = 0.25$, $NP(u_8) = 0.2$ and $NP(u_7)$ is almost zero. Hence, $NP(u_3)$ is high because of high $NP(u_4)$ as well as big $C(u_1 \rightarrow u_3)$ and $C(u_2 \rightarrow u_3)$; $NP(u_1)$ is low due to small $C(u_3 \rightarrow u_1)$; and $NP(u_6)$ is medium with respect to the low importance of its neighbors.

To calculate the node position of the person within the social network the convergent, iterative algorithm is used. Its three versions were proposed and considered in [11] whereas the convergence was proved in [22]. The number of iterations as well as calculation precision can be adjusted by the application of the appropriate stop condition τ that denotes the maximum acceptable difference $NP^{(k)}(x) - NP^{(k-1)}(x)$ separately for each member x in the network. Alternatively, threshold τ may concern the differences in sums of all members' node positions instead of individual members.

The commitment function $C(y \rightarrow x)$ is a very important element in the process of node position assessment. It reflects the strength of the connection from node y to x . In other words, it denotes the part of y 's activity that is passed to x .

The value of commitment function $C(y \rightarrow x)$ in the social network must satisfy the following set of criteria:

1. The value of commitment is from the range $[0; 1]$: $\forall(x, y \in Nodes) C(y \rightarrow x) \in [0; 1]$.
2. The sum of all commitments has to equal 1, separately for each node of the network:

$$\forall(x \in Nodes) \sum_{x \in Nodes} C(x \rightarrow y) = 1 \quad (4)$$

3. If there is no relationship from y to x then $C(y \rightarrow x) = 0$.
4. If a member y is not active to anybody and other n members x_i , $i = 1, \dots, n$ are active to y , then in order to satisfy criterion 3, the sum 1 is distributed equally among all the y 's acquaintances x_i , i.e.

$$\forall(x_i \in Nodes) C(y \rightarrow x_i) = 1/n \quad (5)$$

Since the relationships are not reflexive and with respect to criterion 3, the commitment function to itself equals 0: $\forall(x \in Nodes) C(x \rightarrow x) = 0$.

According to the above criteria all values of commitment are from the range $[0; 1]$ (criterion 1) as well as the sum of all commitments equals 1, separately for each member of the network (criterion 2). Moreover, there is no relationship u_2 to u_1 so $C(u_2 \rightarrow u_1) = 0$ (criterion 3). Note also that node u_3 is not active to anybody but two others u_2 and u_4 are active to u_3 , so according to condition 4, the commitment of u_3 is equally distributed among all u_3 's connections $C(u_3 \rightarrow u_2) = C(u_3 \rightarrow u_4) = 1/2$.

The commitment function $C(x \rightarrow y)$ of member x within activity of their acquaintance y can be evaluated as the normalized sum of all activities from x to y in relation to all

activities of x :

$$C(x \rightarrow y) = \frac{A(x \rightarrow y)}{\sum_{j=1}^m A(x \rightarrow y_j)} \quad (6)$$

where: $A(x \rightarrow y)$ – the function that denotes the activity of node x directed to node y , e.g. the number of emails sent by x to y ; m – the number of all nodes within the social network.

2.2 Centrality

Centrality measures enable to find users who are extensively involved in relationships with other network members. Usually, centrality indices are applied to undirected graphs while it is not important whether the actor is prominent due to being the recipient or the source of many relationships [37]. However, after introducing appropriate modifications it is possible to use centrality methods in the directed graphs and in such a case the outgoing relationships are taken into account. The concept of centrality was first introduced by Bavelas [2].

2.2.1 Centrality Based on Node Degree

Outdegree centrality of the member x takes into account the number of outdegree of the member x for edges which are directed to the given node [31], [35]:

$$ODC(x) = \frac{o(x)}{m-1} \quad (7)$$

where:

$o(x)$ — the number of the first level neighbours to whom x is adjacent

m — the total number of members in the social network.

Users who communicate with the greater number of people obtain the greater outdegree centrality value. Actors with high outdegree are recognized by other network members as a crucial cog that occupies a central location in a network [37]. On the other hand users who have low outdegree centrality are not very open to the external world and do not communicate with many members. ODC and IDC are the simplest and most intuitive measures that can be used in network analysis.

