Control and Cybernetics

vol. **XX** (XXXX) No. X

Evaluation of Node Position Based on Email Communication

by

Przemysław Kazienko[1], Katarzyna Musiał[1] and Aleksander Zgrzywa[1]

[1] Wroclaw University of Technology, Intitute of Applied Informatics
Wyb. Wyspianskiego 27, 50-370 Wroclaw, Poland

Abstract: The rapid development of various kinds of social networks within the Internet has enabled to investigate their properties and analyze their structure. An interesting scientific problem in this domain is the assessment of the node position within the directed, weighted graph that represents the social network of email users.

The new method of node position analysis, which takes into account both the node positions of the neighbors as well as the strength of the connections between network nodes, is presented in the paper. The node position can be used to discover key network users, who are the most important in the population and who have potentially the greatest influence on others. The experiments carried out on two datasets enabled to study main properties of the new measure.

Keywords: node position assessment, social network of email users, social network analysis, network analysis, centrality measure.

1. Introduction

The various kinds of e–commerce and e–business solutions that exist in the market encouraged the users to utilize the Internet and available web–based services more willingly in their everyday life. Many customers look for services and goods that have high quality. Thus, not only the information provided by vendors is important for potential customers but also the opinions of other users who have already bought the goods or used the particular service. It is natural that users, to gather other people opinions, communicate with each other via different communication channels, e.g. by exchanging emails, commenting on forums, using instant messengers, etc. This information flow from one individual to another is the basis for the social network of users. This network can be represented as a directed graph, in which nodes are the users and the edges describe the information flow from one user to another. One of the most meaningful and useful issue in social network analysis is the evaluation of the

node position within the network. Since the social network describes the interactions between people, the problem of assessment the node position becomes very complex because humans with their spontaneous and social behavior are hard predictable. However, the effort should be made to evaluate their status because such analysis would help to find users who are the most influential among community members, possess the highest social statement and probably the highest level of trust (Golbeck, Hendler, 2004), (Rana, Hinze, 2004). These users can be representatives of the entire community. A small group of key persons can initiate new kinds of actions, spread new services or activate other network members (Kazienko, Musiał, 2007). On the other hand, users with the lowest position should be stimulated for greater activity or be treated as the mass, target receivers for the prior prepared services that do not require the high level of involvement. In this paper only the community of email users, called the social network of email users (*ESN*), is considered. In order to calculate the position of the email user, the new measure called node position is introduced in the further sections. It enables to estimate how valuable the particular node is within *ESN*. In contrary to the PageRank algorithm that is designed to assess the importance of the web pages, the presented node position measure take into account not only the significance of the direct connections of a node but also the quality of the connection.

2. Related Work

The main concept of a regular social network appears to be simple as it can be described as a finite set of nodes that are linked with one or more edges (Garton, Haythorntwaite, Wellman, 1997), (Hanneman, Riddle, 2005), (Wasserman, Faust, 1994). A node of the network is usually called an actor, an individual, corporate, collective social unit (Wasserman, Faust, 1994), or customer (Yang, Dia, Cheng, Lin, 2006) whereas an edge named also a tie or relationship, as a linkage between a pair of nodes (Wasserman, Faust, 1994). The range and type of the edge can be extensive and different depending on the type and character of the analyzed actors (Hanneman, Riddle, 2005), (Wasserman, Faust, 1994).

The notation that is widely used to represent a social network is the graph. Nodes of the graph are actors while edges correspond to the relations in the social network (Wasserman, Faust, 1994).

The social networks of internet users somewhat differ from the regular ones and because of that they yield for new approaches to their definition and analysis. This kind of social networks is also called an online social network (Garton, Haythorntwaite, Wellman, 1997), computer-supported social network (Wellman, Salaff, 1996), web community (Gibson, Kleinberg, Raghavan, 1998), (Flake, Lawrence, Lee Giles, 2000), or web-based social network (Golbeck, 2005). Note that there is no one coherent definition of social network in the Internet including email-based social network. Some researchers claim that a web community is also built on the set of web pages relevant to the same, common topic (Gib-

son, Kleinberg, Raghavan, 1998), (Flake, Lawrence, Lee Giles, 2000). Adamic and Adar argue that a web page must be related to the physical individual in order to be treated as a node in the online social network. Thus, they analyze the links between users' homepages and form a virtual community based on this data. Additionally, the equivalent social network can also be created from email communication (Adamic, Adar, 2003), (Culotta, Bekkerman, McCallum, 2004), (Shetty, Adibi, 2005). Others declare that computer-supported social network appears when a computer network connects people or organizations (Garton, Haythorntwaite, Wellman, 1997), (Wellman, Salaff, 1996). On the other hand, Golbeck asserts the view that a web-based social network must fulfil the following criteria: users must explicitly establish their relationships with others, the system must have support for making connections, relationships must be visible and browsable (Golbeck, 2005). Boyd created a social network from people who meet each other using an online system Friendster (Boyd, 2004). Yet another multirelational social network can be established within the multimedia sharing system like Flickr (Musiał, Kazienko, Kajdanowicz, 2008).