2.2.2 Eccentricity Centrality

Eccentricity states that the most central node within the network is the one that minimizes the maximum distance to any other node in the network [8]:

$$EC(x) = \frac{1}{\max\{d(x, y) : y \in M\}} \quad (8)$$

where:

$d(x, y)$ — the length of the shortest path from user x to y

M — the set of all members of the social network

2.2.3 Closeness Centrality

The closeness centrality, in contrast to proximity prestige, pinpoints how close a member is to all the others within the social network [2]. Its main idea is that the member takes the central position if they can quickly contact other members in the network. This measure emphasizes quality (position in a network) rather than quantity (number of links, like in a centrality degree measure). The member with

high CC is a good propagator of ideas and information [3]. A similar idea was studied for hypertext systems [7]. The closeness centrality $CC(x)$ of member x tightly depends on the geodesic distance, i.e. the shortest paths from member x to all other people in the social network [33] and is calculated as follows:

$$CC(x) = \frac{m - 1}{\sum_{y \neq x, y \in M} c(x, y)} \quad (9)$$

where:

$c(x, y)$ — a function describing the distance between nodes x and y (i.e. max, min, mean or median);

m — the number of nodes in a network [13], [14], [27].

2.2.4 Betweenness Centrality

Betweenness centrality BC of member x pinpoints to what extent x is between other members. Members with high BC are very important to the network because others actors can connect with each other only through them. It can be calculated only for undirected relationships by dividing the number of shortest geodesic distances (paths) from y to z by the number of shortest geodesic distances from y to z that pass through member x . This calculation is repeated for all pairs of members y and z , excluding x . Betweenness centrality of the member x is the sum of all the outcomes [13], [14], [16], [17], [34]:

$$BC(x) = \frac{\sum_{i \neq x \neq j, i, j \in M} b_{ij}(x)}{b_{ij}} \quad (10)$$

where:

$b_{ij}(x)$ — the number of shortest paths from i to j that pass through x ;

b_{ij} — the number of all shortest path between i and j ;

m — the number of nodes in a network. If a member obtains high value of BC then it means that he/she is the node without which the network will split into subnetworks.

3. GENERAL FEATURES OF USER POSITION MEASURES

The existing centrality measures can be divided based on the way in which they are calculated into three main groups: degree, shortest paths and rank group (Table 1). The first group takes into consideration only the measures that take into account only the degree of a given node and thus they are local measures. Measures based on the calculation of the shortest paths between vertexes are global ones as they calculate the length of the paths between each pair of the nodes. In the last, third group first the position is calculated according to one of the measures from the first or second group. After that for each node the positions of nodes that communicate with a given one are aggregated. Note also that all presented groups consist of structural indices the same as the node position.

The existing centrality and prestige measures suffer from many shortcomings (Table 3) that result in the need for developing a new methods that could cope of these disadvantages. The most important ones are enumerated below together with the short notion about how the developed node position method will overcome most presented issues.

- **Multirelational networks** — the networks that can consist of more than one type of relationship have not

been yet studied using the methods described above [37]. However, this problem can be no more neglected as most of the networks in the Internet consist of more than one type of relationship [24], [30].

- **Weighted Networks** — none of the enumerated measures deals with the problem of weighted networks. Of course, it is possible to include the strength of the relationship in the calculations, however the issue of assessing the connection strength is not considered. Thus, there is a need to provide the comprehensive method that enables to evaluate the strength of the relationship between user x and y . This strength can be also called commitment from user x to y . The node position method includes the description of evaluating the commitment function in both one- and multirelational networks of users. Additionally, the proposition of calculating the commitment function that also takes into account the time factor is presented.
- **Disconnected Networks** — all methods based on the shortest paths and in consequence eigenvector measures that use as an input one of these methods cannot be applied in the disconnected graphs that exist in the Internet. The calculation of these measures values for a disconnected graph gives as the outcome the values zero for each node. The node position method can be calculated for the disconnected graphs.
- **Complexity of the methods and diversification of measures values** — most of the presented methods are very inefficient when applying them in complex networks which constitute big part of the networks existing in the Internet. For example the calculation of the shortest paths within the large networks is a very time consuming task. Only the degree-based measures are effective. However, these methods tend to diversify users only to limited extent. The performed experiments have revealed that over 95% users obtain the same degree prestige and centrality (see Sec. 4.3). Thus, a criterion that should be met by a new method is to provide the mechanism that enables to make a trade-off between the accuracy of the calculations and the time needed to perform them.
- **Centrality or Prestige** — beside degree centrality, all of the presented methods are strictly dedicated to one of the application, i.e. calculating prestige or centrality of a given node. Although, node position method is the measure that denotes rather the prestige of a node, it can be easily transform to the measure that expresses the centrality of a node.

4. COMPARATIVE EXPERIMENTS

The goal of the last part of the efficiency tests is to compare the processing time of the proposed node position measure to other indexes that serve to assess the position of the user within the network of users.

4.1 Test Environment

The experiments were carried out on real dataset extracted from email communication at Wrocław University of Technology (WUT). It contained logs collected by the university mail server and referred only to the emails incoming to

Table 1: Groups of the centrality indices

Group Name	Name of Centrality Indices	Complexity
Node Degree	Indegree Centrality, Outdegree Centrality	Local measure — considers only the number of the first-level relationships [8] – Does not take into account the <i>ODC</i> of other nodes
Shortest Paths	Eccentricity Centrality, Closeness Centrality, Betweenness Centrality, Proximity Prestige	– Global measure — calculates the length of shortest paths to all nodes – Does not take into account the positions of other nodes
Rank of Actor	Rank Prestige, Node Position	– Global measure — takes into account the positions of other nodes

Table 2: Comparison of User Position Measures

Name	Properties	Example of Application
<i>IDC</i>	– Extracting the most visible actors in the network	– When we are interested in nodes that have the most direct votes – Whenever the graph represents something like a voting results (static situation) [8]
<i>PP</i>	– Extracting users who are reachable by the biggest number of network members – Global measure – It is inversely related to distance between nodes	– Identifying members who are perceived by others as important users
<i>RP, NP</i>	– Global measure – Considers the positions of other nodes	– Identifying members who are perceived by others as important users
<i>ODC</i>	– Measure of the "activity" – High outdegree denotes "where the action is" in the network – Extracting the most visible actors in the network	– When estimating how open is a user in contacts with other people
<i>EC</i>	– Global measure – Determine a node that minimizes the maximum distance to any other node in the network [8]	– Facility location problem that uses min-max criterion [19], e.g. determining the location for an emergency facility
<i>CC</i>	– Node with high <i>CC</i> can be very productive in communicating information to others – Actors with high <i>CC</i> need not rely on others to receive information [2] – It is inversely related to distance between nodes	– Whenever we are interested in people who can spread the information about e.g. special offers – Minimum location problem [19] that minimize the total travel, e.g. determining the location for a shopping mall where the total distance to all customers in the region are minimal
<i>BC</i>	– Express the extent to which a given node controls the communication between two nodes	– Finding users who control the information transfer between other members

Table 3: Advantages and disadvantages of Centrality Indices

Name	Advantages	Disadvantages
<i>IDC</i>	<ul style="list-style-type: none"> – Simple and easy to compute – Quite informative in many applications [37] 	<ul style="list-style-type: none"> – Big number of Duplicates (Sec. 4.3) – Local measure – takes into account only first-level neighborhood [8] – Lack of applicability in multirelational weighted networks
<i>PP</i>	<ul style="list-style-type: none"> – Global measure – Consider the whole topological structure of a network 	<ul style="list-style-type: none"> – The disconnected graph results in value 0 of <i>PP</i> for all nodes – Even if the graph is connected a given node can be not reachable by one of the rest nodes and this results in not relevant outcomes – Very complex and inefficient in large networks – Lack of applicability in multirelational weighted networks
<i>RP</i>	<ul style="list-style-type: none"> – Global measure – Consider the position of other nodes 	<ul style="list-style-type: none"> – The disconnected graph results in value 0 for all nodes in measures based on geodesic distance – Even if the graph is connected a given node can be not reachable by one of the rest nodes and this results in not relevant outcomes – Very complex and inefficient in large networks — especially measures based on geodesic distance – Lack of applicability in multirelational weighted networks
<i>NP</i>	<ul style="list-style-type: none"> – Global measure – Consider the position of other nodes – The need for adjusting the parameters – Small number of duplicates – Can be calculated in the case of weighted, directed and disconnected graphs 	<ul style="list-style-type: none"> – The need for adjusting the parameters – Can be time-consuming when the high ϵ is picked
<i>ODC</i>	<ul style="list-style-type: none"> – Simple and easy to compute – Quite informative in many applications [37] 	<ul style="list-style-type: none"> – Big number of Duplicates (Sec. 4.3) – Local measure – takes into account only first-level neighborhood [8] – Lack of applicability in multirelational weighted networks
<i>EC</i>	<ul style="list-style-type: none"> – Global measure – Consider the whole topological structure of a network 	<ul style="list-style-type: none"> – The disconnected graph results in value 0 of <i>EC</i> for all nodes – Even if the graph is connected a given node can be not reachable by one of the rest nodes and this results in not relevant outcomes – Very complex and inefficient in large networks – Lack of applicability in multirelational weighted networks
<i>CC</i>	<ul style="list-style-type: none"> – Global measure – Consider the whole topological structure of a network 	<ul style="list-style-type: none"> – The disconnected graph results in value 0 of <i>CC</i> for all nodes – Even if the graph is connected a given node can be not reachable by one of the rest nodes and this results in <i>CC</i> is not defined for a given pair of nodes [8] – Very complex and inefficient in large networks – Lack of applicability in multirelational weighted networks
<i>BC</i>	<ul style="list-style-type: none"> – Global measure – Consider the whole topological structure of a network 	<ul style="list-style-type: none"> – The disconnected graph results in value 0 of <i>BC</i> for all nodes – Even if the graph is connected a given node can be not reachable by one of the rest nodes and this results in not relevant outcomes – Not recommended to use in directed graphs – Is not very stable in dynamic graphs [12] – Very complex and inefficient in large networks – Lack of applicability in multirelational weighted networks