Social network analysis (Wasserman, Faust, 1994) provides some measures useful to assess the node position within the social network. To the most commonly used belong: centrality, prestige, reachability, and connectivity (Hanneman, Riddle, 2005), (Wasserman, Faust, 1994). There exist many approaches to evaluation of person centrality (Freeman, 1979): degree centrality, closeness centrality, and betweeness centrality. Degree centrality $DC(x)$ takes into account the number of neighbors $o(x)$ that are directly adjacent from person x (Hanneman, Riddle, 2005), as follows: $DC(x) = \frac{o(x)}{m-1}$, where m – the number of users within the network. The closeness centrality $CC(x)$ pinpoints how close an individual x is to all the others within the network (Bavelas, 1950). It depends on the shortest paths $d(x, y_i)$ from user x to other people y_i in the following way: $CC(x) = \frac{m-1}{\sum_{i=1}^{m} d(x,y_i)}$. The similar idea was studied for hypertext systems (Botafogo, Rivlin, Shneiderman, 1992). Finally, the betweeness centrality $BC(x)$ of member x specifies to what extend member x is between other members in the social network (Freeman, 1979). Member x is more important (in-between) if there are many people in the network whose relationships with other network members must go through x (Hanneman, Riddle, 2005). The second feature that characterizes an individual in the social network and enables to identify the most powerful members is prestige. It also can be calculated in various ways, e.g. degree prestige, proximity prestige, and rank prestige. The degree prestige $DP(x)$ takes into account the number of users $i(x)$ that are adjacent to x (Wasserman, Faust, 1994), as follows: $DP(x) = \frac{i(x)}{m-1}$. Proximity prestige $PP(x)$ shows how close are all the other users to user x (Wasserman, Faust, 1994). The rank prestige $RP(x)$ (Wasserman, Faust, 1994), is measured based on the status of users in the network and depends not only on geodesic distance and number of relationships, but also on the status of users connected with the user (Katz, 1953). Another popular measures used for internet analysis

Figure 1. Two social networks of email users

is PageRank introduced by Brin and Page to assess the importance of web pages (Berkhin, 2005), (Brin, Page, 1998) , (Brinkmeier, 2006). 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. The main difference between PageRank and node position proposed in this paper is the existence and meaning of commitment function. In PageRank, all links have the same weight and importance whereas node position makes the quantitative distinction between the strengths of individual relationships.

3. Evaluation of Node Position Based on Email Communication

Before the new method for node position measurement will be presented the definition of social network of the email users should be established.

3.1. Social Network of Email Users

Numerous and inconsistent definitions of the social networks (see Sec. 2) yields for the creation of one consistent approach dedicated for the network of email users.

DEFINITION 1. *An email social network is a tuple ESN=(EA,R), where EA is a finite set of registered email addresses i.e. the digital representation of a person, organizational unit, group of people, or other social entity, that communicate with one another using email system. R is a finite set of social relationships that join pairs of distinct email addresses: R:EA×EA, i.e.* $R = \{(ea_i, ea_j) : ea_i \in EA, ea_j \in EA, i \neq j\}$. *The set of email addresses EA must not contain isolated members – with no relationships and card(EA)*> 1.

Note that $(ea_i, ea_j) \neq (ea_j, ea_i)$. The example of two separate social network of email users is presented in Fig. 1. An individual human can simultaneously belong to many email–based social networks. Moreover, they can also maintain several email addresses, even in the same email server — see user Bob in Fig. 1. The email address is a digital representation of the physical social entity. These are objects that can be unambiguously ascribed to one person (individual email address), to a group of people or an organization (group email address). This representation must explicitly identify the social entity (a user, group of users or an organization). This mapping enables to define the connections between social entities based on the relationships between their email addresses. An individual email address possesses individuals, whereas a group email address corresponds to a group of people, e.g. family that use only one email account, as

Figure 2. Two fragments of an email social network. The size of the email address node corresponds to the value of its node position. The arrows reflect commitment values. $\varepsilon \approx 1$

well as to an organization, e.g. all employees use one email account to respond customers' requests. Such group email addresses can by identified with the certain probability by email content analysis.

3.2. Node Position Measure

Based on the data derived from the source email system, we can build a graph that represents the connections between users and then analyze the position of each node within such network. Nodes of the graph represent the email users – addresses while edges correspond to the relationships extracted from the data about their common communication or activities.