the staff members as well as entire organizational units registered at the university. Some other experiments on this dataset were presented in [23] and [25].

Table 4: The statistical information for the WUT datasets

Emails before cleansing	8,052,227
Period (after cleansing)	02.2006-09.2007
Emails after cleansing	8,052,227
External emails (sender or recipient outside the WUT domain)	5,252,279
Internal emails (sender and recipient from the WUT domain)	2,799,948
Cleansed email addresses	165,634
Isolated users	0
Cleansed email addresses from the WUT domain without isolated members; final set of the social network members	5,845
Emails per user	479
Network users with no activity	26 (0.45%)
Weighted relationships extracted from emails	149,344
Relationships after application of Eq. (5)	176,504
Relationships per user	30.2
Percentage of all possible relationships	0.517%

First, the data has to be cleansed by removal of bad and unification of duplicated email addresses. Additionally, only emails from and to the WUT domain were left. After data preparation the commitment function $C(x \rightarrow y)$ (strength of the relationship) is evaluated for each pair of members. Every email with more than one recipient was treated as $1/n$ of a regular email, where n is the number of its recipients. It was the extension of Eq. (6). To evaluate approximate values of node position measure, the initial values of this measure were established to 1 for all members and the stop condition $\tau = 0.00001$ was applied separately for each user.

The general statistics related to the processed datasets are presented in Tab. 4.

4.2 Efficiency Tests

The measures that are taken into consideration are indegree centrality (IDC), outdegree centrality (ODC), closeness centrality (CC), eccentricity centrality (EC) and (NP). The reason for choosing these method is that they represent two different presented groups of centrality indices. IDC and ODC are degree-based centralities whereas EC and CC are the indices developed based on the concept of geodesic distances between nodes. Note that EC and CC are the least complex measures from this group. It means, that if this measure will be less effective than NP method then the rest of them will also need more time than NP to be calculated. Note also, that the complexity of the algorithms for assessing the centrality based on eigenvector concept, i.e. all rank prestige measures, is the same as for the measures based

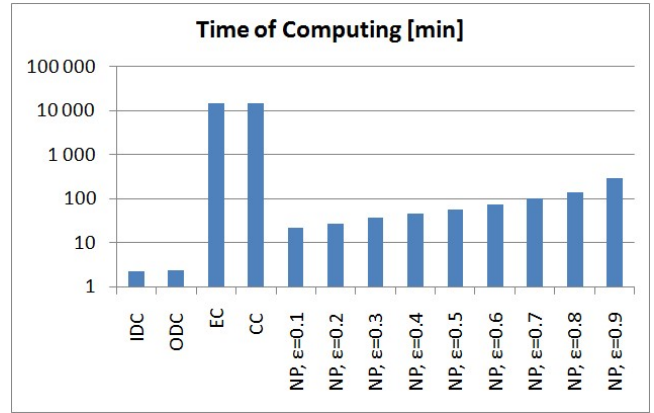


Figure 2: Processing time for different user position measures

on which the input values for eigenvector method were calculated. For example, in the case of eigenvector closeness centrality (CCE) the complexity will be the same as in the case of CC . The outcomes of the experiments are presented