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:

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

where: ε – the constant coefficient from the range $[0, 1]$; y_1,...,y_m – acquaintances of x, i.e. nodes that are in the direct relation to x; m – the number of x's acquaintances; $C(y_1 \rightarrow x)$,...,$C(y_m \rightarrow x)$ – the function that denotes the contribution in activity of y_1,...,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 user x is inherited from the others but the level of inheritance depends on the activity of the users directed to this person, i.e. intensity of mutual communication. Thus, the node position depends both on the number and quality of relationships. A user can possess the high node position if some other people transfer their high NP to them. For example, the node position of user ea_3 in Fig. 2a is 0.9 mostly come from ea_3's high commitment in the activities of user ea_4, $C(ea_4 \rightarrow ea_3) = 0.6$ and $C(ea_4 \rightarrow ea_3) * NP(ea_4)$ equals as much as 0.54. The contribution of two other users ea_1 and ea_2 in $NP(ea_3)$

is only 0.36, even though their commitment values are the greatest possible $C(ea_1 \rightarrow ea_3) = C(ea_2 \rightarrow ea_3) = 1$. On the other hand, despite the very high $NP(ea_3)$, the value of $NP(ea_1)$ is only 0.09 due to very low ea_1's participation in ea_3's activity, $C(ea_3 \rightarrow ea_1) = 0.1$. User ea_3 is the only one who sends emails to user ea_1. The node position of user ea_6 is medium-sized: $NP(ea_6) = 0.4$, although three other persons ea_5, ea_7, and ea_8 pass most of their activities to ea_6: $C(ea_5 \rightarrow ea_6) = 0.8$, $C(ea_7 \rightarrow ea_6) = 0.9$, and $C(ea_8 \rightarrow ea_6) = 1$, Fig. 2b. It results from the low or very low node position of ea_6's acquaintances: $NP(ea_5) = 0.25$, $NP(ea_8) = 0.2$ and $NP(ea_7)$ is almost zero. Hence, $NP(ea_3)$ is high because of high $NP(ea_4)$ as well as big $C(ea_1 \rightarrow ea_3)$ and $C(ea_2 \rightarrow ea_3)$; $NP(ea_1)$ is low due to small $C(ea_3 \rightarrow ea_1)$; and $NP(ea_6)$ is medium with respect to the low importance of its neighbors.

3.3. Node Position Evaluation

To calculate the node position of the person within the social network the convergent, iterative algorithm is used. First, the initial node positions $NP^{(0)}(x)$ are assigned to every node x in the network $ESN(EA, R)$. Next, values of $NP^{(k)}(x)$ are iteratively calculated based on previous node positions of other users $y \in EA$, i.e. $NP^{(k-1)}(y)$, $k = 1, 2, \cdots$. 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 user $x \in EA$. Alternatively, threshold τ may concern the differences in sums of all users' node positions instead of individual users.

3.4. Commitment Function

The commitment function $C(y \rightarrow x)$ is a very important element in the process of node position assessment, thus it needs to be explained more detailed. $C(y \rightarrow x)$ 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 $ESN(EA,R)$ must satisfy the following set of criteria:

1. The value of commitment is from the range $[0; 1]$: $\forall(x, y \in EA)\ C(y \rightarrow x) \in [0; 1]$.
2. The sum of all commitments has to equal 1, separately for each node of the network:
$$\forall(x \in EA) \sum_{x \in EA} C(x \rightarrow y) = 1 \tag{2}$$
3. If there is no relationship from y to x then $C(y \rightarrow x) = 0$.

Figure 3. Example of the social network of email users with the assigned commitment values

4. If a member y is not active to anybody and other n members x_i, $i = 1,...,n$ 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.
$$\forall(x_i \in EA)\ C(y \to x_i) = 1/n \tag{3}$$

Since the relationships are not reflexive (see Definition 1) and with respect to criterion 3, the commitment function to itself equals 0: $\forall(x \in EA)C(x \to x) = 0$.

The example of network of email users with values of commitment function assigned to every edge is presented in Fig. 3. 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 user of the network (criterion 2). Moreover, there is no relationship ea_2 to ea_1 so $C(ea_2 \to ea_1) = 0$ (criterion 3). Note also that node ea_3 is not active to anybody but two others ea_2 and ea_4 are active to ea_3, so according to condition 4, the commitment of ea_3 is equally distributed among all ea_3's connections $C(ea_3 \to ea_2) = C(ea_3 \to ea_4) = 1/2$.

The commitment function $C(x \to y)$ 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:

$$C(x \to y) = \frac{A(x \to y)}{\sum_{j=1}^{m} A(x \to y_j)} \tag{4}$$

where: $A(x \to y)$ – 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 *ESN*. In the above formula the time is not considered. The similar approach is utilized by Valverde et al. to calculate the strength of relationships. It is established as the number of emails sent by one person to another person (Valverde, Theraulaz, Gautrais, Fourcassie, Sole, 2006). However, the authors do not respect the general activity of the given individual. In the proposed approach, this general, local activity exists in the form of denominator in Eq. (4).

Every single email provides information about the sender activity but unfortunately one email can be simultaneously sent to many recipients. An email sent to only one person reflects strong attention of the sender directed to this recipient. The same email sent to 20 people does not respect individual recipients so much. For that reason, the strength of email communication $S(x \to y)$ from x to y has been introduced:

$$S(x \to y) = \sum_{i=1}^{card(EM(x \to y))} \frac{1}{n_j(x \to y)} \tag{5}$$

where: $EM(x \to y)$ – the set of all email messages sent by x to y; $n_j(x \to y)$ – the number of all recipients of the ith email sent from x to y.

Based on the strength of the email communication from one user to another the commitment $C(x \rightarrow y)$ from Eq. (4) can be redefine as follows:

$$C(x \rightarrow y) = \frac{S(x \rightarrow y)}{n(x)} \tag{6}$$

where: $n(x)$ – the total number of emails sent by user x.

In another version of commitment function $C(x \rightarrow y)$ all member's activities are considered with respect to the point of time when the emails were sent. The entire time from the first to the last activity of any member is divided into k periods. For instance, a single period can be a month. Activities (sent emails) in each period are considered separately for each individual:

$$C(x \rightarrow y) = \frac{\sum_{i=0}^{k-1} (\lambda)^i S_i(x \rightarrow y)}{\sum_{i=0}^{k-1} (\lambda)^i n_i(x)} \tag{7}$$

where: i – the index of the period: for the most recent period $i = 0$, for the previous one: $i = 1$, for the earliest $i = k - 1$; $S_i(x \rightarrow y)$ – the function that denotes the activity level of user x directed to user y in the ith time period, i.e. the strength of the email communication from x to y in the ith period — Eq. (5) restricted to only the ith period; $n_i(x)$ – the total number of emails sent by user x in the ith period; $(\lambda)^i$ – the exponential function that denotes the weight of the ith time period, $\lambda \in (0; 1]$; k – the number of time periods.

The activity of user x is calculated in every time period and after that the appropriate weights are assigned to the particular time periods, using $(\lambda)^i$ factor. The most recent period $(\lambda)^i = (\lambda)^0 = 1$, for the previous one $(\lambda)^i = (\lambda)^1 = (\lambda)$ is not greater than 1, and for the earliest period $(\lambda)^i = (\lambda)^{k-1}$ receives the smallest value. The in a sense similar idea was used in the personalized systems to weaken older activities of recent users (Kazienko, Adamski, 2007).

If user x sent many emails to y in comparison to the number of all x's sent emails, then y has the greater commitment within activities of x, i.e. based on Eq. (6) or (7), $C(x \rightarrow y)$ will have greater value. In consequence, the node position of user y will grow. Moreover, if user y is the only recipient of these emails then the node position of user y is even greater.

4. Experiments

4.1. Test Environment

The experiments that illustrate the idea of node position assessment were carried out separately on two datasets: Enron employees' mailboxes and Wrocław University of Technology (WUT) mail server logs. Enron Corporation was the biggest energy company in the USA. It employed around 21,000 people before its bankruptcy at the end of 2001. A number of other researches have been conducted on the Enron email dataset (Priebey, Conroy, Marchette, Park, 2005), (Shetty, Adibi, 2005). Some preliminary analyses on Enron dataset have been

presented in (Kazienko, Musiał, 2008) and (Kazienko, Musiał, Zgrzywa, 2007). The second dataset contains logs collected by the mail server of WUT and refers only to the emails incomming to the staff members as well as entire organizational units registered at the university. First, the data has to be cleansed

Table 1. The statistical information for the Enron and WUT datasets

Emails before cleansing	517,431	8,052,227
Period (after cleansing)	01.1999-07.2002	02.2006-09.2007
Emails after cleansing	411,869	8,052,227
External emails (sender or recipient outside the Enron/WUT domain)	120,180	5,252,279
Internal emails (sender and recipient from the Enron/WUT domain)	311,438	2,799,948
Cleansed email addresses	74,878	165,634
Isolated users	9,390	0
Cleansed email addresses from the Enron/WUT domain without isolated members; set EA in ESN=(EA,R)	20,750	5,845
Emails per user	15	479
Network users in EA with no activity	15,690 (76%)	26 (0.45%)
Commitments extracted from emails	201,580	149,344
Relationships after application of Eq. (3)	250,003	176,504
Relationships per user	12.0	30.2
Percentage of all possible relationships	0.0583%	0.517%

Figure 4. The number of necessary iterations and processing time in relation to ε

by removal of bad and unification of duplicated email addresses. Additionally, only emails from and to the Enron or WUT domain were left. Every email with more than one recipient was treated as $1/n$ of a regular email, where n is the number of its recipients, see Eq. (5). The general statistics related to the processed datasets are presented in Tab. 1.

After data preparation the commitment function is evaluated for each pair of members. To evaluate relationship commitment function $C(x \rightarrow y)$ two formulas (6) and (7) were used. Eq. (6) was utilized to calculate node position without respecting time (NP) whereas Eq. (7) serves to evaluate node position with time factor ($NPwTF$). The initial node positions for all members were

Figure 5. Average NP and $NPwTF$ as well as their standard deviations in the Enron dataset, calculated for different values of ε

established to 1 and the stop condition $\tau = 0.00001$ was applied separately for each user. The node positions without and with time coefficient were calculated for six, different values of the ε coefficient, i.e. $\varepsilon = 0.01$, $\varepsilon = 0.1$, $\varepsilon = 0.3$, $\varepsilon = 0.5$, $\varepsilon = 0.7$, $\varepsilon = 0.9$.