Table 5: Processing time for different user position measures

Measure	Processing Time [min]
IDC	2.17
ODC	2.28
EC	14891.70
CC	14746.86
$NP, \varepsilon = 0.1$	21.81
$NP, \varepsilon = 0.2$	27.50
$NP, \varepsilon = 0.3$	36.49
$NP, \varepsilon = 0.4$	45.36
$NP, \varepsilon = 0.5$	57.04
$NP, \varepsilon = 0.6$	71.71
$NP, \varepsilon = 0.7$	98.62
$NP, \varepsilon = 0.8$	142.63
$NP, \varepsilon = 0.9$	287.30

in Table 5 and Fig. 2. Analysis of processing time of the algorithms that were used to calculate the specific centrality measures reveal that the most efficient measures are the degree-based measures as they are 10 times faster than NP and even 6860 times faster than the measures based on the geodesic distance between users. However, the fact that should be emphasized is that indegree and outdegree centralities take into consideration only the first level neighbors whereas the node position of user depends on the positions of all users within the network. This results in the node position measure is much more diverse than indegree centrality indexes.

NP method is approximately 1470 times faster than measures based on concept of shortest paths. In contrary to measures based on geodesic distance, NP method offers an acceptable in large networks processing time (21.81 [min] for $\varepsilon = 0.1$ and $\tau = 0.000001$ what results in 12.7% of duplicates). In the measures based on geodesic distance the processing time is unacceptable in large networks of Internet users. The analyzed WUT network consists of 5945 users and it is not a big number when comparing for example to

telecommunication networks or different social networking sites or multimedia sharing systems where we can have millions of users.

Another problem when analyzing the processing time of different centrality indices are the structural changes of the network. Note that if a new node joins or leaves the network or a new relationship is established then all methods based on the geodesic distance and eigenvector methods which as an input use methods based on the geodesic distance require the recalculation of relationship strengths and of all shortest paths. This results in these methods become very time consuming. On the other hand, in the case of *NP* the new node in the network causes only that the commitment function values need to be calculated for this node and for nodes that communicate with an added node.

Some other analysis dedicated to different algorithms for user position calculation are presented in [11]

4.3 Analysis of Duplicates

In the last part of the experiments on WUT dataset a comparison between number of duplicates in different centrality indices is presented (Table 6, Fig. 3). The stopping condition for *NP* measure was set to $\tau = 0.000001$. This results in the smallest number of duplicates but on the other hand in the longest processing time in comparison to larger τ values.

Table 6: Number of duplicates for different centrality indices

Measure	Number of Duplicates	Percentage of Duplicates [%]
<i>IDC</i>	5737	96.50
<i>ODC</i>	5703	95.93
<i>EC</i>	1943	32.68
<i>CC</i>	1888	31.76
<i>NP</i> , $\varepsilon = 0.1$	755	12.70
<i>NP</i> , $\varepsilon = 0.2$	631	10.61
<i>NP</i> , $\varepsilon = 0.3$	569	9.57
<i>NP</i> , $\varepsilon = 0.4$	545	9.17
<i>NP</i> , $\varepsilon = 0.5$	549	9.23
<i>NP</i> , $\varepsilon = 0.6$	520	8.75
<i>NP</i> , $\varepsilon = 0.7$	531	8.93
<i>NP</i> , $\varepsilon = 0.8$	521	8.76
<i>NP</i> , $\varepsilon = 0.9$	540	9.08

The biggest number of the users that obtain the same centrality values is in the case of degree measures and exceeds 95% of users. When analyzing the measures based on the shortest paths the number of duplicates is around 30%. This number is smaller than in the case of degree measures but still 3 times bigger than in case node position measure.