4.2. Iterative Data Processing

The conducted experiments revealed that the number of iterations necessary to calculate the node positions for all users tightly depends on the value of the parameter ε, see Eq. (1): the greater ε, the greater the number of iterations (Fig. 4). The number of iterations directly influences the processing time. Thus, much more time is required to fulfill the same stop condition $\tau = 0.00001$ for greater values of ε coefficient (Fig. 4). Obviously, both the processing time and the number of iterations also depend on precision level τ. The smaller value of τ, the more necessary calculations.

Efficiency of calculations can be essential in case of large social networks that contain many millions of nodes. Quantity of calculations can be reduced by application either the greater τ or the smaller ε. However, in both cases, it would happen at the expense of precision, see Sec. 3.3.

4.3. Distribution of Node Position Values

Figure 6. Average NP and its standard deviation in the WUT dataset in relation to ε

Some additional information about the values of node position provide the average node position as well as standard deviation evaluated for the entire email social network ESN. The analyses of node position values (NP) and node positions with time factor ($NPwTF$) for Enron can be found in Fig. 5, whereas statistics of node positions for WUT are placed in Fig. 6.

The average node position does not depend on the value of ε. It equals around 1 in all cases. This feature of node position measure, i.e. convergence of the average to 1 can be formally proved but it would require much reasoning.

On the other hand, the standard deviation substantially differs depending on the value of coefficient ε. The greater ε, the bigger standard deviation. It shows that for greater ε the value of the distance between the members' node positions increases and it is valid for NP and $NPwTF$ and for both datasets. It can be noticed that the value of node position NP for over 93% (Fig. 7) of email users in the Enron community as well as over 70% of users in WUT (Fig. 8)

Figure 7. The percentage of users with $NP < 1$ and $NP \geq 1$ within the Enron social network in relation to ε

is less than 1. It means that only few members exceed the average value that equals 1. The value $NP = 2$ is exceeded by only up to 13% for WUT (Fig. 8) and only up to 90 persons (0.43%) for Enron. This confirms that node position can be the good measure to extract key persons in the social network (Kazienko, Musiał, 2007).

If we analyze the person x in the WUT community, who gained most in the rankings based on NP values from $\varepsilon = 0.01$ to $\varepsilon = 0.9$, we would find out that x was moved 2242 positions from rank position 3347 to 1104. User x had as many as 19 incoming relations. Especially one of x's neighbors with the very high NP and ranking position from 95 for $\varepsilon = 0.01$ and 10 for $\varepsilon = 0.9$ had the relatively high commitment $C = 0.1$ towards user x. Besides, also the other x's neighbors received relatively high node position.

Figure 8. The percentage of users within the WUT social network whose node positions belong to one of three intervals: $NP<1$, $1 \leq NP<2$, and $NP \geq 2$ in relation to ε

4.4. Node Positions with Time Factor

Experiments on influence of time factor on the values of node position have been carried out only on Enron dataset. The number of users who benefited in their node position from the introduction of the time factor ($NPwTF$) is greater than the number of those who lost, Fig. 9. Moreover, this difference is greater for greater values of ε – up to more than 7 times in case of $\varepsilon = 0.9$. Furthermore, the highest gain in ranking for $\varepsilon = 0.9$ was only 2 positions whereas the maximum loss as many as 252 positions. The same tendency can be observed from the values of mean squared error between NP and $NPwTF$, Fig. 10. Overall, the

Figure 9. The percentage contribution of members with $NP \geq NPwTF$ and $NP < NPwTF$ within the Enron social network in relation to ε

greater number of users for whose $NPwTF > NP$ comes from the profile of the Enron dataset. Most users (76%) were not active at all, Tab. 1. Moreover, the majority of the active users was active for the almost entire considered period. That is why most users gain but only a few whereas the minority lost much. This minority were users who received emails only at the beginning of the considered period.

Node positions with time factor $NPwTF$ are more diverse compared to those without time factor – NP for greater ε and less diverse for smaller ε, see standard deviation values in Fig. 5.

Note that node position NP denotes the general position of a node regardless of time. Hence, node position $NP(x)$ for person x who received n emails from y three years ago (only y communicated to x) will be the same as $NP(z)$ for user z who also received n emails from y and only from y but all last month. In case of node position with time factor, $NPwTF(x)$ will be significantly lower than $NPwTF(z)$, because the weight assigned to the earlier period will be lower than the weight assigned to the latest period, see factor $(\lambda)^i$ in Eq. (7).

Figure 10. The mean squared error between NP and $NPwTF$ for the Enron dataset in relation to ε

4.5. Diversity of Node Position Compared to Other Measures

Node position measure appears to be more diverse than the other measures. It can be visible especially while analyzing number of nodes than possess the same centrality value, Fig. 11. Node positions are better for every value of ε, compared to degree prestige (DP) and degree centrality (DC). Note that degree prestige function provides only 286 distinct values for Enron and 208 for WUT. In case of degree centrality, there are only 383 distinct values for Enron and 242 for WUT. For that reason, the percentage of duplicates exceeds 95% for degree measures whereas it is below 60% for node positions in Enron and below 11% for WUT, Fig. 11.

4.6. Ranking Comparison

To compare rankings created upon different measures the Kendall's coefficient of concordance was used. It determines the similarity between two ranking lists. Let X and Y be any n–item rankings, then Kendall's coefficient of concordance $\kappa(X,Y)$ can be evaluated from the following formula (Kendall, 1948): $\kappa(X,Y) = \frac{1}{n(n-1)} \cdot \sum_{i=1}^{n} \sum_{j=1}^{n} sgn(x_j - x_i) \cdot sgn(y_j - y_i)$; where: x_i and y_i are the positions of the same ith item in ranking X and Y, respectively; they range from 1 to n; $sgn(x_j - x_i)$ is the sign of the difference $x_j - x_i$. It means that if item i follows item j in ranking X, then $sgn(x_j - x_i) = -1$; if they are at the same position $sgn(x_j - x_i) = 0$; otherwise $sgn(x_j - x_i) = +1$. When two rankings

Figure 11. Percentage of duplicates within the set of node measures, separately for node position with different values of ε, degree prestige (DP), and degree centrality (DC)

have the same items at every position, Kendall's coefficient for them is equal to +1. However, when two rankings have all the same items but they occur in reverse order, their Kendall's coefficient equals -1.

Kendall's coefficient was calculated separately for each pair of user rankings based on values of degree centrality (DC), degree prestige (DP), and node position for different ε, Tab. 2 and Tab. 3. The similarity of rankings based on node position calculated for different ε provided an obvious correlation: the greater difference in ε, the less similar are rankings. However, for any two values of ε, Kendall's coefficient was relatively high and always greater than 0.82. Hence, node position is the stable measure that depends on ε to limited extent.

Simultaneously, NP–based rankings were different from both DC– and DP–based: κ was from -0.75 (WUT) to only 0.07 (Enron). The closeness between DC– and DP–based ranking was rather high: $\kappa = 0.35$ for Enron and as much as 0.79 for WUT. DC– and DP–based rankings are close each other and differ from NP–based because both DC and DP provide big number of duplicates and do not distinguish users (see Sec. 4.5). It reveals that DC and DP deliver similar, limited knowledge about users in the network whereas node position function is the diverse, meaningful measure.

Table 2. Kendall's coefficient for Enron

	ε=0.01	ε=0.1	ε=0.3	ε=0.5	ε=0.7	ε=0.9	DC
ε=0.1	0.9988						
ε=0.3	0.8727	0.8732					
ε=0.5	0.8623	0.8627	0.9850				
ε=0.7	0.8474	0.8478	0.9681	0.9822			
ε=0.9	0.8308	0.8311	0.9484	0.9620	0.9796		
DC	0.0041	0.0041	0.0084	0.0081	0.0077	0.0074	
DP	0.0052	0.0052	0.0081	0.0079	0.0077	0.0746	0.3517

4.7. Top Network Nodes

After analyzing the individual ranking position based on node position measure, it appears that one of the highest position in the Enron social network occupies Kenneth Lay: the 5th place for ε=0.01, the 2nd for ε=0.1 and ε=0.3, the 1st for ε=0.5 and ε=0.7, and finally the 4th place for ε=0.9. Kenneth Lay was the former Chairman of the Board and Chief Executive Officer who was accused and sentenced for broad range of financial crimes. Another Enron employee Vince Kaminski, who was risk–manager and as one of the first uncovered the frauds in Enron, takes the 9th place for ε=0.01, the 5th for ε=0.1, ε=0.3, and ε=0.5, the 3rd for ε=0.7, and finally the 1st place for ε=0.9.

Table 3. Kendall's coefficient for WUT

	ε=0.01	ε=0.1	ε=0.3	ε=0.5	ε=0.7	ε=0.9	DC
ε=0.1	0.9782						
ε=0.3	0.9399	0.9612					
ε=0.5	0.9054	0.9262	0.9638				
ε=0.7	0.8691	0.8886	0.9237	0.9582			
ε=0.9	0.8197	0.8355	0.8652	0.8967	0.9366		
DC	-0.6874	-0.6946	-0.7083	-0.6931	-0.7353	-0.7497	
DP	-0.6655	-0.6716	-0.6829	-0.7215	-0.7027	-0.7099	0.7919

Top email users in WUT are: 1 faculty library, Science Information Center, 4 individuals - network administrators, 1 trade union, Promotion Department, 1 dean, and Ph.D. Office.