It is worth to notice that *NP* as the only one distinguishes people very good (around 10% of duplicates) and in the same time offers an acceptable in large networks processing time (21.81 [min] for $\varepsilon = 0.1$ and $\tau = 0.000001$). In contrary, the degree measures offer low computational cost but very ineffective way of distinguishing users within the network — more than 95% of duplicates. Finally, in the measures based on geodesic distance the duplicates percentage is at

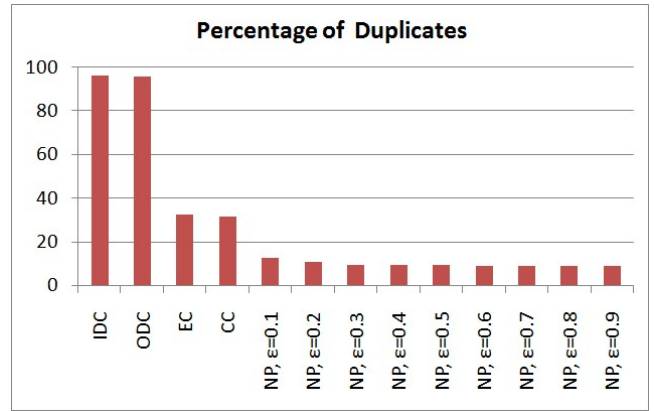


Figure 3: Percentage of duplicates for different user position measures

the level of 30% but the processing time is unacceptable in large networks of Internet users. The analyzed WUT network consists of 5845 users (Tab. 4) and it may not be treated as large when comparing for example to telecommunication networks or different social networking sites or multimedia sharing systems, in which we can have millions of users [32].

5. CONCLUSIONS

Most user position measures used in social network analysis come from other sciences that operate on graphs and networks. However, the discovery of the position of users among other network members, only based on computer data still remains an interesting research domain.

The main contribution of this paper is the analysis of different centrality and prestige measures with respect to their general profile: advantages and disadvantages. Some measures are more general — contain more information about the node's neighborhood or the entire network (e.g. those based on shortest paths or rank of nodes) whereas the others reflect rather local character of the node (e.g. measures based on simple computing of 1-level incoming or outgoing relationships), see Tab. 1.

Additionally, according to the experiments carried out calculation of some measures like eccentricity or closeness centrality as well as node position with greater precision may be absolutely ineffective for larger social networks, see Sec. 4.2. In such cases, node position measure has an advantage due to its flexibility; calculation time can be in some extent tailored by usage larger stop condition and a result the less number of iterations.

Global measures are simultaneously more diverse compared to local indices and provide much less duplicates, see Sec. 4.3.

6. ACKNOWLEDGMENT

The work was supported by The Polish Ministry of Science and Higher Education, grant no. N N516 264935.