The lists of top 10 email users in both organizations are rather stable regardless ranking function, Tab. 4 and 5. It means that top users can change their rank positions in relation to ε. However, the changes are rather insignificant and these users still remain on the top of the ranking lists.

5. Conclusions

Node position is a measure for the importance of a user in the social network that reflects the characteristic of the user's neighborhood. Its value for the given individual respects both node positions of the nearest acquaintances as well as their attention directed to the considered user. Thus, the node position measure NP provides the opportunity to analyze the social network with respect to social behaviors of individuals. In case of the email social network ESN, these behaviors are represented by sent emails and node position values reflect the importance of email users with respect to email communication within the considered community, e.g. users of the single mail server, employees in one organization or company, and the like.

The node position function appears to be a powerful, stable and diverse measure, which can be used to select key users for project teams (Kazienko, Musiał, 2007), find new potential employees, search the best potential consumers for advertising campaigns or recommender systems (Kazienko, Musiał, 2006), and finally for use in target marketing (Yang, Dia, Cheng, Lin, 2006).

Acknowledgements
This work was partly supported by The Polish Ministry of Science and Higher Education, grant no. N516 037 31/3708.

Table 4. Top 10 email users from the Enron company

Pos.		ε=0.01	ε=0.1	ε=0.3	ε=0.5	ε=0.7	ε=0.9	*DC*	*DP*
1	*ID*	305254	291808	332868	299611	299611	337528	335273	299865
	NP	4.065	18.865	31.131	39.804	36.846	20.150	0.074	0.051
2	*ID*	283853	299611	299611	332868	332868	337732	269248	327409
	NP	3.889	18.630	30.776	38.852	35.447	19.513	0.071	0.042
3	*ID*	331364	326239	326239	326239	337528	337735	327409	335273
	NP	3.695	14.849	30.720	37.690	34.305	18.408	0.054	0.042
4	*ID*	291808	332868	325789	325789	325789	299611	266263	323998
	NP	3.687	14.638	29.974	37.268	34.097	18.029	0.051	0.041
5	*ID*	299611	337528	337528	337528	326239	332868	321650	266263
	NP	3.663	14.396	29.862	37.006	33.759	17.036	0.047	0.039
6	*ID*	300777	325789	305254	305254	305254	325789	323998	340221
	NP	3.515	13.893	26.186	32.270	29.361	16.516	0.046	0.039
7	*ID*	326239	305254	291693	291693	291693	326239	290282	337361
	NP	3.285	13.645	22.850	27.670	24.894	15.713	0.045	0.036
8	*ID*	332868	283853	283853	283853	283853	305254	307264	321650
	NP	3.264	11.883	22.637	27.305	24.655	13.942	0.044	0.035
9	*ID*	337528	291693	291808	291808	291808	291693	261852	309676
	NP	3.240	11.533	21.165	25.629	22.598	12.347	0.042	0.034
10	*ID*	280434	295169	331364	331364	331364	283853	273245	290282
	NP	3.239	11.096	18.494	21.869	19.329	11.733	0.041	0.032

6. Bibliography

Adamic, L.A. and Adar, E. (2003) Friends and Neighbors on the Web. *Social Networks* **25**, 3, 211–230.

Bavelas, A. (1950) Communication patterns in task – oriented groups. *Journal of the Acoustical Society of America* **22**, 271–282.

Berkhin, A. (2005) A Survey on PageRank Computing. *Internet Mathematics* **2**, 1, 73–120.

Botafogo, R.A., Rivlin, E. and Shneiderman, B. (1992) Structural analysis of hypertexts: identifying hierarchies and useful metrics. *ACM Transaction on Information Systems* **10**, 2, 142–180.

Boyd, D.M. (2004) Friendster and Publicly Articulated Social Networking. *CHI 2004, ACM Press*, 1279–1282.

Brin, S and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Comp. Networks and ISDN Syst.* **30**, 1–7, 107–117.

Brinkmeier, M. (2006) PageRank Revisited. *ACM Transactions on Internet Technology* **6**, 3, 282–301.

Culotta, A., Bekkerman, R. and McCallum, A. (2004) Extracting social networks and contact information from email and the Web. *CEAS*

Table 5. Top 10 email users from the WUT community

Pos.		ε=0.01	ε=0.1	ε=0.3	ε=0.5	ε=0.7	ε=0.9	*DC*	*DP*
1	*ID*	13265	13265	13265	13265	5679	5679	9686	9686
	NP	1.678	7.191	15.660	18.890	20.795	30.241	0.319	0.104
2	*ID*	3575	3575	3575	3575	846	846	14151	14151
	NP	1.601	6.498	14.193	17.511	17.468	26.565	0.257	0.088
3	*ID*	14151	14151	14151	14151	13265	59745	2578	749
	NP	1.527	5.861	12.917	16.441	16.791	19.974	0.155	0.0618
4	*ID*	54	54	54	54	14151	96	13253	2578
	NP	1.466	5.317	11.764	15.416	16.285	16.989	0.137	0.060
5	*ID*	9686	9686	9686	9686	54	2275	5171	13265
	NP	1.426	4.966	11.015	14.590	16.239	16.925	0.132	0.059
6	*ID*	498	498	498	5679	3575	845	498	5171
	NP	1.423	4.894	10.475	13.341	16.137	16.709	0.125	0.056
7	*ID*	2275	2275	2275	498	9686	1169	13265	1169
	NP	1.273	3.632	8.293	13.122	15.538	15.280	0.122	0.055
8	*ID*	2578	2578	5679	2275	2275	54	4472	59745
	NP	1.253	3.356	7.425	12.184	15.239	13.794	0.115	0.055
9	*ID*	1437	1437	2578	846	498	9686	1066	135
	NP	1.225	3.063	6.920	10.87	12.700	13.207	0.111	0.052
10	*ID*	1066	1066	1437	2578	59745	59841	7650	2786
	NP	1.220	3.030	6.014	8.913	11.888	12.126	0.108	0.051

2004, First Conference on Email and Anti-Spam.

Flake, G., Lawrence, S. and Lee Giles, C. (2000) Efficient identification of web communities. *6th ACM SIGKDD*, 150–160.

Freeman, L.C. (1979) Centrality in social networks: Conceptual clarification *Social Networks* **1**, 3, 215–239.

Garton, L., Haythornthwaite, C. and Wellman, B. (1997) Studying Online Social Networks. *J. of Computer-Mediated Communication* **3**, 1.

Gibson, D., Kleinberg, J. and Raghavan, P. (1998) Inferring Web communities from link topology. *9th ACM Conference on Hypertext and Hypermedia*, 225-234.

Golbeck, J. (2005) Computing and Applying Trust in Web-Based Social Networks. *Dissertation Submitted to the Faculty of the Graduate School of the Universtity of Maryland.*

Golbeck, J. and Hendler, J.A. (2004) Accuracy of Metrics for Inferring Trust and Reputation in Semantic Web-Based Social Networks. *EKAW 2004, LNCS 3257, Springer Verlag*, 116–131.

Hanneman, R. and Riddle, M. (2006) Introduction to social network methods. Online textbook, available from: http://faculty.ucr.edu/∼hanneman /nettext/.

Katz, L. (1953) A new status derived from sociometrics analysis. *Psychometrica* **18**, 39–43.

Kazienko, P. and Adamski, M. (2007) AdROSA - Adaptive Personalization of Web Advertising. *Information Sciences* **177**, 11, 2269–2295.

Kazienko, P. and Musiał, K. (2006) Recommendation Framework for Online Social Networks. *AWIC 2006, Studies in Computational Intelligence, Springer Verlag*, **23**, 111–120.

Kazienko, P. and Musiał, K. (2007) On Utilizing Social Networks to Discover Representatives of Human Communities. *International Journal of Intelligent Information and Database Systems*, **1**, 3/4, 293–310.

Kazienko, P., Musiał K. and Zgrzywa, A. (2007) Evaluation of Node Position Based on Mutual Interaction in Social Network of Internet Users. *TPD 2007, Wydawnictwo Politechniki Poznańskiej*, 265–276.

Kazienko, P. and Musiał K. (2008) Mining Personal Social Features in the Community of Email Users. *SOFSEM 2008, LNCS 4910, Springer Verlag*, 708–719.

Kendall, M.G. (1948) Rank correlation methods. *Charles Griffin & Company, Ltd., London.*

Musiał K., Kazienko, P. and Kajdanowicz T. (2008) Multirelational Social Networks in Multimedia Sharing Systems. Chapter in Knowledge Processing and Reasoning for Information Society. *Academic Publishing House EXIT, Warsaw*, 275-292.

Priebey, C.E., Conroy, J.M., Marchette, D.J. and Park, Y. (2005) Scan Statistics on Enron Graphs. *Computational & Mathematical Organization Theory* **11**, 3, 229–247.

Rana, O.F. and Hinze, A. (2004) Trust and reputation in dynamic scientific communities. *IEEE Distributed Systems Online* **5**, 1.

Shetty, J. and Adibi, J. (2005) Discovering Important Nodes through Graph Entropy The Case of Enron Email Databases. *3rd International Workshop on Link Discovery, ACM Press*, 74–81.

Valverde, S.,Theraulaz, G.,Gautrais, J.,Fourcassie, V.,Sole, R.V. (2006) Self-organization patterns in wasp and open source communities. *IEEE Intelligent Systems* **21**, 2, 36–40.

Wasserman, S. and Faust, K. (1994) Social network analysis: Methods and applications. *Cambridge University Press, New York.*

Wellman, B. and Salaff, J. (1996) Computer Networks as Social Networks: Collaborative Work, Telework, and Virtual Community. *Annual Review Sociol* **22**, 213–238.

Yang, W.S., Dia, J.B., Cheng, H.C. and Lin, H.T. (2006) Mining Social Networks for Targeted Advertising. *HICSS-39, IEEE Comp. Soc.*, 137a.