7. REFERENCES

- [1] Alexander C.N., "A method for processing sociometric data", *Sociometry* 26, pp. 268–269, 1963.
- [2] Bavelas A., "Communication patterns in task – oriented groups", *Journal of the Acoustical Society of America* 22, pp. 271–282, 1950.
- [3] Beauchamp, M.A. "An improved index of centrality", *Behavioral Science* 10, pp.161–163, 1973.
- [4] Berkhin P., "A Survey on PageRank Computing", *Internet Mathematics* 2(1), pp. 73–120, 2005.
- [5] Bonacich P., "Power and centrality: a family of measures", *American Journal of Sociology* 92, pp. 1170–1182, 1987.
- [6] Bonacich P., Lloyd P., "Eigenvector-like Measures of Centrality for Asymetric Relations", *Social Networks* 23(3), pp. 191–201, 2001.
- [7] Botafogo R.A., Rivlin E., Shneiderman B., "Structural analysis of hypertexts: identifying hierarchies and useful metrics", *ACM Transaction on Information Systems* 10(2), pp. 142–180, 1992.
- [8] Brandes U., Erlebach T., "Network Analysis, Methodological Foundations", *Springer – Verlag, Berlin, Heidelberg, Germany*, 2005.
- [9] Brin S., Page L., "The Anatomy of a Large-Scale Hypertextual Web Search Engine", *Computer Networks and ISDN Systems* 30(1-7), pp. 107–117, 1998.
- [10] Brinkmeier M., "PageRank Revisited", *ACM Transactions on Internet Technology* 6(3), pp. 282–301, 2006.
- [11] Bródka P., Musiał K., Kazienko P., "A Performance of Centrality Calculation in Social Networks", *International Conference on Computational Aspects of Social Networks, CASoN 2009, June 24-27, 2009, Fontainebleau, France, IEEE Computer Society*, 2009.
- [12] Carpenter T., Karakostas G., Shallcross D., "Practical Issues and Algorithms for Analyzing Terrorist Networks", *Invited talk at WMC 2002*, 2002.
- [13] Carrington P., Scott J., Wasserman S., "Models and methods in Social Network Analysis", *Cambridge University Press, Cambridge*, 2005.
- [14] Degenne A., Forse M., "Introducing social networks", *London: SAGE Publications Ltd*, 1999.
- [15] Dijkstra E.W., "A note on two problems in connection with graphs", *Numerische Mathematik* 1, pp. 269–271, 1959.
- [16] Freeman L.C., "A set of measures of centrality based on betweenness", *Sociometry* 40, pp.35–41, 1977.
- [17] Freeman L.C., Roeder D., Mulholland R.R., "Centrality in Social Networks: II. Experimental Results", *Social Networks* 2(2), pp.119–141, 1980.
- [18] Friedkin N.E., "Theoretical foundations for centrality measures", *American Journal of Sociology* 96, pp.1478–1504, 1991.
- [19] Hakimi S.L., "Optimum location of switching centers and the absolute centers and medians of a graph", *Operations Research* 12, pp.450–459, 1964.
- [20] Hubbell C.H., "An input–output approach to clique detection", *Sociometry* 28, pp. 277–299, 1965.
- [21] Katz L., "A new status derived from sociometrics analysis", *Psychometrika* 18, pp.39–43, 1953.
- [22] Kazienko P., "Associations: Discovery, Analysis and Applications", *Oficyna Wydawnicza Politechniki Wroclawskiej, Wroclaw*, 2008.
- [23] Kazienko P., Musiał K., "Mining Personal Social Features in the Community of Email Users", *SOFSEM 2008, The 34th International Conference on Current Trends in Theory and Practice of Computer Science, Nový Smokovec, Slovakia, January 19-25, 2008, Springer Verlag, Lecture Notes in Computer Science LNCS 4910*, pp. 708–719, 2008.
- [24] Kazienko P., Musiał K., Kajdanowicz T., "Multidimensional Social Network in the Social Recommender System", *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 2009, accepted for publication.
- [25] Kazienko P., Musiał K., Zgrzywa A., "Evaluation of Node Position Based on Email Communication", *Control and Cybernetics*, 38(1), 2009, in press.
- [26] Knoke D., Burt R.S., "Prominence", In *Burt R.S. and Minor M.J. (eds.) Applied Network Analysis, Newbury Park, CA: Sage*, pp. 195–222, 1983.
- [27] Krebs V., "The Social Life of Routers", *Internet Protocol Journal* 3, pp. 14–25, 2000.
- [28] Kumar R., Raghavan P., Rajagopalan S., Tomkins A., "The Web and Social Networks", *IEEE Computer* 35(11), pp. 32–36, 2002.
- [29] Moreno J.L., "Who shall survive?: Foundations of Sociometry, Group Psychotherapy, and Sociodrama", *Washington, D.C.: Nervous and Mental Disease Publishing Co. Reprinted in 1953 (2nd Edition) and in 1978 (3rd Edition) by Beacon House, Inc., Beacon, New York*, 1934.
- [30] Musiał K., Kazienko P., Kajdanowicz T., "Multirelational Social Networks in Multimedia Sharing Systems", *Chapter 18 in N.T.Nguyen, G.Kolaczek, B.Gabrys (eds.) Knowledge Processing and Reasoning for Information Society, Academic Publishing House EXIT, Warszawa*, pp. 275–292, 2008.
- [31] Proctor C.H., Loomis C.P., "Analysis of sociometric data", *Research Methods in Social Relations, M. Jahoda, M. Deutch, S.W. Cok (eds.), Dryden Press, NewYork*, pp. 561–586, 1951.
- [32] Ruta D., Kazienko P., Bródka P., "Network-Aware Customer Value in Telecommunication Social Networks", *The 2009 International Conference on Artificial Intelligence, ICAI'09, July 13-16, 2009, Las Vegas, Nevada, USA*, 2009, to appear.
- [33] Sabidussi G., "The centrality index of a graph", *Psychometrika* 31(4), 1966.
- [34] Scott J., "Social network analysis: A handbook (2nd ed.)", *London: Sage*, 2000.
- [35] Shaw M.E., "Group structure and the behavior of individuals in small groups", *Journal of Psychology* 38, pp. 139–149, 1954.
- [36] Tutzau F., "Entropy as a measure of centrality in networks characterized by path-transfer flow", *Social Networks* 29(2), pp. 249–265, 2007.
- [37] Wasserman S., Faust K., "Social network analysis: Methods and applications", *New York: Cambridge University Press*, 1994.